BEYOND FITTING: KNOWLEDGE-GUIDED LLM REASONING FOR COMBINATORIAL GENE PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Combinatorial perturbation prediction aims to estimate transcriptional responses induced by simultaneous perturbation of multiple genes, which is essential for advancing a mechanistic understanding of biochemical processes. Existing deep learning methods primarily rely on learned representations to predict perturbation responses, often lacking explicit reasoning and mechanistic interpretability. Meanwhile, emerging LLM-enhanced methods still struggle with continuous-response prediction and inductive generalization to unseen perturbation genes. To address these limitations, we propose **CombPert**, a biological knowledge-guided LLM reasoning framework for **Comb**inatorial **Pert**urbation prediction. Specifically, CombPert first constructs a task-specific biological knowledge graph from curated biological resources and dataset-specific co-expression statistics, and retrieves mechanistic paths connecting the perturbation genes and the target gene. To improve inductive generalization and response magnitude estimation, CombPert further incorporates similarity-based evidence from similar perturbation genes and conditional statistical priors. Finally, CombPert integrates these sources into a structured CoT prompt, supporting Reasoning Elicitation for training-free prediction and mechanistic explanation, and Reasoning Synthesis for CoT data generation and downstream fine-tuning. Extensive experiments on representative benchmarks demonstrate strong predictive performance and inductive generalization to unseen perturbation genes. Our code is available at `https://anonymous.4open.science/r/CombPert-submit-FBA5`.

1 INTRODUCTION

Understanding how cellular transcriptional states respond to simultaneous genetic perturbations is central to drug discovery, synthetic lethality screening, and synthetic biology (O'Neil et al., 2017; Morris et al., 2023; Wei et al., 2026b). Enabled by single-cell sequencing and CRISPR-based perturbation technologies, Perturb-seq (Dixit et al., 2016; Horlbeck et al., 2018) and its combinatorial extensions have made it possible to profile transcriptional responses at scale, revealing diverse interaction patterns such as synergy, suppression, and neomorphism (Norman et al., 2019). Yet exhaustive experimental screening remains infeasible: nearly 20,000 protein-coding genes yield over 200 million possible dual-gene perturbations, making comprehensive wet-lab validation impractical. This challenge has motivated a growing line of deep learning approaches for perturbation prediction, including variational autoencoders (Lotfollahi et al., 2019), compositional latent models (Lotfollahi et al., 2023), graph neural networks (Roohani et al., 2024), and large-scale single-cell foundation models (Cui et al., 2024; Hao et al., 2024). Most of these methods follow a representation learning paradigm, encoding perturbations and cellular context into latent representations and decoding them into transcriptional responses. While effective, such models typically behave as black-box predictors, providing limited insight into the biological evidence or reasoning underlying their predictions.

Recently, large language models (LLMs) have shown strong capabilities in complex reasoning and domain-specific problem solving (Liu et al., 2024; Yang et al., 2025; Peng et al., 2025; Zhang et al., 2025; Wei et al., 2022; Guo et al., 2025; Li et al., 2025). Driven by their extensive biomedical knowledge and reasoning capabilities, LLMs have recently been explored for perturbation prediction in several studies (Lu et al., 2024). For instance, methods such as scLAMBDA (Wang et al., 2024)

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.

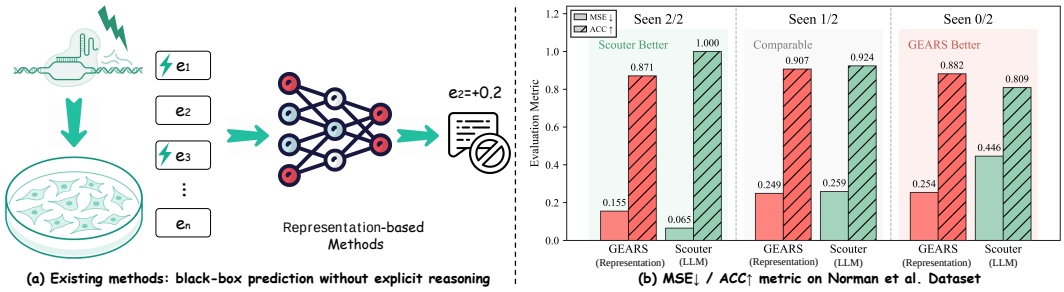

Figure 1: Challenges of LLM-enhanced methods in combinatorial perturbation prediction. (a) Existing methods lack explicit mechanistic reasoning. (b) LLM-enhanced methods perform well in Seen 2/2 but struggle to generalize to Seen 0/2. MSE↓ and ACC↑ denote better performance.

and Scouter (Zhu & Li, 2026) use LLMs as static feature extractors to enhance perturbation or gene representations for downstream prediction. In contrast, methods such as VCWorld (Wei et al., 2026a) and PerturbQA (Wu et al., 2025) employ chain-of-thought (CoT) prompting to elicit step-by-step reasoning over perturbation effects. Despite substantial progress, these LLM-enhanced methods still suffer from some critical limitations.

As illustrated in Figure 1(a), **existing LLM-enhanced methods largely remain black-box regressors**: LLMs only serve as representation encoders, while downstream predictors produce perturbation responses without explicit mechanistic reasoning. Although recent methods attempt to leverage explicit LLM reasoning for perturbation analysis, they mainly target simplified settings, such as single-gene perturbation prediction or gene-level binary classification. **Such methods cannot be directly applied to combinatorial perturbation prediction**, where multi-gene interventions induce a vast perturbation space and non-additive genetic interactions, including synergy, suppression, neomorphism, and redundancy/epistasis. Moreover, Figure 1(b) highlights **the limited inductive generalization of existing LLM-enhanced methods.** On the Norman et al. dataset, although Scouter achieves near-optimal performance in the Seen 2/2 setting and clearly outperforms the representation-based GEARS model, its performance drops sharply in Seen 0/2, where both perturbation genes are unseen during training. This contrast suggests that existing LLM-enhanced methods can fit seen perturbation patterns, but struggle to generalize to unseen combinatorial perturbations without explicit biological grounding. Together, these observations call for biologically grounded reasoning that can capture regulatory mechanisms and support reliable response estimation under unseen perturbation settings. Building on this insight, we pose the following research question to guide our study: *Can reasoning enable LLMs to model nonlinear gene interactions and generalize to unseen perturbations?*

To bridge these gaps, we propose **CombPert**, a biological knowledge-guided LLM reasoning framework for **Comb**inatorial **Pert**urbation prediction. Given two perturbed genes and a target gene, CombPert predicts the target gene's differential expression (DE) response and generates a plausible explanation. Specifically, CombPert first constructs a task-specific biological knowledge graph (KG) from curated resources and dataset-specific co-expression statistics, and retrieves mechanistic paths describing how the perturbed genes may independently or jointly affect the target gene. Furthermore, we introduce similarity-based evidence from similar perturbation genes and conditional statistical priors to support inductive reasoning. Finally, CombPert integrates these sources into a structured CoT prompt. We investigate its effectiveness under two paradigms: *Reasoning Elicitation* for training-free prediction and *Reasoning Synthesis* for CoT data generation and downstream fine-tuning. In summary, the contributions of this paper are as follows:

- We propose CombPert, a biological knowledge-guided LLM reasoning framework for combinatorial perturbation prediction, which predicts continuous target-gene responses and generates mechanistic explanations.
- We address inductive combinatorial perturbation prediction by introducing similarity-based perturbation evidence and conditional response priors, enabling CombPert to calibrate response magnitudes and generalize beyond observed perturbation pairs.
- Extensive experiments across diverse combinatorial perturbation settings demonstrate the effectiveness of CombPert under two paradigms: Reasoning Elicitation and Reasoning Synthesis.

2 Related Work

Existing methods for perturbation prediction can be broadly categorized into two paradigms.

Representation Learning Methods. Early studies focused on relatively simple single-cell perturbation settings and modeled perturbation effects using latent space arithmetic. scGen (Lotfollahi et al., 2019) utilized variational autoencoders to model perturbation effects as translation vectors, implicitly assuming additivity. CPA (Lotfollahi et al., 2023) advanced this by decomposing effects into compositional representations, yet struggled to generalize to entirely unseen genes. Moving beyond latent space modeling, GEARS (Roohani et al., 2024) introduces a GNN-based paradigm that leverages Gene Ontology to propagate information between genes, improving generalization to unseen perturbations. However, its reliance on binary adjacency matrices limits mechanistic expressiveness, and its additive formulation imposes implicit linearity assumptions. Besides, single-cell foundation models (Cui et al., 2024; Hao et al., 2024; Theodoris et al., 2023) further scale up representation learning approaches by pre-training on millions of cells, leading to improved representation capacity. However, as highlighted by Ahlmann-Eltze et al. (2025), both these methods and optimal transport-based approaches (Bunne et al., 2023) remain constrained by statistical pattern matching, lacking explicit modeling of causal mechanisms for accurate combinatorial prediction.

LLM-Enhanced Methods. LLMs have recently shown promise for modeling complex biological relationships in functional genomics. One paradigm treats LLMs as static feature encoders, where methods such as scLAMBDA (Wang et al., 2024), Scouter (Zhu & Li, 2026), and GenePert (Chen & Zou, 2024) convert gene descriptions into embeddings for downstream regression, improving feature quality but failing to exploit LLMs' reasoning capabilities. A second paradigm employs LLMs as direct reasoning engines. VCWorld (Wei et al., 2026a) and PerturbQA (Wu et al., 2025) leverage structured knowledge graphs to prompt step-by-step reasoning. However, these methods simplify continuous prediction into binary classification, precluding quantitative analysis of complex interactions such as synergy or neomorphism. Furthermore, their reasoning architectures are typically restricted to single-path propagation, limiting their ability to model combinatorial perturbations.

3 Preliminaries

3.1 Task Formulation

We formulate combinatorial perturbation prediction as a regression problem over gene expression responses. Let $\mathcal{D} = \{(\mathbf{g}^i, \mathcal{C}^i)\}_{i=1}^N$, where $\mathbf{g}^i \in \mathbb{R}^K$ is the expression profile of cell i over K genes, and $\mathcal{C}^i = (c_1^i, \ldots, c_M^i)$ is the set of applied perturbations, with $M = 0$ denoting the control condition. Following prior work (Roohani et al., 2024), we model responses at the condition level by averaging single-cell measurements. For a perturbation condition $\mathcal{C}$, let $\bar{\mathbf{g}}_{\mathcal{C}} \in \mathbb{R}^K$ denote the mean expression vector, and define the differential expression relative to control as $\mathbf{y}_{\mathcal{C}} = \bar{\mathbf{g}}_{\mathcal{C}} - \bar{\mathbf{g}}_{\text{ctrl}}$. The goal is to predict $\mathbf{y}_{\mathcal{C}}$ using a mapping f_θ from perturbation conditions to gene expression changes. In our framework, an LLM predicts each target gene's differential expression from a perturbation query and relevant biological evidence, yielding gene-wise predictions across the expression profile.

3.2 Knowledge Graph Paths

Given a knowledge graph $\mathcal{G} = \{(e, r, e') \mid e, e' \in \mathcal{E},\ r \in \mathcal{R}\}$, where $\mathcal{E}$ is the set of entities and $\mathcal{R}$ is the set of relation types. A path of length l from a source entity e_s to a target entity e_t is defined as an ordered sequence of typed triples

$$p^{(l)}(e_s, e_t) = \big((e_1, r_1, e_2), (e_2, r_2, e_3), \ldots, (e_l, r_l, e_{l+1})\big), \tag{1}$$

where $e_1 = e_s$, $e_{l+1} = e_t$, and $(e_i, r_i, e_{i+1}) \in \mathcal{G}$ for $i = 1, \ldots, l$. Here, the path length l is equivalent to the number of hops. For example, consider a perturbation query where ***BPGM*** is the perturbed gene and ***GYPA*** is the target gene. The following path provides a possible mechanistic connection:

$$p^{(3)}(\text{BPGM}, \text{GYPA}) = \big((\text{BPGM}, r_1, \text{UBC}), (\text{UBC}, r_2, \text{YY1}), (\text{YY1}, r_3, \text{GYPA})\big), \tag{2}$$

where r_1 = *interacts_with*, r_2 = *activated_by*, and r_3 = *activates*. This path indicates that the perturbation signal may propagate from ***BPGM*** to ***GYPA*** through intermediate biological entities ***UBC*** and ***YY1***, providing interpretable evidence for target-gene response prediction. We further use $\mathcal{P}^{(\leq L)}(e_s, e_t)$ to denote the set of all paths from e_s to e_t with at most L hops in $\mathcal{G}$.

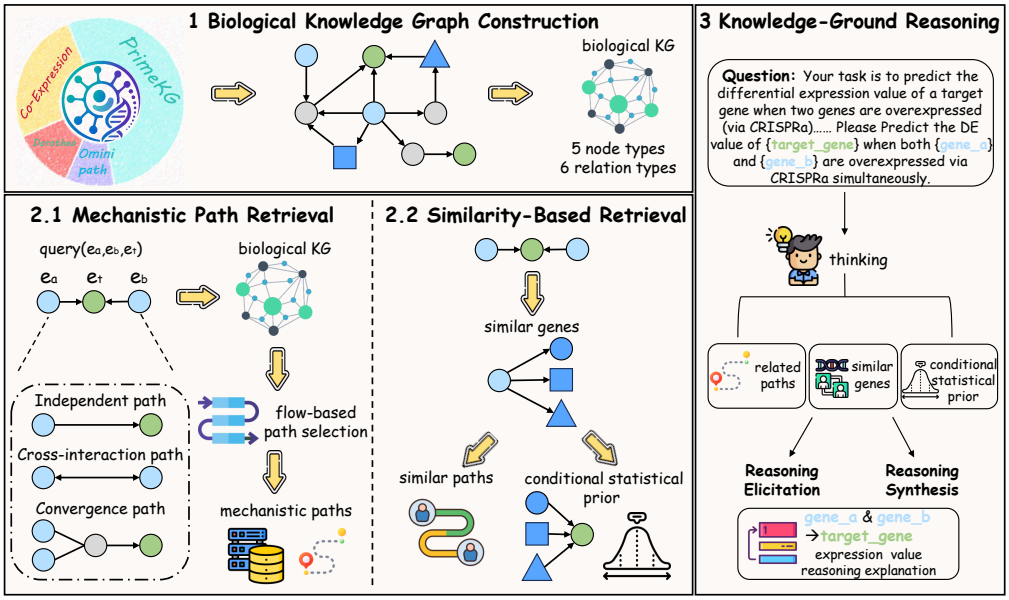

Figure 2: **Overview of CombPert.** The framework integrates biological knowledge graph construction, evidence retrieval, and knowledge-guided LLM reasoning to predict continuous differential expression responses and generate plausible mechanistic explanations under two paradigms.

4 METHOD

Overview. In this section, we introduce CombPert, a biological knowledge-guided LLM reasoning framework for combinatorial perturbation prediction. As illustrated in Figure 2, CombPert first constructs a task-specific biological knowledge graph by integrating curated biological resources with dataset-specific co-expression statistics. Given two perturbed genes and a target gene, CombPert then retrieves mechanistic paths from the KG to characterize independent effects, cross-gene interactions, and convergent regulation patterns. To improve inductive generalization and calibrate response magnitude, CombPert further incorporates similarity-based evidence from similar perturbation genes and conditional statistical priors. Finally, these evidence sources are integrated to guide the LLM in predicting the target gene's differential expression response and generating a mechanistic explanation.

4.1 BIOLOGICAL KNOWLEDGE GRAPH CONSTRUCTION

We first construct a task-specific biological knowledge graph by integrating curated interaction resources with a broader biomedical graph. Formally, we represent biological knowledge as a heterogeneous graph $\mathcal{G} = \{(e, r, e') \mid e, e' \in \mathcal{E},\ r \in \mathcal{R}\}$, where $\mathcal{E}$ denotes entities and $\mathcal{R}$ denotes relation types. Entities include genes/proteins and functional concepts, while relations encode biological interactions and associations. Starting from perturbation-relevant anchor genes, including perturbed genes and candidate target genes with strong differential responses, we extract a constrained PrimeKG (Chandak et al., 2023) subgraph and add external genes only when they connect at least two anchors through protein-protein interactions. We then augment the graph with signaling and physical interaction edges from OmniPath (Türei et al., 2016), transcription factor–target regulation edges from DoRothEA (Garcia-Alonso et al., 2019), and co-expression edges estimated from control cells. The final graph consists of five node types and six relation types, as detailed in Appendix A.1.

4.2 MECHANISTIC PATH RETRIEVAL

Given the constructed biological knowledge graph, we next retrieve mechanism-relevant evidence to explain how perturbations affect downstream gene expression. Formally, given a perturbation query $q(e_A, e_B, e_t)$ with a perturbed gene pair $(e_A, e_B) \in \mathcal{E} \times \mathcal{E}$ and a target gene $e_t \in \mathcal{E}$, we extract a concise set of mechanistic paths to form an informative subgraph $\mathcal{G}_{\text{sub}} \subset \mathcal{G}$. The retrieval procedure consists of two stages: path retrieval and path selection. Specifically, we first retrieve multiple types of

path structures to capture the diverse biological mechanisms underlying combinatorial perturbations, and then refine the resulting candidates using a reliability- and diversity-aware selection strategy.

Complementary Path Retrieval. We retrieve three complementary types of path structures, namely independent paths, cross-interaction paths, and convergence paths. Their definitions are as follows:

- **Independent paths**, which describe the individual effect of each perturbed gene on the target gene. For each $e \in \{e_A, e_B\}$, these are paths in $\mathcal{P}(e, e_t)$.
- **Cross-interaction paths**, which characterize the direct physical or functional crosstalk between the two perturbed genes. These are paths in $\mathcal{P}(e_A, e_B)$.
- **Convergence paths**, which capture the integration of perturbation signals through a shared mediator $e_m \in \mathcal{E}$. Specifically, a convergence structure is formed when there exist paths from e_A to e_m, from e_B to e_m, and from e_m to e_t.

Accordingly, we denote the sets of retrieved independent, cross-interaction, and convergence paths by $\mathcal{P}^{\text{ind}}$, $\mathcal{P}^{\text{cross}}$, and $\mathcal{P}^{\text{conv}}$, respectively.

Flow-based Path Selection. After initial path retrieval, many trivial and redundant candidates may remain. Inspired by redundancy-aware path pruning (Chen et al., 2026), we further select paths based on reliability and diversity. For a candidate path $p = (e_1, r_1, e_2, \ldots, e_l, r_l, e_{l+1})$, we define a flow-based score that estimates the strength of signal propagation along the path. Let $s(e_1) = 1$ be the initial signal at the source node. The propagated signal at node e_i is computed recursively as:

$$s(e_i) = s(e_{i-1}) \cdot \alpha(r_{i-1}) \cdot \beta(e_i), \quad i = 2, \ldots, l+1, \tag{3}$$

where $\alpha(r_{i-1})$ is the relation-specific decay factor and $\beta(e_i)$ is the node-specific penalty. The reliability of p is defined as: $R(p) = \left(\prod_{i=1}^{l+1} s(e_i)\right)^{\frac{1}{l+1}}$. We then apply Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998) to encourage structural diversity: $\text{MMR}(p) = \lambda \widetilde{R}(p) - (1 - \lambda) \max_{p' \in \mathcal{S}} \text{Jaccard}(p, p')$, where $\widetilde{R}(p)$ is the normalized flow-based score, $\mathcal{S}$ is the set of already selected paths, and λ controls the trade-off between reliability and redundancy. Here, $\text{Jaccard}(p, p')$ measures the overlap between two paths based on their intermediate node sets, with a larger value indicating higher structural redundancy. For each path category, we retain the top-k_p selected paths based on the MMR score:

$$\mathcal{P}_{\text{base}} = \arg \text{ top-}k_{p_{p \in \mathcal{P}}} \text{MMR}(p). \tag{4}$$

Finally, we aggregate the selected paths from the three categories as $\mathcal{P}_{\text{base}} = \mathcal{P}^{\text{ind}}_{\text{base}} \cup \mathcal{P}^{\text{cross}}_{\text{base}} \cup \mathcal{P}^{\text{conv}}_{\text{base}}$ for downstream LLM reasoning.

4.3 Similarity-Based Retrieval

While mechanistic paths from observed perturbation genes provide structured evidence for signal propagation, they may be limited for unseen or sparsely observed perturbations and insufficient to capture response magnitude. We therefore complement them with similar-gene information for generalization and a conditional statistical prior for magnitude estimation.

Similar Perturbation Gene Retrieval. To provide empirical grounding for unseen or weakly observed perturbation genes, we retrieve similar genes from the training set under the assumption that genes with similar biological roles, network contexts, and expression characteristics tend to induce similar downstream responses. Formally, for each perturbation gene $e \in \{e_A, e_B\}$ and candidate training gene $e' \in \mathcal{E}_{\text{train}}$, we define a multi-view similarity score as:

$$S(e, e') = w_{\text{topo}} S_{\text{topo}}(e, e') + w_{\text{func}} S_{\text{func}}(e, e') + w_{\text{expr}} S_{\text{expr}}(e, e') + w_{\text{sem}} S_{\text{sem}}(e, e'), \tag{5}$$

where S_{topo}, S_{func}, S_{expr}, and S_{sem} denote topological, functional, co-expression, and semantic similarity, respectively, with $w_{\text{topo}} + w_{\text{func}} + w_{\text{expr}} + w_{\text{sem}} = 1$. Specifically, S_{topo} is computed as the Jaccard similarity between the 1-hop gene neighborhoods of e and e' in the knowledge graph, S_{func} as the Jaccard similarity between their annotated GO terms, S_{expr} as the Pearson correlation between their control-condition expression profiles, and S_{sem} as the cosine similarity between their textual gene-summary embeddings. Candidate genes are ranked by $S(e, e')$, and the top-k_s retrieved genes are retained as similar perturbation genes. For each retrieved similar gene, we further retrieve its mechanistic paths to the target gene e_t following Section 4.2. The resulting path set is denoted by $\mathcal{P}_{\text{sim}}$ and used as similarity-based evidence for downstream reasoning.

Conditional Statistical Prior. To complement mechanistic evidence with quantitative guidance, we construct a conditional statistical prior for each target gene e_t using responses to similar-gene perturbations in the training data. Let $\mathcal{N}_{k_{sp}}(e)$ denote the set of top-k_{sp} similar genes retrieved for the query gene e. We define a local sample set $\mathcal{D}(e_t)$ by collecting training-set perturbation samples whose perturbed gene belongs to $\mathcal{N}_{k_{sp}}(e)$ and for which the differential expression value of e_t is available. For each sample $j \in \mathcal{D}(e_t)$, let $e^{(j)}$ denote the perturbed gene in that sample and let $y_t^{(j)}$ denote the observed differential expression value of e_t. We assign a relevance weight $\omega_j = S(e, e^{(j)})$, where $S(\cdot,\cdot)$ is the multi-view similarity score defined in Eq. equation 5. The conditional mean and standard deviation are then computed as:

$$\mu_{\text{cond}}(e_t) = \frac{\sum_{j\in\mathcal{D}(e_t)} \omega_j y_t^{(j)}}{\sum_{j\in\mathcal{D}(e_t)} \omega_j}, \qquad \sigma_{\text{cond}}(e_t) = \left(\frac{\sum_{j\in\mathcal{D}(e_t)} \omega_j \left(y_t^{(j)} - \mu_{\text{cond}}(e_t) \right)^2}{\sum_{j\in\mathcal{D}(e_t)} \omega_j} \right)^{\frac{1}{2}}. \quad (6)$$

The resulting pair $\mathcal{T}_{\text{cond}} = (\mu_{\text{cond}}(e_t), \sigma_{\text{cond}}(e_t))$ summarizes the expected response and its variability among similar perturbations, providing a local numerical reference to mechanistic reasoning.

4.4 Knowledge-Grounded Reasoning

Finally, we formulate perturbation prediction as a knowledge-grounded CoT reasoning task. Given a query $q(e_A, e_B, e_t)$, we integrate mechanistic paths $\mathcal{P}_{\text{base}}$, similarity-based evidence $\mathcal{P}_{\text{sim}}$, and conditional statistical priors $\mathcal{T}_{\text{cond}}$ into a structured prompt to support two reasoning paradigms.

Reasoning Elicitation. We use the structured prompt to guide the LLM in predicting the target-gene response and generating a plausible mechanistic explanation without updating model parameters:

$$(\hat{y}_t, \hat{E}) = f_{\text{parse}}\big(\text{LLM}\big(f_{\text{prompt}}\big(q(e_A, e_B, e_t), \mathcal{P}_{\text{base}}, \mathcal{P}_{\text{sim}}, \mathcal{T}_{\text{cond}}\big)\big)\big), \quad (7)$$

where $\hat{y}_t$ is the predicted target-gene response magnitude and $\hat{E}$ is the mechanistic explanation.

Reasoning Synthesis. We further use CombPert with a teacher LLM to generate CoT-annotated examples for training queries. We select examples based on prediction accuracy for supervised fine-tuning of student models to improve their generalization to unseen perturbations.

5 Experiments

5.1 Experimental Setups.

We evaluate CombPert on two representative benchmarks: the Norman et al. dataset (Norman et al., 2019) (CRISPRa perturbations in K562 cells) and the Replogle et al. dataset (Replogle et al., 2022) (CRISPRi perturbations in K562 cells). We assess two paradigms: *Reasoning Elicitation* with GPT-5.4, GPT-5.4-mini, Qwen3-32B, Qwen3-8B, and LLaMA-3.1-8B; and *Reasoning Synthesis* using GPT-5.4 as the teacher to fine-tune three open-source LLMs. We compare against two baseline groups: representation learning methods, including Linear (Ahlmann-Eltze et al., 2025), CPA (Lotfollahi et al., 2023), and GEARS (Roohani et al., 2024), and LLM-enhanced methods, including Scouter (Zhu & Li, 2026), scLAMBDA (Wang et al., 2024), and few-shot prompting with GPT-5.4, GPT-5.4-mini, Qwen3-8B, and Qwen3-32B. We also include a Mean baseline that uses each gene's training-set mean DE as its test-set prediction, along with an evidence-matched KG-RAG (Soman et al., 2024) for comparison. Following prior work (Roohani et al., 2024), we evaluate performance using mean squared error (MSE), Pearson correlation coefficient (PCC), and direction accuracy (ACC). Implementation and dataset details are provided in Appendices A.3 and A.4, respectively.

5.2 Main Results

Tables 1 and 2 summarize the main results. Overall, CombPert achieves strong performance under both reasoning paradigms, with GPT-5.4 under Reasoning Elicitation attaining the best overall MSE, PCC, and ACC on both datasets. Compared with representation-based methods, CombPert shows strong generalization to unseen perturbation genes. On Norman, CombPert with GPT-5.4 reduces GEARS's MSE from 0.249 to 0.158 in Seen 1/2 and from 0.254 to 0.232 in Seen 0/2. Compared with

Table 1: Performance comparison on the Norman et al. dataset. ❄ denotes Reasoning Elicitation, while ♣ denotes Reasoning Synthesis. Best results are **bolded**, and second-best results are underlined.

Model / Method	Seen 2/2			Seen 1/2			Seen 0/2			Overall		
	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑
Mean	0.758	0.000	0.621	0.675	0.000	0.650	0.680	0.000	0.523	0.691	0.000	0.620
KG-RAG (Soman et al., 2024)	0.092	0.961	0.986	0.305	0.758	0.907	0.629	0.348	0.640	0.337	0.706	0.863
Linear (Ahlmann-Eltze et al., 2025)	0.183	0.892	0.961	0.361	0.800	0.913	0.489	0.706	0.862	0.357	0.796	0.910
CPA (Lotfollahi et al., 2023)	0.664	0.589	0.779	0.551	0.634	0.813	0.430	0.704	0.806	0.545	0.641	0.805
GEARS (Roohani et al., 2024)	0.155	0.771	0.871	0.249	0.806	0.907	0.254	**0.800**	0.882	0.233	0.799	0.895
Scouter (Zhu & Li, 2026)	**0.065**	**0.982**	**1.000**	0.259	0.824	0.924	0.446	0.607	0.809	0.265	0.805	0.913
scLAMBDA (Wang et al., 2024)	0.153	0.943	0.996	0.266	0.856	0.951	0.414	0.683	0.850	0.278	0.834	0.937
GPT-5.4	0.090	0.975	0.994	0.279	0.785	0.927	0.597	0.282	0.649	0.314	0.710	0.881
GPT-5.4-mini	0.628	0.296	0.614	0.465	0.356	0.667	0.483	0.346	0.656	0.498	0.343	0.655
Qwen3-8B	0.732	0.101	0.575	0.608	0.139	0.614	0.672	0.181	0.503	0.643	0.142	0.584
Qwen3-32B	0.742	0.235	0.579	0.532	0.307	0.615	0.540	0.281	0.606	0.571	0.288	0.606
Llama3.1-8B	1.121	0.041	0.578	0.955	0.001	0.575	1.359	0.053	0.441	1.072	0.019	0.547
CombPert (Our Methods)												
❄ GPT-5.4	0.073	0.971	0.986	**0.158**	**0.879**	**0.960**	**0.232**	0.796	0.917	**0.159**	**0.877**	**0.955**
❄ GPT-5.4-mini	0.076	0.965	0.990	0.172	0.852	0.935	0.249	0.781	0.891	0.172	0.857	0.935
❄ Qwen3-8B	0.128	0.921	0.952	0.237	0.816	0.926	0.345	0.741	0.900	0.241	0.817	0.926
♣ Qwen3-8B	0.117	0.933	0.967	0.221	0.832	0.931	0.284	0.751	0.885	0.216	0.832	0.928
❄ Qwen3-32B	0.099	0.949	0.974	0.197	0.864	0.937	0.321	0.765	0.908	0.207	0.857	0.937
♣ Qwen3-32B	0.122	0.941	0.975	0.196	0.845	0.939	0.245	0.791	**0.921**	0.194	0.851	0.941
❄ Llama3.1-8B	0.201	0.931	0.964	0.352	0.711	0.884	0.407	0.683	0.831	0.337	0.744	0.890
♣ Llama3.1-8B	0.192	0.935	0.972	0.343	0.709	0.890	0.338	0.701	0.845	0.315	0.748	0.895

Table 2: Performance comparison on the Replogle et al. dataset. ❄ denotes Reasoning Elicitation, while ♣ denotes Reasoning Synthesis. Best results are **bolded**, and second-best results are underlined.

Model / Method	Seen 2/2			Seen 1/2			Seen 0/2			Overall		
	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑
Mean	0.206	0.000	0.520	0.211	0.000	0.488	0.196	0.000	0.483	0.208	0.000	0.495
KG-RAG (Soman et al., 2024)	0.037	0.956	0.900	0.125	0.761	0.829	0.749	0.440	0.417	0.196	0.762	0.785
Linear (Ahlmann-Eltze et al., 2025)	0.060	0.823	0.980	0.090	0.774	0.863	0.078	0.480	0.633	0.082	0.742	0.858
CPA (Lotfollahi et al., 2023)	0.145	0.535	0.720	0.121	0.576	0.685	0.209	0.147	0.533	0.141	0.501	0.671
GEARS (Roohani et al., 2024)	0.125	0.367	0.510	**0.074**	0.598	0.563	**0.061**	0.583	0.500	0.085	0.538	0.540
Scouter (Zhu & Li, 2026)	0.023	0.845	0.980	0.085	0.554	0.788	0.169	0.161	0.567	0.082	0.567	0.805
scLAMBDA (Wang et al., 2024)	0.048	0.848	0.980	0.082	0.759	0.875	0.112	0.411	0.700	0.077	0.729	0.875
GPT-5.4	0.026	0.958	0.980	0.177	0.364	0.787	0.533	-0.492	0.383	0.193	0.386	0.775
GPT-5.4-mini	0.156	0.490	0.590	0.206	0.409	0.563	0.361	0.460	0.533	0.217	0.437	0.555
Qwen3-8B	0.299	0.127	0.530	0.342	0.157	0.538	0.945	0.182	0.467	0.422	0.153	0.525
Qwen3-32B	0.189	0.462	0.580	0.259	0.456	0.617	0.457	0.412	0.667	0.271	0.451	0.615
Llama3.1-8B	0.428	0.079	0.580	0.464	0.070	0.538	1.010	0.210	0.450	0.536	0.093	0.535
CombPert (Our Methods)												
❄ GPT-5.4	**0.012**	**0.960**	**1.000**	0.078	**0.778**	0.886	0.176	**0.588**	**0.790**	**0.076**	**0.795**	**0.900**
❄ GPT-5.4-mini	0.033	0.768	0.980	0.085	0.738	**0.908**	0.215	0.546	0.717	0.091	0.717	0.898
❄ Qwen3-8B	0.102	0.692	0.960	0.124	0.621	0.850	0.742	0.330	0.633	0.211	0.597	0.845
♣ Qwen3-8B	0.091	0.704	0.975	0.108	0.637	0.855	0.681	0.340	0.618	0.190	0.609	0.849
❄ Qwen3-32B	0.148	0.695	0.970	0.139	0.603	0.879	0.391	0.426	0.683	0.179	0.598	0.873
♣ Qwen3-32B	0.138	0.687	0.971	0.136	0.584	0.881	0.315	0.452	0.696	0.164	0.590	0.876
❄ Llama3.1-8B	0.115	0.704	0.930	0.118	0.586	0.758	0.380	0.375	0.683	0.157	0.584	0.790
♣ Llama3.1-8B	0.106	0.708	0.938	0.109	0.584	0.764	0.311	0.393	0.697	0.139	0.586	0.797

LLM-enhanced methods, CombPert with GPT-5.4 outperforms Scouter and scLAMBDA across all overall metrics on both datasets. Both reasoning paradigms also improve upon few-shot inference. Under Reasoning Elicitation, integrating biological and statistical evidence improves GPT-5.4's Seen 0/2 ACC on Norman from 0.649 to 0.917. Under Reasoning Synthesis, fine-tuned students outperform their few-shot counterparts across all overall metrics on both datasets. For example, Qwen3-8B on Norman reduces MSE from 0.643 to 0.216 and improves PCC from 0.142 to 0.832 and ACC from 0.584 to 0.928. Reasoning Synthesis further reduces overall MSE relative to Reasoning Elicitation for all three student backbones on both datasets, suggesting that the synthesized examples provide useful supervision for learning evidence-guided prediction through parameter updates. Together, these results support the utility of CombPert for both inference-time guidance and downstream model training. To further assess the robustness and effectiveness of CombPert, we report statistical significance and full-panel analyses in Appendices B.5 and B.9, respectively.

5.3 Ablation Study

We conduct ablation experiments to assess each component's contribution to CombPert with GPT-5.4 as the backbone. Table 3 reports the results on both datasets. The results show that removing any key component consistently degrades performance, with more pronounced drops on the Seen 0/2 subset.

Effect of Mechanistic Path Retrieval. We examine the role of mechanistic path retrieval, which provides biological grounding by connecting perturbation genes and target genes. Removing all

Table 3: Ablation study on the key components of CombPert across two datasets. "*w/o*" means removing or replacing the module. Best results are **bolded**, and second-best results are underlined.

Model Variant	Norman et al. Dataset						Replogle et al. Dataset					
	Overall			Seen 0/2			Overall			Seen 0/2		
	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑
CombPert	**0.159**	**0.877**	**0.955**	**0.232**	**0.796**	**0.917**	**0.076**	**0.795**	0.900	**0.176**	**0.588**	**0.790**
w/o KG Paths	0.202	0.799	0.902	0.341	0.712	0.852	0.108	0.707	0.865	0.250	0.552	0.733
w/o Complementary Paths	0.193	0.814	0.912	0.317	0.744	0.872	0.106	0.707	0.885	0.293	0.498	0.700
w/o Flow-based Path Selection	0.234	0.802	0.917	0.369	0.696	0.835	0.100	0.737	0.885	0.290	0.518	0.716
w/o Conditional Statistical Prior	0.272	0.757	0.877	0.387	0.652	0.798	0.111	0.751	0.890	0.321	0.576	0.617
w/o Similar Gene Retrieval	0.255	0.810	0.902	0.376	0.661	0.808	0.099	0.715	**0.903**	0.247	0.548	0.733
w/o Biological Guidance	0.182	0.835	0.935	0.283	0.760	0.903	0.112	0.683	0.880	0.318	0.482	0.700

KG paths substantially degrades overall performance, increasing MSE from 0.159 to 0.202 on the Norman et al. dataset and from 0.076 to 0.108 on the Replogle et al. dataset. Moreover, removing complementary paths by retaining only independent paths, or replacing flow-based path selection with random scoring also consistently hurts performance, indicating the importance of diverse mechanistic evidence and relevance-aware selection for capturing combinatorial effects and reducing noise.

Effect of Similarity-Based Retrieval. We analyze the role of similarity-based retrieval, which provides evidence and numerical calibration from similar seen genes. *w/o Conditional Statistical Prior* removes the training-set DE distribution calibration, while *w/o Similar Gene Retrieval* removes examples from similar seen genes. Removing either component leads to clear performance drops, especially in the Seen 0/2 setting. These results suggest that calibration and similar examples provide complementary signals for predicting responses to unseen perturbations.

Effect of Biological Guidance. We remove the task-specific biological guidance so that the model directly predicts the response from the assembled context. The resulting performance drop highlights the value of biological guidance in integrating retrieved evidence for continuous response prediction.

5.4 Analysis

We further assess the effectiveness of CombPert through the following research questions (RQ): (1) **RQ1:** Can CombPert capture complex genetic interactions? (2) **RQ2:** How effective is the conditional statistical prior for response magnitude estimation under inductive settings? (3) **RQ3:** Can CombPert achieve accurate predictions with limited training data? Besides, we analyze single-gene perturbations, hyperparameter sensitivity, and mechanistic reasoning in Appendices B.2, B.3, and B.4, respectively. We evaluate cost and latency, explanation quality and faithfulness, and synthetic data utility in Appendices B.6, B.7, and B.8, respectively, and present a case study in Appendix C.

5.4.1 Gene Interaction Analysis (RQ1)

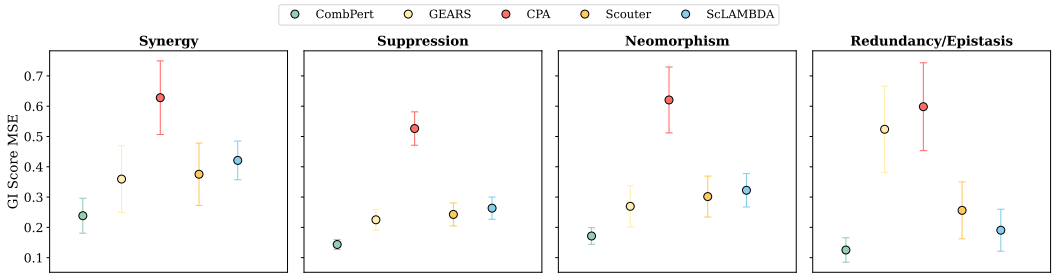

Figure 3: Stratified evaluation of complex genetic interaction prediction across genetic interaction types. GI Score MSE (↓) measures the error of non-additive interaction residuals.

To evaluate whether CombPert can capture complex genetic interactions in combinatorial perturbations, we compute the Genetic Interaction Score MSE (GI Score MSE), with details provided in Appendix B.1. Following prior definitions (Roohani et al., 2024), we stratify perturbation pairs into Synergy, Suppression, Neomorphism, and Redundancy/Epistasis. As shown in Figure 3, CombPert achieves the lowest GI Score MSE across all interaction types, indicating its stronger ability to model

complex non-additive effects. These consistent gains suggest that grounding LLM reasoning in biological evidence improves its ability to characterize diverse gene–gene interaction patterns.

5.4.2 Statistical Prior Analysis (RQ2)

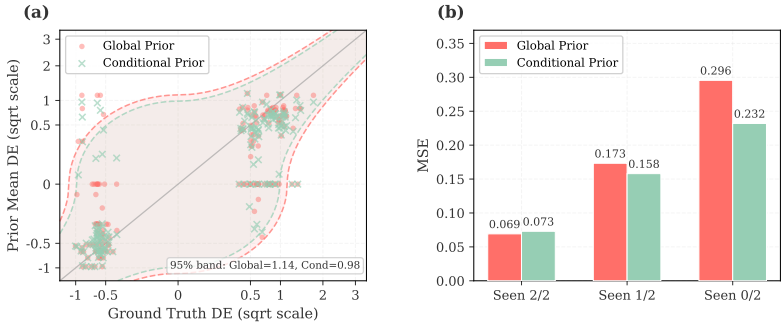

Figure 4: Effectiveness of the conditional statistical prior. (a) Prior mean comparison under the inductive Seen 0/2 setting. (b) MSE comparison across Seen 2/2, Seen 1/2, and Seen 0/2 splits.

To assess whether the conditional statistical prior improves response magnitude estimation under inductive settings, we compare it against a global prior that uses the overall response distribution without conditioning on similar perturbation genes. As shown in Figure 4(a), on the challenging Seen 0/2 split, the conditional prior mean aligns more closely with ground-truth differential expression values, yielding a narrower 95% confidence band than the global prior (b=0.98 vs. b=1.14). Figure 4(b) further shows that the conditional prior consistently reduces MSE across all novelty splits, with the largest gain on Seen 0/2. These results indicate that similar-gene conditioning provides a more informative prior for response magnitude estimation and improves inductive generalization.

5.4.3 Data Efficiency Analysis (RQ3)

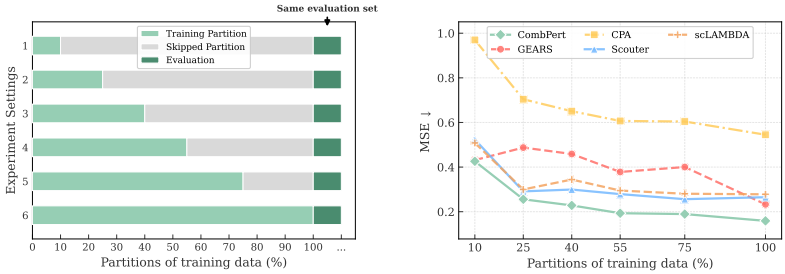

Figure 5: Data efficiency analysis. Left: training-data fraction settings with a shared evaluation set. Right: overall MSE of different methods under varying training fractions.

To assess the robustness of CombPert with limited training data, we vary the training fraction (10%, 25%, 40%, 55%, 75%, 100%) while keeping the held-out test set fixed. As shown in Figure 5, CombPert consistently achieves the lowest overall MSE across all fractions, outperforming baselines even with only a small portion of training data. These results demonstrate that CombPert achieves competitive performance with substantially fewer supervised perturbation examples.

6 Conclusion

In this paper, we proposed CombPert, a biological knowledge-guided LLM reasoning framework for combinatorial perturbation prediction. By integrating mechanistic paths, similarity-based evidence, and conditional statistical priors into structured CoT prompts, CombPert supports Reasoning Elicitation for training-free prediction and mechanistic explanation, and Reasoning Synthesis for CoT data generation and downstream fine-tuning. Extensive experiments demonstrate strong generalization to unseen perturbations, improved prediction of non-additive interactions, and competitive performance with limited data. Additional evaluations further support the framework's generality and interpretability, demonstrating its effectiveness in single-gene and full-panel prediction, sensitivity to mechanistic evidence, and ability to generate plausible and evidence-consistent explanations.

AI USE STATEMENT

We used generative AI tools to polish the manuscript and generate synthetic training data to evaluate the data synthesis capability of CombPert. We assessed the utility of the generated data by fine-tuning models and evaluating them on held-out perturbations. The authors reviewed the AI-polished text for accuracy and checked the experimental results supporting the reported conclusions. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with AI assistance.

REFERENCES

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.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. Nature methods, 20(11):1759–1768, 2023.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 335–336, 1998.

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

Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. Pathrag: Pruning graph-based retrieval augmented generation with relational paths. In AAAI, pp. 30183–30191. AAAI Press, 2026.

Yiqun Chen and James Zou. Genepert: Leveraging genept embeddings for gene perturbation prediction. bioRxiv, pp. 2024–10, 2024.

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

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.

Elena A Feshchenko, Evgeniya V Smirnova, Gayathri Swaminathan, Anjali M Teckchandani, Rachana Agrawal, Hamid Band, Xiaolong Zhang, Roland S Annan, Steven A Carr, and Alexander Y Tsygankov. Tula: an sh3-and uba-containing protein that binds to c-cbl and ubiquitin. Oncogene, 23(27):4690–4706, 2004.

Luz Garcia-Alonso, Christian H Holland, Mahmoud M Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome research, 29(8):1363–1375, 2019.

Joseph L Goldstein and Michael S Brown. Regulation of the mevalonate pathway. Nature, 343(6257): 425–430, 1990.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. Nature methods, 21(8):1481–1491, 2024.

Max A Horlbeck, Albert Xu, Min Wang, Neal K Bennett, Chong Y Park, Derek Bogdanoff, Britt Adamson, Eric D Chow, Martin Kampmann, Tim R Peterson, et al. Mapping the genetic landscape of human cells. Cell, 174(4):953–967, 2018.

Jay D Horton, Joseph L Goldstein, Michael S Brown, et al. Srebps: activators of the complete program of cholesterol and fatty acid synthesis in the liver. The Journal of clinical investigation, 109(9):1125–1131, 2002.

Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. arXiv preprint arXiv:2505.21318, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

Joseph Longo, Peter J Mullen, Rosemary Yu, Jenna E van Leeuwen, Mehdi Masoomian, Dixon TS Woon, Yuzhuo Wang, Eric X Chen, Robert J Hamilton, Joan M Sweet, et al. An actionable sterol-regulated feedback loop modulates statin sensitivity in prostate cancer. Molecular metabolism, 25: 119–130, 2019.

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.

Zhiyong Lu, Yifan Peng, Trevor Cohen, Marzyeh Ghassemi, Chunhua Weng, and Shubo Tian. Large language models in biomedicine and health: current research landscape and future directions. Journal of the American Medical Informatics Association, 31(9):1801–1811, 2024.

John A Morris, Christina Caragine, Zharko Daniloski, Júlia Domingo, Timothy Barry, Lu Lu, Kyrie Davis, Marcello Ziosi, Dafni A Glinos, Stephanie Hao, et al. Discovery of target genes and pathways at gwas loci by pooled single-cell crispr screens. Science, 380(6646):eadh7699, 2023.

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.

Nigel J O'Neil, Melanie L Bailey, and Philip Hieter. Synthetic lethality and cancer. Nature Reviews Genetics, 18(10):613–623, 2017.

Miao Peng, Nuo Chen, Zongrui Suo, and Jia Li. Rewarding graph reasoning process makes llms more generalized reasoners. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 2257–2268, 2025.

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.

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

Karthik Soman, Peter W. Rose, John H. Morris, Rabia E. Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A. Nelson, Sui Huang, and Sergio E. Baranzini. Biomedical knowledge graph-optimized prompt generation for large language models. Bioinform., 40(9), 2024.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. Nature, 618(7965):616–624, 2023.

Christine B. F. Thien and Wallace Y. Langdon. Cbl: many adaptations to regulate protein tyrosine kinases. Nature Reviews Molecular Cell Biology, 2(4):294–307, 2001. doi: 10.1038/35067100.

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

Katarina Vukojević, Violeta Šoljić, Vlatka Martinović, Fila Raguž, and Natalija Filipović. The ubiquitin-associated and sh3 domain-containing proteins (ubash3) family in mammalian development and immune response. International journal of molecular sciences, 25(3):1932, 2024.

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with sclambda. bioRxiv, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. VCWorld: A biological world model for virtual cell simulation. In The Fourteenth International Conference on Learning Representations, 2026a. URL https://openreview.net/forum?id=hhq89Hs7T3.

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, 2026b.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. Contextualizing biological perturbation experiments through language. arXiv preprint arXiv:2502.21290, 2025.

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

Bohan Zhang, Xiaokang Zhang, Jing Zhang, Jifan Yu, Sijia Luo, and Jie Tang. Cot-based synthesizer: Enhancing llm performance through answer synthesis. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6286–6303, 2025.

Ouyang Zhu and Jun Li. Scouter predicts transcriptional responses to genetic perturbations with large language model embeddings. Nature Computational Science, 6(1):21–28, 2026.

A Experimental Details

A.1 Details of Graph Construction

We build the graph in three stages: initialization, augmentation, and assembly. Tables 4 and 5 summarize the types, sources, and descriptions of nodes and relations, respectively.

Graph Initialization. We start from task-specific anchor genes derived from the perturbation dataset, including perturbed genes and candidate target genes with strong differential responses. Using these anchors, we extract a constrained subgraph from PrimeKG (Chandak et al., 2023). To improve task relevance, we retain only node types useful for mechanistic reasoning and include external genes only when they connect at least two anchor genes through protein–protein interactions.

Knowledge Augmentation. We augment the initialized graph with mechanistic relations not fully covered by PrimeKG. Specifically, we add directed signaling and physical interaction edges from OmniPath (Türei et al., 2016), high-confidence transcription factor–target regulation edges from DoRothEA (Garcia-Alonso et al., 2019), and co-expression edges estimated from control cells to capture dataset-specific gene dependencies.

Final Graph Assembly. The final graph is a unified heterogeneous graph with **five node types**: Gene/Protein, Pathway, and three Gene Ontology categories—Biological Process, Molecular Function, and Cellular Component. It contains **six relation types**: protein–protein interaction, signaling, transcription factor regulation, co-expression, gene–GO association, and gene–pathway membership, derived from PrimeKG, OmniPath, DoRothEA, and dataset-specific co-expression statistics. We standardize all relations into these edge categories while preserving attributes such as direction, sign, confidence, and correlation strength when available.

Table 4: Node types in the biological knowledge graph.

Node Type	Source	Description
Gene/Protein	PrimeKG	Dataset genes and additional bridge genes that connect at least two dataset genes.
Biological Process	PrimeKG (Gene Ontology)	GO biological process terms.
Molecular Function	PrimeKG (Gene Ontology)	GO molecular function terms.
Cellular Component	PrimeKG (Gene Ontology)	GO cellular component terms.
Pathway	PrimeKG (Reactome)	Signaling and metabolic pathway entities.

Table 5: Edge types in the biological knowledge graph.

Edge Type	Source	Description
protein-protein interaction	PrimeKG (STRING, BioGRID, HuRI, CORUM)	Undirected protein-protein interaction edges with source database annotations.
signaling	OmniPath	Directed signaling edges with activation/inhibition signs.
transcription factor regulation	DoRothEA	Directed transcriptional regulation edges with regulation type and confidence level.
co-expression	Control cells	Undirected gene–gene co-expression edges computed from control cells using Pearson correlation filtering with a threshold of at least 0.4.
gene-GO association	PrimeKG (Gene Ontology)	Undirected gene-to-GO annotation edges.
gene-pathway membership	PrimeKG (Reactome)	Undirected gene-to-pathway membership edges.

A.2 KNOWLEDGE-GUIDED COT PROMPT DESIGN

To enable LLMs to perform biologically grounded reasoning, we design a knowledge-guided CoT prompt for combinatorial perturbation prediction. As shown in Figure 6, the prompt is organized into five parts: role specification, task instruction, perturbation context, retrieved biological evidence, and structured output requirements. Together, these components specify the role of the LLM, clarify the combinatorial perturbation prediction task, provide gene-level and condition-level contextual information, incorporate retrieved biological evidence, and regulate the final response into a structured prediction with an explicit reasoning trace. Besides, we emphasize that all perturbation-response examples and conditional priors in the prompt are derived exclusively from the training partition. Held-out perturbation responses are neither included in the prompt nor used for evidence retrieval or prior construction.

A.3 EXPERIMENTAL DETAILS

Implementation Details. For all LLM backbones, we set the temperature to 0.6 and the maximum generation length to 2048 tokens. For mechanistic path retrieval, we set the maximum path length L to 4 and retain the top k_p=3 paths per category using MMR with a diversity trade-off λ=0.7. For similarity-based retrieval, we retrieve k_s=3 functionally similar genes, weighting the multi-view similarity dimensions as w_{topo}=0.3, w_{func}=0.1, w_{expr}=0.1, and w_{sem}=0.5. To compute semantic similarity, we obtain functional summaries from the National Center for Biotechnology Information (NCBI) (Zhu & Li, 2026) and encode them using OpenAI's `text-embedding-3-large` model [1]. For the conditional statistical prior, we use k_{sp}=20 similar perturbation examples to estimate the corresponding prior mean and standard deviation. For Reasoning Synthesis, we use CombPert with GPT-5.4 as the teacher to generate reasoning traces and response predictions separately for each dataset. After excluding outputs with missing traces or invalid numerical predictions, we retain the 50% of valid examples with the lowest squared errors against observed training responses within each dataset. The retained traces and generated predictions are used to fine-tune Qwen3-32B, Qwen3-8B, and LLaMA-3.1-8B. All generation and filtering are restricted to the training split. Experiments are conducted on a server equipped with eight NVIDIA GeForce RTX 4090 GPUs.

Metrics. We evaluate prediction performance using mean squared error (MSE), Pearson correlation coefficient (PCC), and direction accuracy (ACC). For each perturbation condition, let $\{y_i\}_{i=1}^n$ and $\{\hat{y}_i\}_{i=1}^n$ denote the ground-truth and predicted differential expression responses of the n evaluated target genes, respectively.

Mean Squared Error. MSE measures the discrepancy between predicted and ground-truth response magnitudes:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2. \tag{8}$$

A lower MSE indicates more accurate estimation of perturbation-induced expression changes.

Pearson Correlation Coefficient (PCC). PCC measures the linear agreement between predicted and ground-truth response patterns:

$$\text{PCC} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{9}$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote the mean ground-truth and predicted responses, respectively. A higher PCC indicates better agreement in the transcriptional response pattern.

Direction Accuracy (ACC). ACC measures the proportion of target genes whose predicted differential expression has the same direction as the ground truth:

$$\text{ACC} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[\text{sign}(\hat{y}_i) = \text{sign}(y_i)\right], \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function. A higher ACC indicates better consistency in predicting up- and down-regulation.

[1] `https://platform.openai.com/docs/guides/embeddings`

Inference Prompt Template for Combo Perturbation Prediction

POSITION: You are an expert computational biologist specializing in gene perturbation response prediction. The perturbation experiments use {perturbation_technique} in {cell_line}. {perturbation_technique_description}. When gene X is perturbed via {perturbation_technique}: Gene X itself is overexpressed/inhibited and downstream genes may be up-regulated or down-regulated depending on the biological function of gene X.

TASK DESCRIPTION: Your task is to predict the differential expression (DE) value of a target gene when one or two genes are perturbed (via {perturbation_technique}) in human {cell_line}. The DE value is defined as: DE = mean(perturbed cells) − mean(control cells). A positive DE means the target gene is up-regulated by the perturbation; a negative DE means it is down-regulated.

You will be given:
1. The target gene and its baseline expression level.
2. Knowledge-graph paths connecting the perturbed gene(s) to the target gene (signaling cascades, protein interactions, transcription-factor regulation, co-expression, etc.).
3. For combinatorial perturbations: known single-perturbation effects (when available from training data) and the additive baseline.
4. Statistical distribution of the target gene's DE values across training-set perturbations (as a calibration reference).
5. For unseen (OOD) perturbation genes: analogical examples from functionally similar seen genes, including their observed DE values and KG paths.

You must reason step-by-step through the biological mechanism and then output your final prediction as a single floating-point number on the last line in exactly this format: PREDICTION: <number>. For example: PREDICTION: -0.73.

USER PROMPT: Predict the DE value of {target_gene} when both {gene_a} and {gene_b} are perturbed via {perturbation_technique} simultaneously.

Baseline Expression Level: {target_gene} is at a level of {target_expr_desc} (raw value: {target_expr:.4f})
Known Single-Perturbation Effects are {single_pert_info}
Knowledge-Graph Paths include {kg_paths_text}
Similar cases and data include {statistical_prior_section} and {example_retrieval_section}

Please think and follow these five steps:

Step 1 Gene Function & Pathway Context. Describe the known biological functions of {gene_a}, {gene_b}, and {target_gene}. Identify relevant pathways and regulatory relationships. Analyze the baseline expression level of {target_gene}. Note that {target_gene} is {target_expr_desc}, and you should consider whether the baseline expression of {target_gene} may constrain the plausible response magnitude.

Step 2 Independent Effects. Based on the single-perturbation data (if available) and the KG paths, estimate how perturbing {gene_a} alone and {gene_b} alone would each affect {target_gene}.

Step 3 Interaction Topology. Analyze the cross-interaction paths between {gene_a} and {gene_b}. Are they in the same pathway? Do they converge on shared regulators of {target_gene}? Consider synergy, suppression, or additive effects.

Step 4 Quantitative Integration. Combine the independent effects with the interaction analysis. Carefully consider the additive baseline ({additive_baseline}) and whether the actual combinatorial effect is likely to slightly deviate from it (and in which direction).{statistical_calibration_hint}

Step 5 Final Prediction. State your predicted DE value for {target_gene} under the combined perturbation of {gene_a} and {gene_b} in exactly the following format: PREDICTION: <your numerical prediction>.

Figure 6: Inference prompt template for combinatorial perturbation prediction.

A.4 Dataset Details

We evaluate our proposed method on two large single-cell combinatorial perturbation datasets with distinct genetic perturbation mechanisms: the Norman et al. (Norman et al., 2019) dataset and the Replogle et al. (Replogle et al., 2022) dataset. For both datasets, we compute the differential expression of each gene by subtracting the mean expression of control cells from the mean expression of perturbed cells. Following common practice, the main experiments focus on the top 20 differentially expressed genes ranked by absolute differential expression for each perturbation condition. To assess

generalization beyond these highly responsive genes, we also evaluate the full measured gene panel on the Norman dataset, with results reported in Appendix B.9.

Norman et al. Dataset. This dataset profiles the transcriptional responses of K562 human erythroleukemia cells under CRISPRa (activation) perturbations. It encompasses 131 two-gene combinatorial perturbations and 105 unique perturbation genes. We split the dataset based on the generalizability to unseen genes. Specifically, 70 genes are randomly designated as training genes, while the remaining 35 genes are held out as out-of-distribution genes. To systematically evaluate the model's combinatorial generalization capability, the test set is stratified into three distinct difficulty levels based on the occurrence of the perturbed genes in the training combinatorial set:

- **Seen 2/2**: Both genes in the combinatorial perturbation have been independently observed in other training combinations.
- **Seen 1/2**: Only one gene in the combination has been observed in the training combinatorial set.
- **Seen 0/2**: Neither gene has been observed in the training combinatorial set, representing the most challenging out-of-distribution generalization scenario.

The detailed data split statistics for the Norman dataset are summarized in Table 6.

Table 6: Data split statistics for the Norman et al. dataset.

Perturbation Type	Category	Training & Validation Set	Test Set
Single Perturbations	Seen Single Gene Data	2300	0
Single Perturbations	Unseen Single Gene Data	0	740
Combinatorial Perturbations	Seen 2/2 Data	1040	280
Combinatorial Perturbations	Seen 1/2 Data	0	960
Combinatorial Perturbations	Seen 0/2 Data	0	340
Total Conditions		**3340**	**2320**

Replogle et al. Dataset. To further validate the robustness of our framework across different perturbation modalities, we use the experiment 6 subset from the Replogle et al. dataset. This dataset employs CRISPRi (interference) to perturb genes associated with cholesterol biosynthesis and DNA repair in K562 cells. The exact split statistics for the Replogle dataset are detailed in Table 7.

Table 7: Data split statistics for the Replogle et al. dataset.

Perturbation Type	Category	Training & Validation Set	Test Set
Single Perturbations	Seen Single Gene Data	680	0
Single Perturbations	Unseen Single Gene Data	0	140
Combinatorial Perturbations	Seen 2/2 Data	440	100
Combinatorial Perturbations	Seen 1/2 Data	0	240
Combinatorial Perturbations	Seen 0/2 Data	0	60
Total Conditions		**1120**	**540**

B Additional Experimental Results

B.1 Details of Gene Interaction Analysis

Given a perturbation query $q(e_A, e_B, e_t)$, where (e_A, e_B) denotes the perturbed gene pair and e_t denotes the target gene, we define the genetic interaction score (GI Score) as $\delta_{e_t}^{e_A+e_B} = \mathrm{DE}_{e_t}^{e_A+e_B} - (\mathrm{DE}_{e_t}^{e_A} + \mathrm{DE}_{e_t}^{e_B})$, where $\mathrm{DE}_{e_t}^{e_A+e_B}$ denotes the observed differential expression of e_t under the combinatorial perturbation (e_A, e_B), while $\mathrm{DE}_{e_t}^{e_A}$ and $\mathrm{DE}_{e_t}^{e_B}$ denote its single-perturbation effects under e_A and e_B, respectively. We then compute GI Score $\mathrm{MSE} = \frac{1}{|\mathcal{Q}|}\sum_{q(e_A,e_B,e_t)\in\mathcal{Q}}(\hat{\delta}_{e_t}^{e_A+e_B} - \delta_{e_t}^{e_A+e_B})^2$, where $\hat{\delta}_{e_t}^{e_A+e_B}$ is the predicted GI score and $\mathcal{Q}$ denotes the evaluated query set. This metric

directly evaluates whether a model can capture non-additive interaction effects beyond the sum of individual perturbations, providing a targeted measure of complex genetic interaction modeling.

B.2 Single-Gene Perturbation Analysis

Inference Prompt Template for Single Perturbation Prediction

POSITION: You are an expert computational biologist specializing in gene perturbation response prediction. The perturbation experiments use {perturbation_technique} in {cell_line}. {perturbation_technique_description}. When gene X is perturbed via {perturbation_technique}: Gene X itself is overexpressed/inhibited and downstream genes may be up-regulated or down-regulated depending on the biological function of gene X.

TASK DESCRIPTION: Your task is to predict the differential expression (DE) value of a target gene when one or two genes are perturbed (via {perturbation_technique}) in human {cell_line}. The DE value is defined as: DE = mean(perturbed cells) − mean(control cells). A positive DE means the target gene is up-regulated by the perturbation; a negative DE means it is down-regulated.

You will be given:
1. The target gene and its baseline expression level.
2. Knowledge-graph paths connecting the perturbed gene(s) to the target gene (signaling cascades, protein interactions, transcription-factor regulation, co-expression, etc.).
3. For combinatorial perturbations: known single-perturbation effects (when available from training data) and the additive baseline.
4. Statistical distribution of the target gene's DE values across training-set perturbations (as a calibration reference).
5. For unseen (OOD) perturbation genes: analogical examples from functionally similar seen genes, including their observed DE values and KG paths.

You must reason step-by-step through the biological mechanism and then output your final prediction as a single floating-point number on the last line in exactly this format: PREDICTION: <number>. For example: PREDICTION: -0.73.

USER PROMPT: Predict the DE value of {target_gene} when {gene_a} is perturbed via {perturbation_technique}.

Baseline Expression Level: {target_gene} is at a level of {target_expr_desc} (raw value: {target_expr:.4f})
Knowledge-Graph Paths include {kg_paths_text}
Similar cases and data include {statistical_prior_section} and {example_retrieval_section}

Please think and follow these three steps:

Step 1 Gene Function & Pathway Context. Describe the known biological functions of {gene_a} and {target_gene}. Identify relevant pathways and regulatory relationships. Analyze the baseline expression level of {target_gene}. Note that {target_gene} is {target_expr_desc}, and you should consider whether the baseline expression of {target_gene} may constrain the plausible response magnitude.

Step 2 Based on the pathway analysis and the KG paths, estimate the direction (up or down) and approximate magnitude of the DE change in {target_gene} when {gene_a} is perturbed via {perturbation_technique}.

Step 3 Final Prediction. State your predicted DE value for {target_gene} in exactly the following format: PREDICTION: <your numerical prediction>.

Figure 7: Inference prompt template for single perturbation prediction.

To further examine whether the proposed framework can generalize beyond combinatorial perturbations, we conduct an additional analysis on unseen single-gene perturbations. Although CombPert is primarily designed for combinatorial perturbation prediction, the single-gene setting provides a simpler but important testbed for evaluating whether the model can capture gene-specific regulatory effects under inductive conditions. Given a perturbation gene e_p and a target gene e_t, the model is asked to predict the transcriptional response of e_t after perturbing e_p. As shown in Figure 7, we adapt the knowledge-guided prompt to the single-perturbation scenario by removing the cross-gene interaction component and focusing on the biological context, mechanistic paths, and similar perturbation evidence related to e_p and e_t.

Table 8 reports performance on the top 20 DEGs for each perturbation. Under Reasoning Elicitation, CombPert with GPT-5.4 achieves the best results across all metrics, with an MSE of 0.108, a PCC of

Table 8: Performance comparison on unseen single-gene perturbations in the Norman et al. dataset. ❄ denotes Reasoning Elicitation; ♣ denotes Reasoning Synthesis. Best results are **bolded**, and second-best results are underlined.

Model / Method	Norman et al. Dataset (Unseen Single)		
	MSE↓	PCC↑	ACC↑
Mean	0.410±0.022	0.000±0.000	0.542±0.017
Representation-based Methods			
Linear (Ahlmann-Eltze et al., 2025)	0.264±0.020	0.507±0.033	0.789±0.014
CPA (Lotfollahi et al., 2023)	0.298±0.022	0.354±0.046	0.620±0.021
GEARS (Roohani et al., 2024)	0.130±0.015	0.698±0.037	0.881±0.022
LLM-enhanced Methods			
Scouter (Zhu & Li, 2026)	0.274±0.024	0.448±0.042	0.739±0.023
scLAMBDA (Wang et al., 2024)	0.233±0.020	0.577±0.039	0.809±0.021
GPT-5.4	0.304±0.018	0.299±0.039	0.578±0.020
GPT-5.4-mini	0.308±0.019	0.284±0.035	0.554±0.018
Qwen3-8B	0.421±0.024	-0.015±0.031	0.467±0.013
CombPert (Our Methods)			
❄ GPT-5.4	**0.108**±0.013	**0.869**±0.020	**0.937**±0.014
❄ GPT-5.4-mini	0.191±0.014	0.568±0.023	0.781±0.012
❄ Qwen3-8B	0.246±0.016	0.480±0.030	0.778±0.015
♣ Qwen3-8B	0.216±0.010	0.530±0.021	0.797±0.019

0.869, and an ACC of 0.937. Compared with GEARS, the strongest representation-based baseline, it reduces MSE from 0.130 to 0.108 and improves PCC from 0.698 to 0.869. It also substantially outperforms few-shot prompting with the same GPT-5.4 backbone, which obtains an MSE of 0.304, a PCC of 0.299, and an ACC of 0.578. Gains are also observed with smaller backbones: for example, CombPert reduces Qwen3-8B's MSE from 0.421 to 0.246. Under Reasoning Synthesis, fine-tuning Qwen3-8B on CombPert-generated data further reduces MSE from 0.246 to 0.216 and improves PCC from 0.480 to 0.530 and ACC from 0.778 to 0.797. These results support the utility of CombPert under both reasoning paradigms and suggest that its evidence-guided approach extends to single-gene as well as combinatorial perturbation prediction.

B.3 Hyperparameter Analysis

To assess CombPert's sensitivity to retrieval and path-scoring configurations, we examine six factors: the MMR diversity–relevance trade-off λ, edge-type priority, the number of retrieved similar genes k_s, relation decay, node penalty, and multi-view similarity weights. For edge-type priority, we consider five relation types: signaling (Sig.), transcription factor regulation (TF), protein–protein interaction (PPI), co-expression (Co.), and gene–GO association (GO).

As shown in Figure 8(a), an intermediate λ achieves the best performance, suggesting that balancing relevance and diversity is important: overly relevance-biased or diversity-biased retrieval may introduce redundant or noisy paths. Figure 8(b) further shows that prioritizing signaling edges yields the lowest MSE and highest direction accuracy, indicating that signaling relations provide more informative evidence for perturbation propagation, whereas less directly causal relations may introduce additional noise. Figure 8(c) shows that retrieving too few analogous examples provides insufficient evidence, while retrieving too many may introduce less relevant examples; thus, k_s=3 provides the best trade-off.

Figures 8(d) and 8(e) examine relation decay and node penalties, respectively. Replacing relation-specific decay with uniform values or scaling the default values degrades performance. Removing node penalties produces the largest deterioration among the tested penalty settings, while halving or increasing them also reduces accuracy. These results support the use of both components in flow-based path scoring. Figure 8(f) compares the default similarity weights with uniform weighting and the removal of individual similarity views. Each removal degrades both metrics, with removing

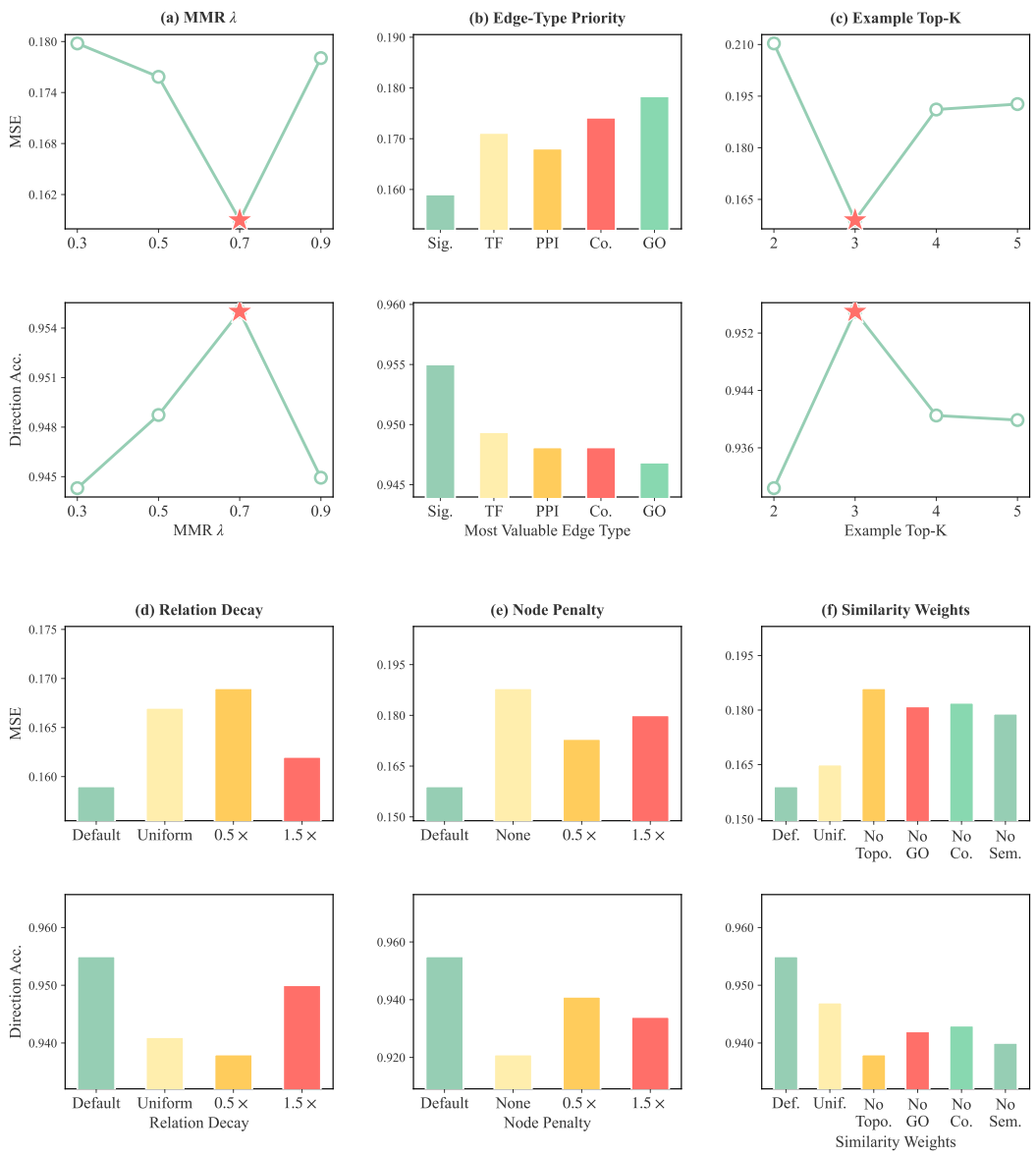

Figure 8: Hyperparameter sensitivity analysis of CombPert: (a) MMR trade-off λ, (b) edge-type priority, (c) number of retrieved similar genes k_s, (d) relation decay, (e) node penalty, and (f) similarity weights. We report MSE and direction accuracy (ACC).

topological similarity causing the largest MSE increase, suggesting that the different views provide complementary retrieval signals. Overall, the default configuration achieves the best performance across the tested settings.

B.4 Validation of Mechanistic Reasoning

To assess whether CombPert effectively incorporates retrieved mechanistic evidence into its reasoning, we conduct counterfactual interventions on the Norman et al. dataset. Specifically, we keep numerical priors and analogous evidence fixed while modifying only the retrieved mechanisms under three settings: key-path removal, key-path reversal, and random-path modification. The key path is the highest-ranked signed path selected according to Eq. equation 4, whereas the random path is a

matched lower-ranked path retrieved for the same query. We quantify prediction shift as the mean absolute difference between the original predictions of CombPert and those obtained under each intervention, and assess predictive accuracy using MSE.

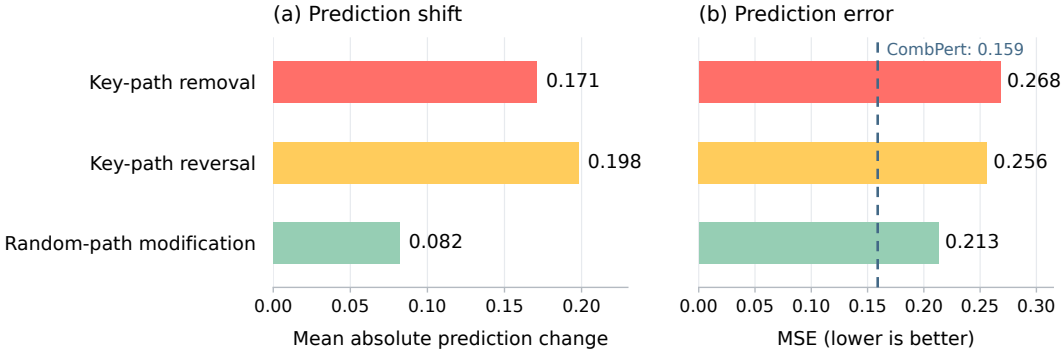

Figure 9: Validation of mechanistic reasoning: (a) mean absolute prediction shift; (b) mean squared error (MSE), with the dashed line indicating the original CombPert baseline.

As shown in Figure 9, key-path removal and reversal induce prediction shifts of 0.171 and 0.198, respectively, both more than twice that induced by random-path modification (0.082). These interventions also increase MSE from 0.159 to 0.268 and 0.256, respectively, compared with 0.213 under random-path modification. Together, these results suggest that CombPert effectively incorporates retrieved mechanistic evidence into its reasoning, with key paths playing an important role in predictive accuracy.

B.5 Statistical Significance Analysis

To evaluate the statistical significance of the models' performance, we report the standard error alongside the mean performance metrics. In our experiments, the standard error is calculated on a per-sample basis across the test set. Specifically, for each evaluation metric (MSE, PCC, ACC), we first compute the metric for each individual perturbation sample in the test set and then calculate the standard error of these sample-level metrics. This standard error reflects the variance in the model's prediction quality across different genetic perturbations, providing insights into whether a model performs consistently well across all target genes or if its average performance is skewed by a few extreme predictions.

Table 9: Performance on the Norman et al. dataset (mean ± standard error). ❄ denotes inference with CombPert (Reasoning Elicitation); ♣ denotes a model fine-tuned on CombPert-generated CoT data (Reasoning Synthesis). Best results are **bolded**, and second-best results are underlined.

Model / Method	Seen 2/2			Seen 1/2			Seen 0/2			Overall		
	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑
Mean	0.758±0.047	0.000±0.000	0.621±0.010	0.675±0.035	0.000±0.000	0.650±0.015	0.680±0.022	0.000±0.000	0.523±0.015	0.691±0.023	0.000±0.000	0.620±0.010
KG-RAG	0.092±0.018	0.961±0.005	0.986±0.004	0.305±0.025	0.758±0.034	0.907±0.021	0.629±0.036	0.348±0.051	0.640±0.042	0.337±0.028	0.706±0.032	0.863±0.021
Representation-based Methods												
Linear	0.183±0.024	0.892±0.012	0.961±0.006	0.361±0.029	0.800±0.020	0.913±0.014	0.489±0.023	0.706±0.023	0.862±0.008	0.357±0.019	0.796±0.013	0.910±0.009
CPA	0.664±0.054	0.589±0.023	0.779±0.013	0.551±0.054	0.634±0.032	0.813±0.016	0.430±0.035	0.704±0.022	0.806±0.012	0.545±0.035	0.641±0.020	0.805±0.010
GEARS	0.155±0.017	0.771±0.024	0.871±0.024	0.249±0.033	0.806±0.023	0.907±0.016	0.254±0.017	**0.800**±0.024	0.882±0.024	0.233±0.021	0.799±0.015	0.895±0.012
LLM-enhanced Methods												
Scouter	**0.065**±0.008	**0.982**±0.002	**1.000**±0.000	0.259±0.022	0.824±0.018	0.924±0.011	0.446±0.032	0.607±0.038	0.809±0.022	0.265±0.015	0.805±0.014	0.913±0.008
scLAMBDA	0.153±0.019	0.943±0.008	0.996±0.001	0.266±0.021	0.856±0.015	0.951±0.008	0.414±0.025	0.683±0.037	0.850±0.018	0.278±0.014	0.834±0.012	0.937±0.006
GPT-5.4	0.090±0.011	0.975±0.002	0.994±0.001	0.279±0.019	0.785±0.024	0.927±0.014	0.597±0.027	0.282±0.022	0.649±0.014	0.314±0.013	0.710±0.015	0.881±0.009
GPT-5.4-mini	0.628±0.045	0.296±0.018	0.614±0.012	0.465±0.021	0.356±0.024	0.667±0.014	0.483±0.026	0.346±0.016	0.656±0.012	0.498±0.034	0.343±0.021	0.655±0.013
Qwen3-8B	0.732±0.048	0.101±0.019	0.575±0.013	0.608±0.038	0.139±0.027	0.614±0.014	0.672±0.036	0.181±0.019	0.503±0.010	0.643±0.026	0.142±0.017	0.584±0.009
Qwen3-32B	0.742±0.048	0.235±0.019	0.579±0.013	0.532±0.038	0.307±0.027	0.615±0.014	0.540±0.036	0.281±0.019	0.606±0.010	0.571±0.026	0.288±0.017	0.606±0.009
Llama3.1-8B	1.121±0.097	0.041±0.052	0.578±0.010	0.955±0.084	0.001±0.025	0.575±0.014	1.359±0.073	0.053±0.019	0.441±0.009	1.072±0.086	0.019±0.023	0.547 ±0.013
CombPert (Our Methods)												
❄ GPT-5.4	0.073±0.007	0.971±0.004	0.986±0.003	**0.158**±0.010	**0.879**±0.012	**0.960**±0.008	**0.232**±0.013	0.796±0.013	0.917±0.006	**0.159**±0.007	**0.877**±0.008	**0.955**±0.005
❄ GPT-5.4-mini	0.076±0.007	0.965±0.005	0.990±0.003	0.172±0.010	0.852±0.010	0.935±0.007	0.249±0.021	0.781±0.020	0.891±0.008	0.172±0.008	0.857±0.008	0.935±0.004
❄ Qwen3-8B	0.128±0.011	0.921±0.012	0.952±0.008	0.237±0.017	0.816±0.016	0.926±0.010	0.345±0.019	0.741±0.017	0.900±0.006	0.241±0.000	0.817±0.010	0.926±0.006
♣ Qwen3-8B	0.117±0.009	0.933±0.016	0.967±0.010	0.221±0.010	0.832±0.009	0.931±0.021	0.284±0.015	0.751±0.021	0.885±0.010	0.216±0.007	0.832±0.011	0.928±0.006
❄ Qwen3-32B	0.099±0.013	0.949±0.012	0.974±0.006	0.197±0.012	0.864±0.011	0.937±0.008	0.321±0.013	0.765±0.012	0.908±0.006	0.207±0.008	0.857±0.008	0.937±0.005
♣ Qwen3-32B	0.122±0.007	0.941±0.022	0.975±0.006	0.196±0.010	0.845±0.019	0.939±0.008	0.245±0.010	0.791±0.020	**0.921**±0.019	0.194±0.009	0.851±0.020	0.941±0.011
❄ Llama3.1-8B	0.201±0.014	0.931±0.007	0.964±0.006	0.352±0.022	0.711±0.020	0.884±0.009	0.407±0.020	0.683±0.016	0.831±0.007	0.337±0.014	0.744±0.012	0.890±0.006
♣ Llama3.1-8B	0.192±0.014	0.935±0.007	0.972±0.006	0.343±0.022	0.709±0.020	0.890±0.009	0.338±0.020	0.701±0.016	0.845±0.007	0.315±0.014	0.748±0.012	0.895±0.006

As shown in Table 9 and Table 10, CombPert maintains relatively low standard errors across most scenarios, particularly in the Seen 2/2 and Seen 1/2 splits. These results indicate low uncertainty

Table 10: Performance on the Replogle et al. dataset (mean ± standard error). ❄ denotes inference with CombPert (Reasoning Elicitation); ♣ denotes a model fine-tuned on CombPert-generated CoT data (Reasoning Synthesis). Best results are **bolded**, and second-best results are underlined.

Model / Method	Seen 2/2			Seen 1/2			Seen 0/2			Overall		
	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑
Mean	0.206±0.009	0.000±0.000	0.520±0.032	0.211±0.009	0.000±0.000	0.488±0.024	0.196±0.011	0.000±0.000	0.483±0.025	0.208±0.006	0.000±0.000	0.495±0.017
KG-RAG	0.037±0.017	0.956±0.014	0.900±0.000	0.125±0.029	0.761±0.047	0.829±0.031	0.749±0.018	0.440±0.016	0.417±0.014	0.196±0.068	0.762±0.050	0.785±0.046
Representation-based Methods												
Linear	0.060±0.002	0.823±0.014	0.980±0.004	0.090±0.005	0.774±0.024	0.863±0.023	0.078±0.004	0.480±0.043	0.633±0.035	0.082±0.003	0.742±0.016	0.858±0.015
CPA	0.145±0.017	0.535±0.038	0.720±0.020	0.121±0.010	0.576±0.026	0.685±0.017	0.209±0.013	0.147±0.027	0.533±0.020	0.141±0.008	0.501±0.019	0.671±0.012
GEARS	0.125±0.011	0.367±0.029	0.510±0.017	**0.074**±0.002	0.598±0.025	0.563±0.019	**0.061**±0.001	0.583±0.012	0.500±0.015	0.085±0.003	0.538±0.016	0.540±0.012
LLM-enhanced Methods												
Scouter	0.023±0.001	0.845±0.020	0.980±0.002	0.085±0.006	0.554±0.034	0.788±0.015	0.169±0.014	0.161±0.034	0.567±0.025	0.082±0.004	0.567±0.021	0.805±0.009
scLAMBDA	0.048±0.002	0.848±0.020	0.980±0.004	0.082±0.005	0.759±0.024	0.875±0.015	0.112±0.009	0.411±0.012	0.700±0.008	0.077±0.003	0.729±0.016	0.875±0.009
GPT-5.4	0.026±0.001	0.958±0.002	0.980±0.001	0.177±0.021	0.364±0.050	0.787±0.024	0.533±0.032	-0.492±0.014	0.383±0.014	0.193±0.013	0.386±0.030	0.775±0.015
GPT-5.4-mini	0.156±0.007	0.490±0.017	0.590±0.016	0.206±0.009	0.409±0.019	0.563±0.016	0.361±0.021	0.460±0.020	0.533±0.029	0.217±0.013	0.437±0.019	0.555±0.019
Qwen3-8B	0.299±0.005	0.127±0.018	0.530±0.011	0.342±0.013	0.157±0.020	0.538±0.013	0.945±0.056	0.182±0.022	0.467±0.027	0.422±0.011	0.153±0.013	0.525±0.009
Qwen3-32B	0.189±0.008	0.462±0.012	0.580±0.020	0.259±0.007	0.456±0.024	0.617±0.019	0.457±0.015	0.412±0.017	0.667±0.017	0.271±0.012	0.451±0.021	0.615±0.019
Llama3.1-8B	0.428±0.011	0.079±0.024	0.580±0.017	0.464±0.010	0.070±0.015	0.538±0.014	1.010±0.059	0.210±0.013	0.450±0.021	0.536±0.032	0.093±0.018	0.535±0.017
CombPert (Our Methods)												
❄ GPT-5.4	**0.012**±0.001	**0.960**±0.003	**1.000**±0.000	0.078±0.005	**0.778**±0.020	0.886±0.009	0.176±0.004	**0.588**±0.007	**0.790**±0.011	**0.076**±0.003	**0.795**±0.012	**0.900**±0.005
❄ GPT-5.4-mini	0.033±0.004	0.768±0.031	0.980±0.004	0.085±0.006	0.738±0.020	**0.908**±0.008	0.215±0.009	0.546±0.006	0.717±0.008	0.091±0.004	0.717±0.014	0.898±0.005
❄ Qwen3-8B	0.102±0.004	0.692±0.039	0.960±0.006	0.124±0.009	0.621±0.029	0.850±0.011	0.742±0.017	0.330±0.011	0.633±0.009	0.211±0.006	0.597±0.020	0.845±0.007
♣ Qwen3-8B	0.091±0.004	0.704±0.039	0.975±0.006	0.108±0.009	0.637±0.029	0.855±0.011	0.681±0.017	0.340±0.011	0.618±0.009	0.190±0.006	0.609±0.020	0.849±0.007
❄ Qwen3-32B	0.148±0.004	0.695±0.041	0.970±0.006	0.139±0.009	0.603±0.029	0.879±0.012	0.391±0.013	0.426±0.009	0.683±0.006	0.179±0.006	0.598±0.020	0.873±0.007
♣ Qwen3-32B	0.138±0.004	0.687±0.041	0.971±0.006	0.136±0.009	0.584±0.029	0.881±0.012	0.315±0.013	0.452±0.009	0.696±0.006	0.164±0.006	0.590±0.020	0.876±0.007
❄ Llama3.1-8B	0.115±0.007	0.704±0.044	0.930±0.012	0.118±0.006	0.586±0.030	0.758±0.019	0.380±0.016	0.375±0.018	0.683±0.002	0.157±0.005	0.584±0.021	0.790±0.011
♣ Llama3.1-8B	0.106±0.007	0.708±0.044	0.938±0.012	0.109±0.006	0.584±0.030	0.764±0.019	0.311±0.016	0.393±0.018	0.697±0.002	0.139±0.005	0.586±0.021	0.797±0.011

in the estimated mean performance of CombPert under both Reasoning Elicitation and Reasoning Synthesis.

B.6 Cost and Latency Analysis

We measured inference latency and estimated cost using the OpenRouter API[2]. Input prompts generally contained no more than 5,000 tokens across all models. For consistent cost comparisons, we assumed 5,000 input tokens per query. Table 11 summarizes the estimated cost and average latency per query across model backbones.

Table 11: Model Cost and Latency(s).

Model	Input Price ($/1M tokens)	Output Price ($/1M tokens)	Input Tokens	Output Tokens	Latency (s)	Cost ($/query)
GPT-5.4	2.50	15.00	5000	2000	1.02	4.25×10^{-2}
GPT-5.4-mini	0.75	4.50	5000	1300	0.83	9.60×10^{-3}
Qwen3-32B	0.08	0.24	5000	2000	0.55	8.80×10^{-4}
Qwen3-8B	0.05	0.40	5000	2000	0.76	1.05×10^{-3}
Llama-3.1-8B	0.02	0.05	5000	1300	0.36	1.70×10^{-4}

B.7 Evaluation of Mechanistic Explanation Quality and Faithfulness

To assess the quality and reliability of the generated explanations of our model, we conduct a human evaluation with three independent evaluators with biological backgrounds. We randomly sample 100 test queries and evaluate each explanation along two dimensions: *biological plausibility* and *evidence consistency*, both using a 3-point scale.

Biological Plausibility. This metric evaluates whether the explanation provides a biologically reasonable mechanism connecting the perturbation genes to the target gene:

- **1**: Biologically implausible or contains major mechanistic errors.
- **2**: Partially plausible but contains incomplete or questionable mechanistic claims.
- **3**: Biologically plausible and provides a coherent mechanistic interpretation.

Evidence Consistency. This metric evaluates whether the explanation is consistent with the provided biological evidence, including mechanistic paths, similar-gene evidence, and statistical priors:

- **1**: Substantially contradicts or ignores the provided evidence.

[2] https://openrouter.ai/

- **2**: Partially supported but contains unsupported or inconsistent claims.
- **3**: Well aligned with the provided evidence without major contradictions.

To assess the reliability of CombPert-generated explanations, we compare them with those from direct few-shot prompting using the same GPT-5.4 backbone. Results in Table 12 show that CombPert consistently achieves higher scores on both dimensions, indicating that its biologically guided reasoning produces plausible and evidence-grounded explanations.

Table 12: Human evaluation of mechanistic explanation quality.

Method	Evaluator	Plausibility↑	Evidence Consistency↑
Few-shot prompting	E1	2.10	1.92
	E2	2.18	1.98
	E3	2.14	1.95
	Average	2.14	1.95
CombPert	E1	2.52	2.43
	E2	2.58	2.49
	E3	2.55	2.46
	Average	**2.55**	**2.46**

B.8 Synthetic Data Utility Comparsion

To assess whether biological evidence improves the utility of synthesized CoT data, we compare data generated by CombPert and direct few-shot prompting. Data generation and filtering follow the protocol in Appendix A.3. We fine-tune Qwen3-8B and Qwen3-32B on the resulting datasets under matched settings and evaluate them on held-out perturbations in the Norman et al. dataset. As shown in Table 13, fine-tuning on CombPert-generated data outperforms fine-tuning on few-shot-generated data across both backbones and all metrics. Relative to training-free CombPert, it improves all three metrics for Qwen3-8B and reduces MSE while improving ACC for Qwen3-32B, with a slight decrease in PCC. These results demonstrate the utility of CombPert-generated CoT data for improving prediction on held-out perturbations.

Table 13: Overall comparison of synthetic data utility on the Norman et al. dataset. Reasoning Synthesis reports student performance after fine-tuning on data generated by GPT-5.4 using each method. Training-free results are included for reference. Best results for each backbone are **bolded**.

Method	Qwen3-8B			Qwen3-32B		
	MSE↓	PCC↑	ACC↑	MSE↓	PCC↑	ACC↑
Training-free Inference						
Few-shot prompting	0.643	0.142	0.584	0.571	0.288	0.606
CombPert	0.241	0.817	0.926	0.207	**0.857**	0.937
Reasoning Synthesis						
Few-shot prompting	0.562	0.354	0.781	0.394	0.663	0.805
CombPert	**0.216**	**0.832**	**0.928**	**0.194**	0.851	**0.941**

B.9 Full Gene Panel Evaluation

To assess the generalizability and robustness of CombPert beyond the top 20 DEGs, we evaluate it across the full measured gene panel on the Norman et al. dataset. For comparison, we include CPA, GEARS, direct few-shot prompting with GPT-5.4-mini, and CombPert with the same GPT-5.4-mini backbone. As shown in Table 14, CombPert consistently achieves the highest PCC across all evaluated settings, indicating that its predictive advantage extends beyond highly responsive genes to the full measured gene panel.

Table 14: Prediction performance across the full gene panel on the Norman et al. dataset, measured by PCC (↑) of expression changes relative to control cells. Best results are **bolded**.

Method	Seen 2/2	Seen 1/2	Seen 0/2	Overall
CPA	0.563 ± 0.000	0.583 ± 0.001	0.520 ± 0.002	0.571 ± 0.001
GEARS	0.649 ± 0.002	0.592 ± 0.001	0.549 ± 0.002	0.593 ± 0.001
GPT-5.4-mini	0.793 ± 0.001	0.562 ± 0.002	0.215 ± 0.000	0.575 ± 0.001
CombPert (GPT-5.4-mini)	$\mathbf{0.894 \pm 0.001}$	$\mathbf{0.714 \pm 0.002}$	$\mathbf{0.550 \pm 0.002}$	$\mathbf{0.711 \pm 0.001}$

C Case Study

To demonstrate the biological validity and interpretability of CombPert, we analyze two challenging combinatorial perturbation scenarios from the Norman et al. and Replogle et al. datasets. Both cases exemplify how CombPert transcends black-box representation learning methods by providing traceable, mechanistic hypotheses that align with established literature. The complete generated reasoning steps for both cases are detailed in Figure 10 and Figure 11.

Case 1. In the Norman et al. dataset, we examine the simultaneous overexpression of *CBL* and *UBASH3A*. The generated response reveals how CombPert leverages the biological knowledge graph to construct a coherent signaling axis. The model correctly identifies *CBL* as an E3 ubiquitin ligase that negatively regulates receptor tyrosine kinase (*RTK*) signaling, particularly the *EGFR* pathway (Thien & Langdon, 2001). Similarly, it recognizes *UBASH3A* as an inhibitory signaling adaptor. Notably, CombPert successfully identifies the direct protein-protein interaction between *CBL* and *UBASH3A* (Feshchenko et al., 2004; Vukojević et al., 2024), inferring that both perturbations converge on the shared *EGFR→E2F1* inhibitory circuitry. Instead of naively summing the inferred single effects, the model reasons about genetic interactions, noting that the functional overlap between *CBL* and *UBASH3A* suggests pathway redundancy. Consequently, the model predicts a sub-additive combinatorial effect, correctly inferring a moderate down-regulation of the downstream target *TEX30* . This reasoning perfectly captures the biological reality that simultaneous overexpression of redundant pathway suppressors yields diminishing returns due to signaling saturation.

Case 2. To further test zero-shot combinatorial generalization, we analyze a Seen 0/2 perturbation from the Replogle dataset. This case includes the simultaneous CRISPRi knockdown of *HMGCR* and *SREBF2*. The generated reasoning demonstrates CombPert's profound understanding of metabolic feedback loops. The model accurately identifies *HMGCR* as the rate-limiting enzyme of the mevalonate/cholesterol biosynthesis pathway (Goldstein & Brown, 1990) and *SREBF2* as the master sterol-responsive transcription factor (Horton et al., 2002). It correctly maps the regulatory circuit where *SREBF2* directly activates the downstream target *FDFT1* (squalene synthase). Crucially, CombPert exhibits advanced reasoning regarding genetic epistasis. It infers that knocking down *HMGCR* alone would typically reduce sterol synthesis and trigger a compensatory up-regulation of cholesterol biosynthesis genes via SREBP signaling, a homeostatic feedback mechanism (Longo et al., 2019). However, because the master regulator *SREBF2* is simultaneously knocked down, the model deduces that this compensatory response is blunted or abolished. Therefore, CombPert correctly predicts that the *SREBF2* knockdown dominates the interaction, overriding the *HMGCR*-induced compensation and leading to an overall down-regulation of *FDFT1*. This sophisticated deduction mirrors the established biological reality of sterol homeostasis, proving that CombPert can accurately resolve conflicting regulatory signals and infer non-additive epistatic interactions.

D Limitations

Our method relies on external biological knowledge resources, whose incompleteness or noise may affect evidence retrieval and prediction quality. In addition, retrieval-augmented reasoning introduces extra computational cost compared with purely representation learing models. Our evaluation is mainly conducted on CRISPR perturbation datasets in K562 cells, and further studies are needed to verify its generalization across cell types, perturbation technologies, and broader biological contexts.

Generated Response for Case "*CBL + UBASHA* ➡ *TEX30*" from Norman et al.

Gene A: *CBL*; **Gene B:** *UBASH3A*; **Target Gene:** *TEX30*; **Ground Truth:** -0.546
Baseline Prediction: GEARS (-0.368), CPA (-0.265), scLAMBDA (-0.207), Scouter (-0.204)
Response:
Step 1 — Gene Function & Pathway Context
CBL is an E3 ubiquitin ligase and classic negative regulator of receptor tyrosine kinase signaling, especially EGFR/MET-type pathways. Overexpressing CBL would generally be expected to dampen signaling outputs downstream of those receptors. UBASH3A is also broadly inhibitory/adaptor-like in signaling, especially immune/TCR-related contexts, and functionally resembles UBASH3B/TULA-family negative regulators. So both perturbed genes lean toward attenuating signaling rather than activating transcriptional output. TEX30 is poorly characterized, with only generic hydrolase-related annotation here. Its baseline expression is moderate so there is room for either direction, but the provided conditional prior strongly indicates TEX30 tends to go down under similar perturbations. Moderate expression also does not argue for strong induction; if anything, the empirical prior suggests this target is usually slightly to moderately repressed.

Relevant paths repeatedly route through E2F1 and FOXA1: CBL → EGR1 / MET / EGFR → E2F1 or FOXA1 → TEX30, UBASH3A → TP53BP2 / UBC / EGFR / CBLB → E2F1 or FOXA1 → TEX30. Given CBL inhibits EGFR and MET, and EGFR is listed as activating E2F1, CBL overexpression plausibly reduces an EGFR→E2F1→TEX30 axis. The FOXA1 links are weaker/unknown-sign, but analog data with FOXA1 itself gives negative TEX30 DE, so that branch also does not support strong upregulation.

Step 2 — Independent Effects
CBL alone: no direct training data, so infer from function and analogs. The strongest mechanistic clue is: CBL inhibits EGFR, EGFR activates E2F1, E2F1 connects to TEX30. If E2F1 tends to support TEX30-associated expression, then more CBL should reduce TEX30. The CBL analog set is dominated by negative outcomes on TEX30, including: FOXA1 single: about -0.29 to -0.32, PTPN13 single: -0.2638, SLC4A1 single: -0.3988, combo analogs like SGK1/TGFBR2 around -0.40 to -0.58. This suggests CBL alone is likely mildly to moderately negative, roughly around -0.25 to -0.40. I'd center single near -0.32.
UBASH3A alone: best analog is UBASH3B, same family, with: UBASH3B single → TEX30 = -0.2131. Also PTPN13 and ZBTB25 analogs are negative (~ -0.26). UBASH3A paths also feed into TP53BP2→E2F1 and UBC/HLA-DQA1→TEX30, with no compelling positive evidence. So UBASH3A alone also likely mildly negative, perhaps around -0.22 to -0.30.

Step 3 — Interaction Topology
CBL and UBASH3A directly interact by PPI, and both connect to EGFR and CBLB before reaching TEX30. This argues they are not independent random perturbations but partially overlapping negative signaling regulators.Important convergence: CBL inhibits EGFR, UBASH3A interacts with EGFR, EGFR activates E2F1, E2F1 links to TEX30.This supports a shared pathway where CBL overexpression suppresses EGFR-driven output and UBASH3A may reinforce inhibitory/adaptor control around the same node. The direct CBL–UBASH3A interaction suggests partial pathway redundancy rather than strong multiplicative synergy. So the combo should likely be more negative than either single, but not dramatically more negative than a simple combined expectation for two related suppressors.The CBLB/FOXA1 convergence also supports shared routing into TEX30-associated regulation, again pointing to coherent negative pressure rather than antagonism.

Step 4 — Quantitative Integration
The target-specific prior is very informative: calibrated mean -0.5274. That prior is for functionally similar perturbations and should heavily anchor the estimate. Since both genes independently look mildly negative and they converge on shared inhibitory signaling circuitry, the combo should likely land near the prior mean or slightly below it.A rough mental additive expectation from estimated singles (-0.32 and -0.24) would be around -0.56 before accounting for redundancy. Because CBL and UBASH3A are pathway-related and interact physically, some redundancy should compress the magnitude a bit, keeping the combo from becoming too extreme. That suggests a final value around -0.48 to -0.58. Given UBASH3B single = -0.2131, UBASH3B + PTPN9 combo = -0.5236. A CBL + UBASH3A combo being close to -0.52 seems very plausible.

Step 5 — Final Prediction
Predicted DE for TEX30 under CBL + UBASH3A overexpression: a moderate down-regulation near the calibrated prior mean, slightly reinforced by the shared EGFR/E2F1 inhibitory topology. PREDICTION: -0.54

Figure 10: Generated Response for Case from Norman et al. Dataset.

Generated Response for Case "*HMGCR + SREBF2 ➡ FDFT1*" from Replogle et al.

Gene A: *HMGCR*; **Gene B:** *SREBF2*; **Target Gene:** *FDFT1*; **Ground Truth:** -0.307
Baseline Prediction: GEARS (0.000), CPA (-0.188), scLAMBDA (-0.015), Scouter (-0.076)
Response:
Step 1 — Gene Function & Pathway Context
HMGCR is the rate-limiting enzyme of the mevalonate/cholesterol biosynthesis pathway. Knocking it down reduces mevalonate production and cholesterol synthesis flux. SREBF2 is a master sterol-responsive transcription factor that activates many cholesterol biosynthetic genes. FDFT1 is explicitly a cholesterol-pathway enzyme, and the KG includes a direct activating edge SREBF2 → FDFT1, plus indirect activation via SREBF1 and NFYA/NR1H3-related convergence. FDFT1 encodes squalene synthase, downstream in the mevalonate pathway and typically part of the sterol-regulated transcriptional program. Since baseline expression is moderate, both moderate up- and down-regulation are plausible, with no severe ceiling constraint.

Step 2 — Independent Effects
SREBF2 knockdown. Strongest evidence is direct: SREBF2 —[activates]→ FDFT1. Also indirect support through SREBF1, which activates FDFT1. So CRISPRi knockdown of SREBF2 should decrease FDFT1. This is the clearest effect in the query. HMGCR knockdown: The KG paths are mostly of the form "HMGCR —[activated by]→ SREBF1/SREBF2," meaning upstream sterol regulators activate HMGCR, not necessarily HMGCR activating FDFT1. Biologically, lowering HMGCR reduces sterol synthesis, which often triggers compensatory activation of SREBP-mediated cholesterol biosynthesis genes. That would tend to increase FDFT1 expression. This matches the conditional prior: similar perturbations show FDFT1 almost always up-regulated, mean around +0.21. Thus HMGCR knockdown alone is likely mildly positive for FDFT1, probably around +0.15 to +0.25. So estimated singles: HMGCR KD → FDFT1: mild up-regulation; SREBF2 KD → FDFT1: moderate down-regulation.

Step 3 — Interaction Topology
HMGCR and SREBF2 are tightly linked in the same cholesterol homeostasis circuit: SREBF2 activates HMGCR, both connect to SREBF1 and other sterol regulators, both converge on FDFT1 through SREBF1, NFYA, and NR1H3 This suggests non-independence. The key biological point is that the likely compensatory up-regulation caused by HMGCR knockdown is mediated largely through SREBP signaling, especially SREBF2. If SREBF2 is simultaneously knocked down, that compensatory response is blunted or abolished. Therefore the double perturbation should not be additive in the sense of "HMGCR-up plus SREBF2-down"; instead SREBF2 knockdown should dominate and suppress the HMGCR-induced compensation. That implies a negative effect overall, likely somewhat more negative than SREBF2 KD alone or at least not rescued by HMGCR KD.

Step 4 — Quantitative Integration
The prior for similar perturbations is positive (~+0.21), reflecting perturbations of cholesterol-pathway enzymes that induce compensatory FDFT1 up-regulation. That prior is useful mainly for HMGCR-like metabolic knockdown, but here one perturbation is the key transcriptional activator SREBF2 itself. Since SREBF2 directly activates FDFT1, this prior should be overridden downward. A reasonable decomposition: HMGCR KD alone: about +0.20, SREBF2 KD alone: about -0.35 to -0.50 because of direct TF loss on a pathway target Double KD: HMGCR compensation largely blocked by SREBF2 loss, so net effect remains negative, perhaps around -0.35, not extremely large because FDFT1 also has additional regulators. Given the direct SREBF2→FDFT1 edge and shared-pathway suppression of compensation, a moderate negative value is most plausible.

Step 5 — Final Prediction
PREDICTION: -0.350

Figure 11: Generated Response for Case from Replogle et al. Dataset.