LEARNING ADAPTIVE PERTURBATION-CONDITIONED CONTEXTS FOR ROBUST TRANSCRIPTIONAL RESPONSE PREDICTION

Yinhua Piao*
AI Co-Research & Education for
innovative Drug Institute, KAIST

Hyomin Kim
Kim Jaechul Graduate School of
Artificial Intelligence, KAIST

Seonghwan Kim
AI Co-Research & Education for
innovative Drug Institute, KAIST

Yunhak Oh
Graduate School of Data Science
KAIST

Junhyeok Jeon & Sang-Yeon Hwang & Jaechang Lim
HITS
Seoul, Republic of Korea

Woo Youn Kim
Department of Chemistry, KAIST
HITS

Chanyoung Park
Department of Industrial and
Systems Engineering, KAIST

Sungsoo Ahn †
Kim Jaechul Graduate School of
Artificial Intelligence, KAIST

ABSTRACT

Predicting high-dimensional transcriptional responses to genetic perturbations is challenging due to severe experimental noise and sparse gene-level effects. Existing methods often suffer from *mean collapse*, where high correlation is achieved by predicting global average expression rather than perturbation-specific responses, leading to many false positives and limited biological interpretability. Recent approaches incorporate biological knowledge graphs into perturbation models, but these graphs are typically treated as dense and static, which can propagate noise and overlook true perturbation signals. We propose ADAPERT, a perturbation-conditioned framework that addresses mean collapse by explicitly modeling sparsity and biological structure. ADAPERT learns perturbation-specific subgraphs from biological knowledge graphs and applies adaptive learning to separate true signals from noise. Across multiple genetic perturbation benchmarks, ADAPERT consistently outperforms existing baselines and achieves substantial improvements on DEG-aware evaluation metrics, indicating more accurate recovery of perturbation-specific transcriptional changes.

1 INTRODUCTION

Predicting how cells respond to genetic perturbations is a key problem in functional genomics (Shalem et al., 2015; Przybyla & Gilbert, 2022). It supports many downstream tasks, such as understanding gene function, analyzing regulatory effects, and identifying potential therapeutic targets. Recent progress in single-cell perturbation experiments, including Perturb-seq and CRISPR-based screens (Datlinger et al., 2017; Bock et al., 2022), now allows gene expression to be measured across thousands of genes under many perturbation conditions (Dixit et al., 2016; Norman et al., 2019; Replogle et al., 2022). As a result, there is growing interest in computational models that can predict transcriptional responses to perturbations that have not been experimentally tested.

*yinhua.piao@kaist.ac.kr
†corresponding author: sungsoo.ahn@kaist.ac.kr

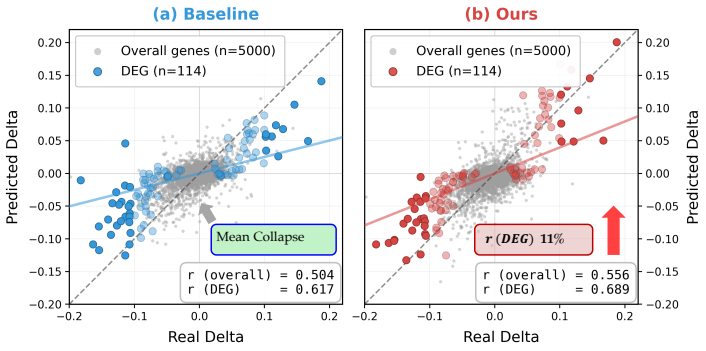


Figure 1: Mean-collapse as a common failure mode in perturbation modeling. For the UQCRB perturbation ($n = 114$ DEGs), a standard perturbation model shows mean-collapse, where predicted expression changes shrink toward zero and large effects are underestimated (Left). Our method reduces this bias by using perturbation-specific context and better tracks gene-level expression changes, especially for strongly expressed genes (Right).

A central challenge in this task is that single-cell measurements are inherently noisy, making it difficult to learn perturbation-specific effects from the samples (Brennecke et al., 2013). Recent efforts have addressed this challenge primarily through input-level improvements: scaling up training data or input shape (Cui et al., 2024), enriching features with biological annotations (Chen & Zou, 2024), and incorporating knowledge graphs to improve generalization to unseen perturbations (Roohani et al., 2024). While these directions have shown progress, we observe that a fundamental failure mode persists across many existing methods.

Instead of capturing perturbation-specific changes, models tend to predict expression shifts close to the global average, a behavior we call **mean-collapse**. As illustrated in Figure 1(a), baseline model (Wenkel et al., 2025) collapses predictions toward the center of the distribution, where non-DEG genes dominate (*Gray* dots). This can produce high overall correlation, but expression changes of biologically important genes (*Red* and *Blue* dots) are strongly underestimated, which aligns with the recent findings (Mejia et al., 2025). As a result, these models yield many false positives and provide limited insight into the true effects of perturbations.

We argue that mean-collapse arises not from insufficient data or features, but from a mismatch between model design and the sparse nature of perturbation responses. For each perturbation, only *a small subset of genes* shows strong changes, and these genes are typically related to the perturbed gene through biological pathways (Wu et al., 2009). A useful perturbation model should therefore meet two requirements. First, it should isolate true signals from noise under high variability. Second, it should use biological structure to guide this separation. Most existing methods treat knowledge graphs as dense and static (Wenkel et al., 2025; He et al., 2025), using them for global embedding rather than selecting perturbation-relevant genes. In noisy settings, this can spread the signal across many unrelated genes and increase false positives.

Based on these observations, we propose AdaPert, a perturbation-conditioned method that addresses mean-collapse by modeling sparsity and biological structure. Rather than treating knowledge graphs as fixed templates, AdaPert learns a perturbation-conditioned context for each perturbation. Starting from control cells, the method selects genes related to the perturbed gene and forms a compact subgraph from the full graph. An adaptive learning scheme then limits variation in non-responsive genes and uses responsive genes to refine the subgraph representation. This design enables robust modeling of perturbation-specific transcriptional changes under noisy settings.

We evaluate AdaPert on multiple genetic perturbation benchmarks. Across datasets, AdaPert consistently outperforms existing methods on multiple metrics, with the largest gains on DEG-aware metrics, indicating better modeling of perturbation-specific transcriptional changes. We further provide a comprehensive analysis across perturbations with different effect-size. These results show that learning adaptive perturbation-conditioned context on biological knowledge graphs improves perturbation prediction in noisy settings.

2 RELATED WORKS

2.1 DATA-DRIVEN GENETIC PERTURBATION MODELING

Genetic perturbation modeling has been extensively studied through data-driven learning approaches, which aim to reconstruct gene expression responses under perturbation conditions (Lopez et al., 2018; Lotfollahi et al., 2019; Bunne et al., 2023; Lotfollahi et al., 2023; Cui et al., 2024; Hao et al., 2024; Adduri et al., 2025). Most existing methods formulate this task as an end-to-end prediction problem, optimizing objectives such as mean squared error or correlation between predicted and observed expression profiles. These approaches have demonstrated strong performance on standard quantitative metrics and are widely adopted as baselines for perturbation prediction tasks. However, by primarily focusing on overall reconstruction accuracy, they do not explicitly model which genes are causally or specifically affected by a given perturbation, motivating the exploration of additional sources of inductive bias (Wenteler et al., 2024; Wu et al., 2024; Li et al., 2024).

2.2 KNOWLEDGE-DRIVEN GENETIC PERTURBATION MODELING

To address the limitations of purely data-driven approaches, recent work has explored incorporating prior knowledge to provide structural (Wenkel et al., 2025; Roohani et al., 2024; He et al., 2025) or semantic constraints (Cui et al., 2024; Chen & Zou, 2024; Istrate et al., 2024). Such priors aim to guide models toward biologically plausible solutions, improve robustness under noisy single-cell measurements, and better capture perturbation-specific regulatory effects. Existing approaches leverage structured biological resources, including curated networks and textual knowledge, but the integration of these priors is often static and global.

Biological knowledge graphs, such as protein–protein interaction networks and pathway databases, have been widely used to encode relationships among genes. In perturbation modeling, these graphs are typically incorporated to generate gene embeddings, constrain message passing, or regularize model parameters (Wenkel et al., 2025; Roohani et al., 2024; He et al., 2025). By propagating information along known biological interactions, graph-based methods introduce inductive biases that reflect prior biological knowledge. However, most existing approaches treat knowledge graphs as dense and static structures that are shared across all perturbations. As a result, the same global graph is applied regardless of the specific perturbation, without explicitly identifying which substructures are relevant to a given perturbation condition.

More recently, large language models (LLMs) have been explored for extracting biological knowledge from unstructured sources, including scientific literature and curated databases (Chen & Zou, 2024; Istrate et al., 2024). In biological applications, LLMs have been applied to gene annotation, relationship scoring, and semantic retrieval (Wu et al.; Istrate et al., 2025; He et al., 2025). These models provide a complementary source of prior knowledge that is difficult to encode in structured graphs alone. However, most LLM-based approaches do not directly model perturbation-response data and instead rely on inferred associations or reasoning over textual knowledge. Consequently, LLMs are often better suited as auxiliary components that provide prior guidance, rather than as standalone models for predicting perturbation-induced transcriptional responses.

3 METHODOLOGY

3.1 PROBLEM DEFINITION AND PRELIMINARIES

We formalize the task of predicting transcriptional responses to genetic perturbations within a conditional generative framework. Let $\mathbf{X}^c \in \mathbb{R}^N$ denote the gene expression profile of a control cell c, where N is the number of observed genes. A perturbation targeting a specific gene p is selected from the gene set $\mathcal{K}$. Our objective is to learn a predictive mapping $\mathcal{F} : (\mathbf{X}^c, p) \to \hat{\mathbf{X}}^p$, where $\hat{\mathbf{X}}^p \in \mathbb{R}^N$ is the predicted expression profile post-perturbation. Existing state-of-the-art methods typically implement $\mathcal{F}$ using a conditional autoencoder backbone consisting of three functional components: a control encoder, a condition encoder, and a perturbation decoder. An encoder $\texttt{ENC}_\theta$ maps the control profile into a lower-dimensional latent representation $\mathbf{z}_c$, capturing the baseline state:

$$\mathbf{z}_c = \texttt{ENC}_\theta(\mathbf{X}^c), \quad \mathbf{z}_c \in \mathbb{R}^d \tag{1}$$

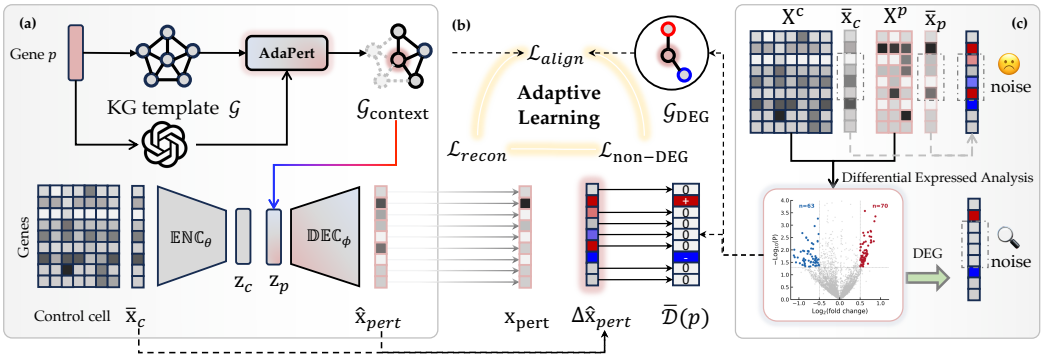


Figure 2: Overview of AdaPert. (a) The model takes a control cell expression profile $\bar{\mathbf{x}}_c$ and a perturbation gene p as input. A perturbation-conditioned subgraph $\mathcal{G}_{\text{context}}$ is extracted from a biological knowledge graph template $\mathcal{G}$, producing a context representation $\mathbf{z}_p$ that is combined with the encoded control state $\mathbf{z}_c$ to predict the perturbed expression profile $\hat{\mathbf{x}}_{\text{pert}}$ via encoder ENC$_\theta$ and decoder DEC$_\phi$. (b) The adaptive learning scheme separates signal from noise using three loss terms: a reconstruction loss $\mathcal{L}_{\text{recon}}$ for overall expression fidelity, a non-DEG loss $\mathcal{L}_{\text{non-DEG}}$ that suppresses spurious changes in non-responsive genes, and an alignment loss $\mathcal{L}_{\text{align}}$ that guides the learned subgraph representation $\mathcal{G}_{\text{DEG}}$ to encode perturbation-specific differential expression patterns. (c) Illustration of single-cell perturbation data structure showing control cells $\mathbf{X}^c$, perturbed cells $\mathbf{X}^p$, and their mean profiles $\bar{\mathbf{x}}_c$ and $\bar{\mathbf{x}}_p$. Differentially expressed genes (DEGs) are identified through statistical testing, distinguishing true perturbation signals from experimental noise.

The perturbed gene p (condition) is represented by an embedding $\mathbf{z}_p \in \mathbb{R}^d$. $\mathbf{z}_p$ is a learnable vector initialized by one-hot encoding or from recent gene foundation models (Cui et al., 2024; Theodoris et al., 2023). In more recent knowledge-guided models (Wenkel et al., 2025), it is derived from a global biological knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ via a graph neural network (Veličković et al., 2018):

$$\mathbf{z}_p = \texttt{GNN}(\mathcal{G}, p) \tag{2}$$

where nodes $\mathcal{V}$ represent genes and edges $\mathcal{E}$ represent functional interactions. Finally, a decoder $\texttt{DEC}_\phi$ reconstructs the perturbed transcriptional response by integrating the cell state $\mathbf{z}_c$ and the perturbation signal $\mathbf{z}_p$:

$$\hat{\mathbf{X}}^p = \texttt{DEC}_\phi(\mathbf{z}_c, \mathbf{z}_p) \tag{3}$$

The model is trained by minimizing a reconstruction loss $\mathcal{L}(\mathbf{X}^p, \hat{\mathbf{X}}^p)$, defined as the mean squared error between the predicted and observed gene expression profiles.

However, this objective treats all genes equally. In genetic perturbation data, true responses are sparse, with only a small subset of genes showing strong changes while most genes remain near baseline. Let $\mathcal{D}(p)$ denote the set of responsive genes (DEGs) under perturbation p, and $\bar{\mathcal{D}}(p)$ its complement, with $|\mathcal{D}(p)| \ll |\bar{\mathcal{D}}(p)|$. The reconstruction loss can be decomposed as

$$\mathcal{L}_{\text{rec}} = \underbrace{\sum_{i \in \mathcal{D}(p)} \left(\hat{\mathbf{X}}_i^p - \mathbf{X}_i^p\right)^2}_{\text{responsive genes}} + \underbrace{\sum_{i \in \bar{\mathcal{D}}(p)} \left(\hat{\mathbf{X}}_i^p - \mathbf{X}_i^p\right)^2}_{\text{non-responsive genes}}. \tag{4}$$

Since the second term dominates, minimizing mean squared error is driven mainly by non-responsive genes, encouraging predictions to shrink toward zero. As a result, large perturbation effects are systematically underestimated, leading to a failure mode we refer to as **mean-collapse**.

3.2 OVERVIEW OF ADAPERT

To address the mean-collapse induced by dense reconstruction objectives, we propose ADAPERT, a perturbation-conditioned framework (fig. 2) that explicitly models sparsity of signals and biological structure related to the perturbation. The model consists of two components. **First**, ADAPERT extracts a *perturbation-conditioned subgraph* from a unified biological knowledge graph template. Rather than using the full graph as a static prior, this module selects a compact subgraph that captures genes biologically related to the perturbed gene, providing a structured hypothesis space for perturbation response modeling. **Second**, ADAPERT employs an *adaptive learning* scheme to separate the true signal from noise. This module constrains spurious variations in non-differentially expressed genes while leveraging differentially expressed genes to guide the alignment and refinement of subgraph representations. Together, these two components enable perturbation-specific modeling that reduces noise propagation and preserves sparse transcriptional responses with high fidelity.

3.3 PERTURBATION-CONDITIONED SUBGRAPH EXTRACTION

We extract a perturbation-specific subgraph from a unified biological knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Instead of propagating messages over the full graph, our approach selects a sparse set of *perturbation-relevant* nodes to construct a subgraph centered around the perturbed gene. This design integrates semantic information beyond graph structure and reduces overfitting by restricting message passing to a small, condition-dependent context.

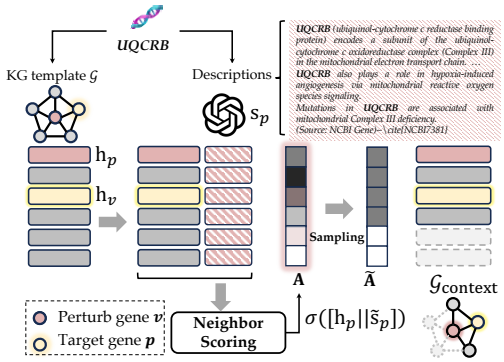


Figure 3: Perturbation-Conditioned Subgraph Extraction. Given a perturbed gene p (e.g., UQCRB), the module extracts a perturbation-specific subgraph from the knowledge graph template $\mathcal{G}$. First, a textual description of the perturbed gene is retrieved from NCBI and encoded using a language model to obtain a semantic embedding $\mathbf{s}_p$. Each node v in the graph is represented by a structural embedding $\mathbf{h}_v$ computed via message passing. Differentiable Gumbel-Softmax sampling is then applied to select a sparse set of perturbation-relevant nodes, yielding a compact subgraph $\mathcal{G}_{\text{context}}$ centered around genes related to the perturbed gene.

Node Representations. Each gene node $v \in \mathcal{V}$ is represented by a structural embedding that captures graph topology. We initialize each node with a one-hot vector $\mathbf{x}_v$ and apply message passing:

$$\mathbf{h}_v^{(0)} = \mathbf{x}_v, \tag{5}$$

$$\mathbf{h}_v^{(l+1)} = \sum_{u \in \mathcal{N}(v)} \frac{1}{|\mathcal{N}(v)|} \mathbf{W}^{(l)} \mathbf{h}_u^{(l)}, \tag{6}$$

where $\mathcal{N}(v)$ denotes the neighbors of v and $\mathbf{W}^{(l)}$ are learnable weights. After L layers, we obtain the structural embedding $\mathbf{h}_v = \mathbf{h}_v^{(L)} \in \mathbb{R}^{d_s}$.

Perturbation Semantic Embedding. We use language model embeddings to provide a basic semantic understanding of each perturbation gene. By encoding textual descriptions of the perturbed gene, the language model captures information that is not explicitly represented in the knowledge graph, such as gene family membership, functional similarity, and naming-related associations. This semantic information complements graph structure and enables the model to identify relevant genes that may be weakly connected or disconnected in the graph. For each perturbation gene p, we retrieve its textual description from NCBI and encode it using a language model (`GPT-4o` (OpenAI, 2024)):

$$\mathbf{s}_p = \text{LM}(\text{desc}(p)), \quad \mathbf{s}_p \in \mathbb{R}^{d_t}. \tag{7}$$

To align semantic and structural spaces, we project the perturbation embedding into the graph embedding space:

$$\tilde{\mathbf{s}}_p = \mathbf{W}_s \mathbf{s}_p, \tag{8}$$

where $\mathbf{W}_s \in \mathbb{R}^{d_s \times d_t}$.

Perturbation-Conditioned Node Scoring. To identify nodes relevant to a given perturbation, we condition node selection on the perturbation embedding. For each node v, we construct a joint representation by concatenating its structural embedding with the perturbation embedding:

$$\mathbf{c}_v = [\mathbf{h}_v \,\|\, \tilde{\mathbf{s}}_p] \,. \tag{9}$$

A perturbation-conditioned node relevance score is computed via a multilayer perceptron:

$$a_v = \mathbf{w}^\top \sigma(\mathbf{W}_c \mathbf{c}_v) \,, \tag{10}$$

where $\sigma(\cdot)$ denotes a non-linear activation. Scores are normalized across all nodes:

$$\alpha_v = \frac{\exp(a_v)}{\sum_{u \in \mathcal{V}} \exp(a_u)}. \tag{11}$$

Differentiable Node Sampling. To enforce sparsity while preserving differentiability, we apply Gumbel-Softmax sampling (Jang et al., 2017) over node scores:

$$\tilde{\alpha}_v = \frac{\exp\left((\log \alpha_v + g_v)/\tau\right)}{\sum_{u \in \mathcal{V}} \exp\left((\log \alpha_u + g_u)/\tau\right)}, \tag{12}$$

where $g_v \sim \mathrm{Gumbel}(0, 1)$ and τ is a temperature parameter. Nodes with $\tilde{\alpha}_v > T$ are selected, yielding a perturbation-specific node-induced subgraph $\mathcal{G}_p$.

Perturbation Context Representation. Finally, we summarize the selected subgraph by aggregating the embeddings of the selected nodes:

$$\mathbf{z}_{\text{context}} = \sum_{v \in \mathcal{V}(\mathcal{G}_p)} \mathbf{h}_v. \tag{13}$$

Because node selection is explicitly conditioned on the perturbation, different perturbations induce distinct subgraphs, allowing the model to focus on causally relevant genes while filtering out unrelated graph structure.

3.4 Adaptive Learning for Signal–Noise Separation

We explicitly separate signal and noise during training by leveraging perturbation-specific differential expression information. For each perturbation p, let $\mathbf{X}^c, \mathbf{X}^p \in \mathbb{R}^N$ denote the control and perturbed expression profiles, and define the perturbation effect $\Delta\mathbf{X}^p = \mathbf{X}^p - \mathbf{X}^c$. Using the training data, we perform a statistical test for each gene and obtain a p-value $q_i^{(p)}$. We define the DEG and non-DEG sets as

$$\mathcal{D}(p) = \{i \mid q_i^{(p)} < 0.05\}, \quad \bar{\mathcal{D}}(p) = \{i \mid q_i^{(p)} \geq 0.05\}. \tag{14}$$

Rather than treating all genes equally, we introduce three complementary loss terms with distinct roles: (i) a global reconstruction loss to preserve overall expression fidelity, (ii) a robust penalty that suppresses spurious changes on non-DEG genes, and (iii) a response-aware alignment loss that encourages the extracted subgraph representation to encode perturbation-specific DEG signals. Together, these objectives promote explicit separation between signal and noise during training.

Global reconstruction loss. We first match the full perturbed expression profile using a mean squared error:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}\left[\|\hat{\mathbf{X}}^p - \mathbf{X}^p\|_2^2\right]. \tag{15}$$

This term ensures global consistency and stabilizes optimization, but alone is insufficient to distinguish true perturbation effects from noisy fluctuations.

Non-DEG robust loss. For non-responsive genes $\bar{\mathcal{D}}(p)$, the expected perturbation change is close to zero, while experimental measurements can be noisy. To reduce spurious deviations without being overly sensitive to outliers, we penalize predicted perturbation changes on $\bar{\mathcal{D}}(p)$ using a Huber loss (Huber, 1992):

$$\mathcal{L}_{\text{non}} = \mathbb{E}_p \left[\sum_{i \in \bar{\mathcal{D}}(p)} \rho_\delta \left(\Delta \hat{\mathbf{X}}_i^p \right) \right], \qquad \Delta \hat{\mathbf{X}}^p = \hat{\mathbf{X}}^p - \mathbf{X}^c. \tag{16}$$

The Huber penalty is defined as

$$\rho_\delta(r) = \begin{cases} \frac{1}{2} r^2, & |r| \le \delta, \\ \delta \left(|r| - \frac{1}{2}\delta \right), & |r| > \delta. \end{cases} \tag{17}$$

The threshold δ controls the transition between quadratic and linear penalties, allowing small residuals to be strongly suppressed while preventing large but noisy deviations from dominating the loss. In practice, δ is set proportional to the empirical standard deviation of non-DEG effects, and is fixed across perturbations.

Adaptive subgraph representation alignment. Beyond expression-level supervision, we explicitly guide the learned subgraph representation to reflect perturbation-specific responses. Let $\mathbf{z}_{\text{context}}^{(p)}$ denote the context representation produced by the extracted subgraph for perturbation p (Section 3.3). We construct a response-driven target by summarizing DEG signals:

$$\mathbf{y}^{(p)} \in \mathbb{R}^N, \qquad \mathbf{y}_i^{(p)} = \begin{cases} \Delta \mathbf{X}_i^p, & i \in \mathcal{D}(p), \\ 0, & i \in \bar{\mathcal{D}}(p), \end{cases} \tag{18}$$

which preserves signed effect sizes while masking non-DEG genes. This vector is mapped into the representation space via a projection head $g(\cdot)$:

$$\mathbf{t}^{(p)} = g\left(\mathbf{y}^{(p)} \right) \in \mathbb{R}^d. \tag{19}$$

We then align the subgraph context with the response-driven target using a cosine-distance loss:

$$\mathcal{L}_{\text{align}} = \mathbb{E} \left[\left\| \frac{\mathbf{z}_{\text{context}}^{(p)}}{\|\mathbf{z}_{\text{context}}^{(p)}\|_2} - \frac{\mathbf{t}^{(p)}}{\|\mathbf{t}^{(p)}\|_2} \right\|_2^2 \right]. \tag{20}$$

This alignment encourages the extracted subgraph to encode perturbation-relevant DEG structure, rather than generic graph features.

Overall objective. The final training objective combines all three terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{non}} \mathcal{L}_{\text{non}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}. \tag{21}$$

4 Experiments

4.1 Experiment Setup

We evaluate AdaPert on a single-cell genetic perturbation prediction task. The goal is to predict the transcriptional response of cells after a target gene is perturbed. All experiments are conducted under the *unseen perturbation* setting, where perturbations in the test set are not observed during training. We use two single-cell CRISPR perturbation datasets from Replogle *et al.* (Replogle et al., 2022) The first dataset, *K562.Replogle*, consists of single-gene knockouts measured by single-cell RNA sequencing in the K562 cell line. The second dataset, *RPE1.Replogle*, is collected using the same experimental protocol in the RPE1 cell line. Both datasets include control cells and perturbed cells for each target gene, enabling direct evaluation of perturbation-induced transcriptional changes. For each dataset, we follow standard preprocessing and data splitting protocols used in prior work. Training, validation, and test sets are constructed such that perturbations in the test set are entirely unseen during training. More details about the dataset are provided in the Appendix.

Table 1: Performance comparison across perturbation datasets. Results are reported as mean ± std over all test perturbations. Δ denotes correlation on differential expression relative to control, and PDS denotes perturbation discriminative score.

Category	Method	K562.Replogle		RPE1.Replogle	
		Pearson-Δ	PDS	Pearson-Δ	PDS
w/o KG	scVI (Lopez et al., 2018)	$0.171_{\pm 0.195}$	$0.502_{\pm 0.290}$	$0.369_{\pm 0.175}$	$0.501_{\pm 0.289}$
	CPA (Lotfollahi et al., 2023)	$0.288_{\pm 0.157}$	$0.503_{\pm 0.290}$	$0.486_{\pm 0.210}$	$0.503_{\pm 0.290}$
	STATE (Adduri et al., 2025)	$0.247_{\pm 0.188}$	$0.506_{\pm 0.289}$	$0.583_{\pm 0.281}$	$0.520_{\pm 0.292}$
w/ KG	GEARS (Roohani et al., 2024)	$0.298_{\pm 0.181}$	$0.529_{\pm 0.287}$	$0.396_{\pm 0.216}$	$0.524_{\pm 0.298}$
	TxPert (Wenkel et al., 2025)	$0.580_{\pm 0.255}$	$0.665_{\pm 0.310}$	$0.655_{\pm 0.300}$	$0.618_{\pm 0.290}$
	MorPH (He et al., 2025)	$0.442_{\pm 0.242}$	$0.664_{\pm 0.299}$	$0.541_{\pm 0.299}$	$\mathbf{0.688}_{\pm \mathbf{0.268}}$
	AdaPert	$\mathbf{0.619}_{\pm \mathbf{0.262}}$	$\mathbf{0.711}_{\pm \mathbf{0.296}}$	$\mathbf{0.674}_{\pm \mathbf{0.291}}$	$0.663_{\pm 0.281}$

Table 2: DEG-aware evaluation on *K562.Replogle*. Results are reported as mean ± std over all test perturbation genes.

Category	Method	Differential Expressed Score		DE-metrics		
		@50	@100	Spearman-sig	Spearman-lfc-sig	Direction-match
w/o KG	scVI	$0.071_{\pm 0.072}$	$0.064_{\pm 0.060}$	0.443	$0.372_{\pm 0.355}$	$0.577_{\pm 0.186}$
	CPA	$0.116_{\pm 0.098}$	$0.104_{\pm 0.083}$	0.474	$0.360_{\pm 0.104}$	$0.645_{\pm 0.104}$
	STATE	$0.044_{\pm 0.041}$	$0.050_{\pm 0.045}$	0.468	$0.298_{\pm 0.198}$	$0.643_{\pm 0.108}$
w/ KG	GEARS	$0.126_{\pm 0.107}$	$0.124_{\pm 0.106}$	0.540	$0.381_{\pm 0.341}$	$0.603_{\pm 0.177}$
	TxPert	$0.220_{\pm 0.173}$	$0.205_{\pm 0.165}$	0.536	$0.640_{\pm 0.241}$	$0.843_{\pm 0.133}$
	MorPH	$0.057_{\pm 0.083}$	$0.090_{\pm 0.097}$	0.546	$0.448_{\pm 0.239}$	$0.757_{\pm 0.128}$
	AdaPert	$\mathbf{0.263}_{\pm \mathbf{0.190}}$	$\mathbf{0.252}_{\pm \mathbf{0.182}}$	**0.622**	$\mathbf{0.688}_{\pm \mathbf{0.240}}$	$\mathbf{0.867}_{\pm \mathbf{0.141}}$

4.2 Baselines and Training Protocol

We compare AdaPert against two categories of baseline methods: (1) models without a knowledge graph, including scVI (Lopez et al., 2018), CPA (Lotfollahi et al., 2023), and STATE (Adduri et al., 2025); and (2) models that incorporate a knowledge graph, including GEARS (Roohani et al., 2024), TxPert (Wenkel et al., 2025), and MorPH (He et al., 2025). These baselines represent state-of-the-art approaches for genetic perturbation prediction. All models are trained under the same experimental setup and computational budget to ensure fair comparison. Unless otherwise specified, we use identical data splits, training procedures, and evaluation protocols across all methods. Additional implementation details are provided in the Appendix.

4.3 Evaluation Metrics

To evaluate perturbation prediction performance, we use a set of complementary metrics that capture different aspects of model behavior. We report two global metrics, Pearson-Δ and the Perturbation Discrimination Score (PDS), which measure overall agreement between predicted and observed perturbation effects. In addition, we include DEG-aware metrics that focus on differential expression accuracy, including Differential Expression Score@K (DES@K) and Spearman correlation of log fold changes and their directions. These metrics are sensitive to false positive predictions and better reflect the recovery of perturbation-specific gene responses. Together, this metric suite allows us to assess both global reconstruction accuracy and the ability to capture biologically meaningful perturbation effects. All metrics are computed using the latest version of the `cell-eval` (Adduri et al., 2025) evaluation framework.

5 MAIN RESULTS

5.1 GLOBAL PERFORMANCE OF PERTURBATION PREDICTION

We conduct genetic perturbation prediction on two CRISPR perturbation datasets, *K562.Replogle* and *RPE1.Replogle*. We report Pearson correlation on differential expression relative to the control (Pearson-Δ) and the perturbation discriminative score (PDS). PDS measures how well a model distinguishes different perturbations. As shown in Table 1, methods without biological knowledge graphs show limited performance on both datasets. Their PDS values are close to chance level, indicating a weak ability to separate different perturbations. Methods that use biological knowledge graphs perform better, showing that the use of biological prior knowledge is important. However, these methods rely on dense and mostly static graph structures, which can still spread noise across genes. ADAPERT achieves the best performance on both datasets. The gains are consistent for Pearson-Δ and PDS. Importantly, the improvement on PDS is larger than that on Pearson-Δ. This shows that ADAPERT improves perturbation discrimination, rather than only increasing global correlation. The results suggest that ADAPERT reduces mean-collapsed predictions and focuses on perturbation-specific transcriptional changes.

5.2 DEG-AWARE COMPARISONS

As shown in Table 2, we report the Differential Expression Score (DES@K), which measures how well true DEGs are ranked among top predicted genes. Methods without knowledge graphs achieve very low DES, indicating poor separation between signal and noise. KG-based methods improve DEG recovery. GEARS shows moderate gains, and TxPert further improves DES, but its scores remain limited, suggesting that noise still affects non-DEG genes. ADAPERT achieves the highest DES at both $k = 50$ and $k = 100$, with consistent improvements over TxPert. This indicates more accurate ranking of true DEGs. ADAPERT also performs better on DEG-specific metrics, including Spearman correlation, the agreement of log-fold changes, and direction consistency. These results show that ADAPERT produces sparse and reliable perturbation-specific gene responses.

5.3 COMPARISON ON MEAN-COLLAPSE

We analyze model sensitivity to mean-collapse by grouping perturbations into small-, medium-, and large-effect sets based on ground-truth effect size. Results are shown in Table 3. For small-effect perturbations, where mean-collapse is most severe, TxPert and ADAPERT achieve similar Pearson-Δ, but ADAPERT shows much higher DES. This shows better separation of true signal from noise, despite similar overall correlation. For medium- and large- effect perturbations, ADAPERT improves all metrics, including Pearson-Δ, DES, and PDS. Overall, these results show that ADAPERT is more robust to mean-collapse, especially when true perturbation effects are weak.

Table 3: Sensitivity to mean bias under different perturbation effect sizes on *K562.Replogle*.

Effect size	Model	P-Δ	DES	PDS
Small-effect ($<5\%$)	TxPert	0.457	0.090	0.740
	ADAPERT	0.462	0.115	0.737
	Improvement	$\uparrow$ 1.1%	$\uparrow$ 28%	$\downarrow$ 0.4%
Medium-effect (5%–10%)	TxPert	0.541	0.173	0.663
	ADAPERT	0.623	0.222	0.740
	Improvement	$\uparrow$ 15%	$\uparrow$ 28%	$\uparrow$ 12%
Large-effect ($>10\%$)	TxPert	0.741	0.353	0.592
	ADAPERT	0.771	0.420	0.657
	Improvement	$\uparrow$ 4.0%	$\uparrow$ 19%	$\uparrow$ 11%

5.4 EFFECT OF $\mathcal{L}_{\text{NON}}$ AND $\mathbf{z}_{\text{CONTEXT}}$ ACROSS PERTURBATIONS

We conduct an ablation study on three perturbation groups categorized by effect size to evaluate the roles of perturbation-specific context $\mathbf{z}_{\text{context}}$ and the non-DEG loss $\mathcal{L}_{\text{non}}$ (Figure 4). Overall, the full model performs well across all metrics and effect-size regimes, with the strongest performance observed for small and medium perturbations. Removing $\mathbf{z}_{\text{context}}$ consistently degrades performance across all comparisons, showing that enriching perturbation context is critical. Removing the $\mathcal{L}_{\text{non}}$

has a strong negative impact for small and medium perturbations, where signals are weak, and noise is high, indicating that adaptive separation of signal and noise is necessary in this regime. For large perturbations ($> 10\%$ DE genes), the effect of the $\mathcal{L}_{\text{non}}$ becomes less pronounced, showing that the balance between signal and noise varies with perturbation effect size.

5.5 Effect-Size–Dependent Behavior of the $\mathcal{L}_{\text{non}}$

We analyze the interaction between $\mathcal{L}_{\text{non}}$ and perturbation effect size by varying the weight λ_{non} of the non-DEG loss across different perturbation groups (Figure 5). For small and medium perturbations, performance consistently improves as λ_{non} increases from $0 \rightarrow 0.01$ across both global and DEG-based metrics. This trend indicates that assigning more weight to the non-DEG loss helps suppress noise and improve signal recovery when perturbation effects are weak. In contrast, for large perturbations, smaller values of λ_{non} yield better performance, suggesting that strong signals require less regularization. Across all perturbation groups, setting λ_{non} too large (e.g., $\lambda = 0.1$) leads to clear performance degradation. This behavior suggests that excessive smoothing over-suppresses perturbation signals and harms both gene-level recovery and global reconstruction.

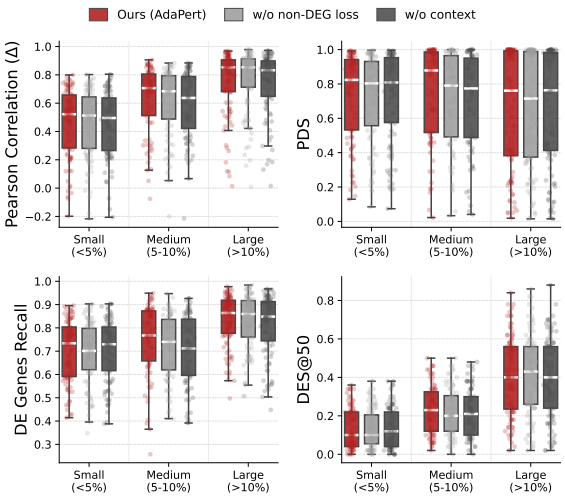


Figure 4: Comparison of model variants across small, medium, and large perturbations.

5.6 Pathway-level Validation of Predicted Profiles

To validate our model at the pathway level, we performed Gene Set Enrichment Analysis (GSEA) on predicted differential expression profiles for HIRA knockdown and compared them to experimental ground truth (Figure 9). The predicted pathway enrichment scores showed significant positive correlation with ground truth, demonstrating that our model captures coordinated pathway-level responses. Notably, the model accurately predicted the downregulation of cell cycle-related pathways, including Myc Targets V1 and E2F Targets, consistent with HIRA's known role in chromatin regulation. These results indicate that our perturbation prediction model preserves biologically meaningful pathway signatures beyond individual gene-level accuracy.

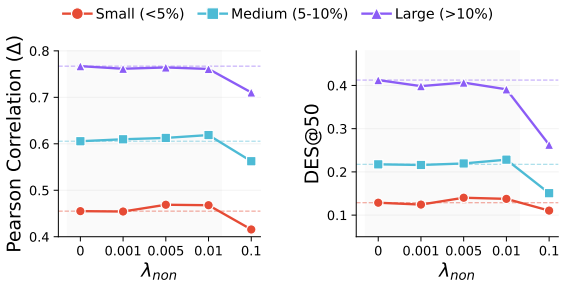


Figure 5: Sensitivity of model performance to λ_{non} across perturbation effect sizes.

6 Conclusion

We identify mean-collapse as a key failure mode in perturbation prediction and propose AdaPert to address it through perturbation-specific context modeling and adaptive signal–noise separation. AdaPert consistently improves performance across benchmarks, especially for perturbations with small effects, and reveals the effect-size–dependent role of regularization. These results highlight the importance of adaptive modeling for robust perturbation prediction.

ACKNOWLEDGMENTS

This work was supported by the InnoCORE program of the Ministry of Science and ICT (N10250153).

REFERENCES

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, et al. Predicting cellular responses to perturbation across diverse contexts with state. *BioRxiv*, pp. 2025–06, 2025.

Christoph Bock, Paul Datlinger, Florence Chardon, Matthew A Coelho, Matthew B Dong, Keith A Lawson, Tian Lu, Laetitia Maroc, Thomas M Norman, Bicna Song, et al. High-content crispr screening. *Nature Reviews Methods Primers*, 2(1):8, 2022.

Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 10(11):1093–1095, 2013.

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.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Yanshuo Chen, Zhengmian Hu, Wei Chen, and Heng Huang. Fast and scalable wasserstein-1 neural optimal transport solver for single-cell perturbation prediction. *Bioinformatics*, 41 (Supplement_1):i513–i522, 2025.

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

M Chevalley, Y Roohani, A Mehrjou, J Leskovec, and P Schwab. Causalbench: A large-scale benchmark for network inference from single-cell perturbation data. arxiv, 2022.

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.

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

Advait Dixit, Oren Parnas, Bing Li, Jernej Chen, Charles Fulco, and et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7): 1853–1866, 2016.

Mingze Dong, Bao Wang, Jessica Wei, Antonio H de O. Fonseca, Curtis J Perry, Alexander Frey, Feriel Ouerghi, Ellen F Foxman, Jeffrey J Ishizuka, Rahul M Dhodapkar, et al. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nature methods*, 20(11): 1769–1779, 2023.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Claudia Feng, Elin Madli Peets, Yan Zhou, Luca Crepaldi, Sunay Usluer, Alistair Dunham, Jana M Braunger, Jing Su, Magdalena E Strauss, Daniele Muraro, et al. A genome-scale single cell crispri map of trans gene regulation across human pluripotent stem cell lines. *bioRxiv*, pp. 2024–11, 2024.

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.

Chujun He, Jiaqi Zhang, Munther Dahleh, and Caroline Uhler. Morph predicts the single-cell outcome of genetic perturbations across conditions and data modalities. *bioRxiv*, 2025.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Ana-Maria Istrate, Donghui Li, and Theofanis Karaletsos. scgenept: Is language all you need for modeling single-cell perturbations? *bioRxiv*, pp. 2024–10, 2024.

Ana-Maria Istrate, Fausto Milletari, Fabrizio Castrotorres, Jakub M Tomczak, Michaela Torkar, Donghui Li, and Theofanis Karaletsos. rbio1-training scientific reasoning llms with biological world models as soft verifiers. *bioRxiv*, pp. 2025–08, 2025.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. *arXiv preprint arXiv:1611.01144*, 2017.

Lanxiang Li, Yue You, Yunlin Fu, Wenyu Liao, Xueying Fan, Shihong Lu, Ye Cao, Bo Li, Wenle Ren, Jiaming Kong, et al. A systematic comparison of single-cell perturbation response prediction models. *bioRxiv*, pp. 2024–12, 2024.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

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

Mohammad Lotfollahi et al. Predicting cellular responses to genetic perturbations using scrna-seq data. *Nature Biotechnology*, 41:1234–1245, 2023.

Gabriel M Mejia, Henry E Miller, Francis JA Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Diversity by design: Addressing mode collapse improves scrna-seq perturbation modeling on well-calibrated metrics. *arXiv preprint arXiv:2506.22641*, 2025.

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.

OpenAI. Gpt-4o system card. `https://openai.com/research/gpt-4o-system-card`, 2024.

James D Pearce, Sara E Simmonds, Gita Mahmoudabadi, Lakshmi Krishnan, Giovanni Palla, Ana-Maria Istrate, Alexander Tarashansky, Benjamin Nelson, Omar Valenzuela, Donghui Li, et al. A cross-species generative cell atlas across 1.5 billion years of evolution: The transcriptformer single-cell model. *bioRxiv*, pp. 2025–04, 2025.

Laralynne Przybyla and Luke A Gilbert. A new era in functional genomics screens. *Nature Reviews Genetics*, 23(2):89–103, 2022.

Joseph M Replogle, Trevor M Norman, Angela Xu, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(15):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.

Yanay Rosen, Yusuf Roohani, Ayush Agarwal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pp. 2023–11, 2023.

Ophir Shalem, Neville E Sanjana, and Feng Zhang. High-throughput functional genomics using crispr–cas9. *Nature Reviews Genetics*, 16:299–311, 2015.

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

Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E Peluso, Qingyao Huang, Tao Fang, et al. The string database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, 2025.

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.

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

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

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Magdalena Huebner, Victor Curean, WT Dee, WT Connell, Alex Hawkins-Hooker, Siu Pui Chung, Yasha Ektefaie, et al. Perteval-scfm: benchmarking single-cell foundation models for perturbation effect prediction. *bioRxiv*, pp. 2024–10, 2024.

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

Michael C Wu, Lingsong Zhang, Zhaoxi Wang, David C Christiani, and Xihong Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.

Yan Wu, Esther Wershof, Sebastian M Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. *arXiv preprint arXiv:2408.10609*, 2024.

Table 4: Statistics of the single-cell perturbation datasets in K562 and RPE1 cell lines from *Replogle et al.* dataset (Replogle et al., 2022)) used in this work.

Dataset	Split	# Cells	# Perturbations	# HVGs
K562.Replogle	Train	111,770	734	5,000
	Val	10,918	82	
	Test	38,475	272	
RPE1.Replogle	Train	74,474	771	3,352
	Val	26,073	308	
	Test	50,593	464	

Table 5: Dataset statistics stratified by perturbation effect size. Effect size categories (Small, Medium, Large) are defined based on the percentage of differentially expressed genes: Small ($< 5\%$), Medium (5–10%), and Large ($> 10\%$).

Split	Effect Size	*K562.Replogle*		*RPE1.Replogle*	
		# Perts	# Cells	# Perts	# Cells
Train	Small	232	34,402	252	41,264
	Medium	251	31,008	259	17,859
	Large	250	35,669	260	15,351
Validation	Small	29	4,206	105	14,033
	Medium	27	3,595	104	7,014
	Large	26	3,117	99	5,026
Test	Small	98	16,003	152	31,405
	Medium	91	11,206	162	10,960
	Large	83	11,266	150	8,228

A Data Statistics

A.1 Single-Cell Genetic Perturbation Data

Predicting how cells respond to genetic perturbations is a key problem in functional genomics (Shalem et al., 2015). It supports many downstream tasks, such as understanding gene function, analyzing regulatory effects, and identifying potential therapeutic targets. Recent progress in single-cell perturbation experiments, including Perturb-seq and CRISPR-based screens (Bock et al., 2022), now allows gene expression to be measured across thousands of genes under many perturbation conditions (Dixit et al., 2016; Replogle et al., 2022; Lotfollahi et al., 2023). As a result, there is growing interest in computational models that can predict transcriptional responses to perturbations that have not been experimentally tested (Bunne et al., 2024; Chevalley et al., 2022).

We use publicly available single-cell CRISPR perturbation (Bock et al., 2022; Feng et al., 2024) datasets generated using Perturb-seq experiments (Dixit et al., 2016; Replogle et al., 2022; Lotfollahi et al., 2023). Gene expression profiles are measured under single-gene perturbations with matched control cells. Differential expression relative to controls is used as the prediction target. We evaluate on the *K562.Replogle* and *RPE1.Replogle* datasets (Replogle et al., 2022). Both datasets contain thousands of cells across hundreds of perturbations and span a wide range of perturbation effect sizes. Models are trained on highly variable genes (HVGs) only. Train, validation, and test splits are defined at the perturbation level, such that cells from held-out perturbations are excluded from training. Dataset statistics are summarized in Table 4, and effect-size–stratified statistics are reported in Table 5. Perturbations are grouped into small-, medium-, and large-effect categories based on the fraction of differentially expressed genes identified at a significance threshold of $p < 0.05$. This stratification reflects substantial heterogeneity in perturbation responses and enables a more fine-grained evaluation under both sparse and strong transcriptional effect regimes.

Extended analysis. *(1) Gene expression distributions in control and perturbed cells.* We compare log1p-transformed expression distributions between all genes and differentially expressed genes (DEGs) in K562 and RPE1 cells (Figure 6). Overall expression exhibits the expected right-skewed distributions in both K562 (mean = 0.54) and RPE1 (mean = 0.59). n DEGs are biased toward higher expression levels. In K562, DEGs (n = 3,015) show a modest shift in expression (mean = 0.64), whereas in RPE1, DEGs (n = 228) exhibit substantially higher expression (mean = 1.34) and a more symmetric distribution. Red and green dashed lines indicate the mean and median, respectively. *(2) Perturbation effect size and DEG distributions.* To describe variation in perturbation responses, we group perturbations into three categories: small, medium, and large. Effect size is defined as the fraction of differentially expressed genes with $p_{value} < 0.05$. Figure 7 shows how the number of DEGs changes with effect size in K562 and RPE1. In both cell lines, the number of DEGs increases with effect size. In K562, small-effect perturbations affect 65 genes on average (1.3% of 5,000 HVGs), and large-effect perturbations affect 482 genes (9.6%). In RPE1, DEG counts increase from 106 genes (3.1% of 3,352 HVGs) to 649 genes (19.4%). The histograms in Figure 7 (bottom) show clear separation between effect-size groups. Small- and large-effect perturbations show little overlap. Across all groups, RPE1 has higher DEG ratios than K562. This suggests that RPE1 cells are more sensitive to genetic perturbations. This difference matters for evaluation, where large-effect perturbations affect many genes, while small-effect perturbations affect few genes. These two cases require different prediction behavior.

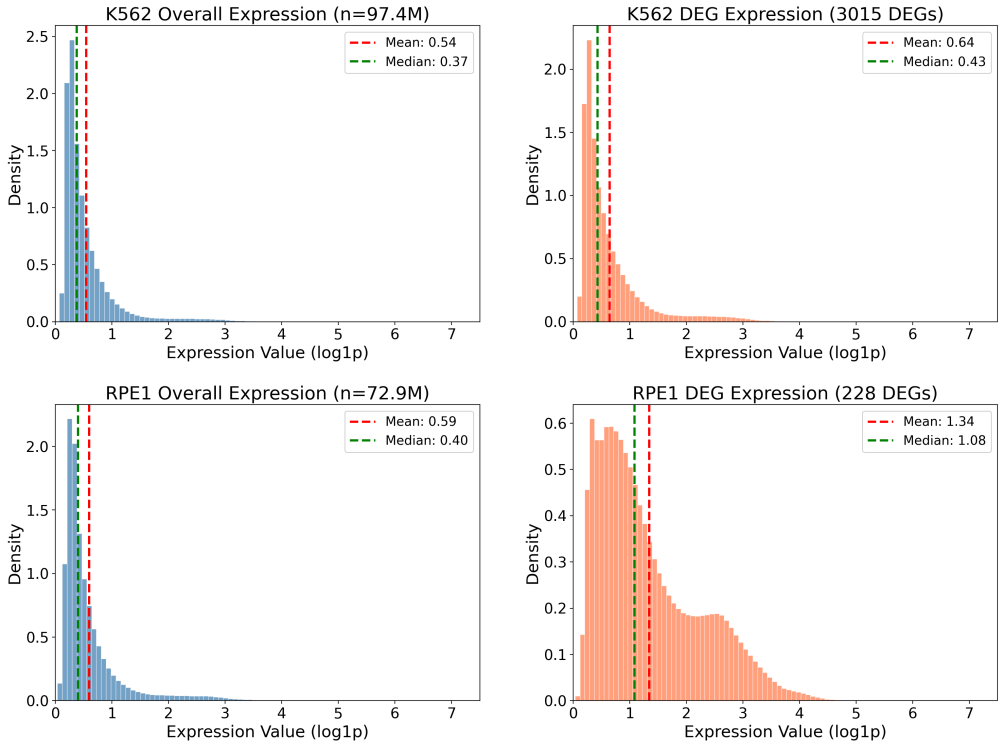


Figure 6: **Expression distribution of overall genes and differentially expressed genes in K562 and RPE1 cells.** Comparison of gene expression distributions between overall genes and differentially expressed genes (DEGs) in two cell lines. (A-B) K562 cells show right-skewed overall expression (mean=0.54, median=0.37) with 3,015 DEGs exhibiting slightly higher expression (mean=0.64, median=0.43). (C-D) RPE1 cells display similar overall expression patterns (mean=0.59, median=0.40), while 228 DEGs show markedly higher and more symmetric expression distribution (mean=1.34, median=1.08). Expression values are log1p-transformed. DEGs in RPE1 were computed from 50 perturbation conditions using top-20 genes ranked by absolute expression change.

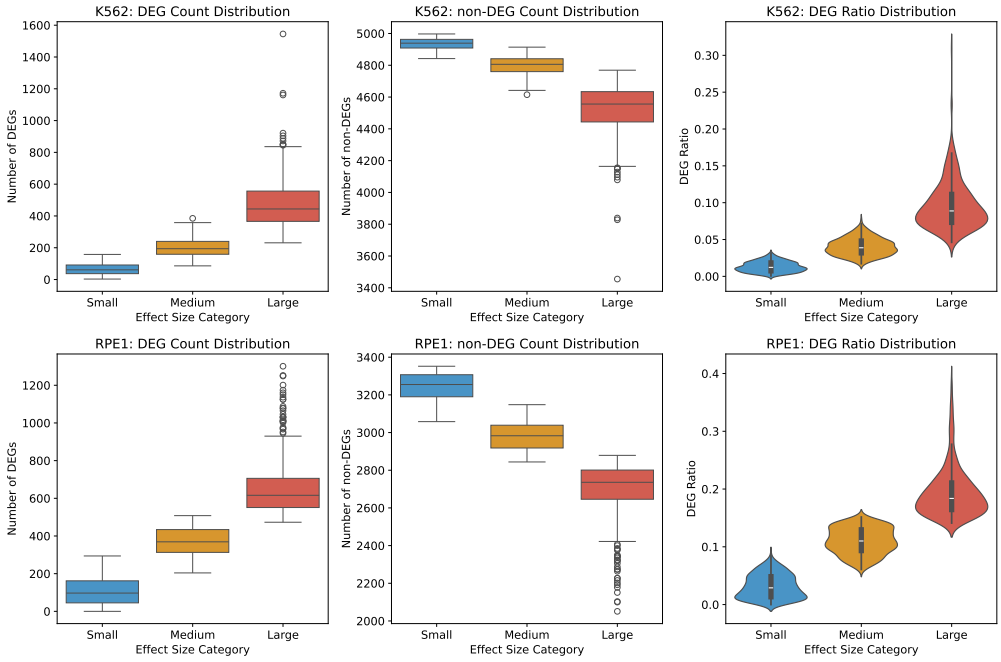


Figure 7: **Distribution of differentially expressed genes (DEGs) across perturbation effect size categories.** (Top row) Stacked bar plots showing the mean number of DEGs (red) and non-DEGs (blue) for each effect size category in K562 (left) and RPE1 (right) datasets. Numbers indicate the mean gene count per category. (Bottom row) Histograms showing the distribution of DEG counts across individual perturbations, colored by effect size category (Small: blue, Medium: orange, Large: red). Effect size categories are defined by tertiles of mean absolute differential expression. DEGs are identified using an absolute expression change threshold of 0.1.

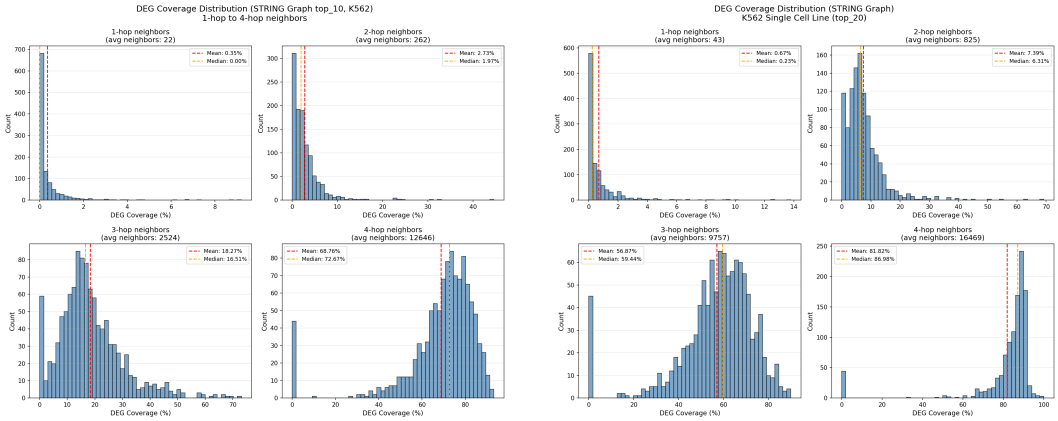
A.2 KNOWLEDGE GRAPH

We use a protein–protein interaction (PPI) knowledge graph constructed from STRING v11.5 (Szklarczyk et al., 2025; 2023). Nodes correspond to genes, and edges represent reported interactions between gene entities. The raw graph includes all interactions provided by STRING and is highly dense. To align the graph with the perturbation datasets, we restrict the graph to genes measured in the experiments. Specifically, we retain only highly variable genes (HVGs) used as model inputs. This step substantially reduces graph size while preserving genes relevant to perturbation modeling. After HVG filtering, the graph remains dense. To further control graph complexity, we apply top-k edge filtering, retaining only the k highest-confidence edges per gene. We consider $k = 10$ and $k = 20$. Graph statistics for the raw graph, the HVG-filtered graph, and the top-k graphs are reported in Table 6. Top-k filtering substantially reduces node degree, yielding much sparser graph structures. This motivates learning perturbation-conditioned subgraphs from localized graph neighborhoods, rather than operating on the full dense graph.

Table 6: Statistics of the STRING knowledge graph. The raw graph contains all protein–protein interactions from STRING v11.5. HVG-filtered graphs are restricted to highly variable genes used in experiments. Top-k variants retain the k highest-confidence edges per gene.

Graph	# Nodes	# Edges	Avg. Degree	Med. Degree
Raw (Full STRING)	18,382	11,257,696	1,224.9	970
HVG	4,509	1,090,554	483.7	386
HVG + Top-20	4,509	89,793	39.8	36
HVG + Top-10	4,509	45,013	20.0	18

DEG coverage in the knowledge graph. We assess whether the knowledge graph captures perturbation-relevant genes by measuring DEG coverage in graph proximity to the perturbed gene. For each perturbation in the test set, we compute the fraction of true DEGs reachable within a small number of hops (e.g., 1–3) from the perturbed gene node. As shown in Figure 8, a large fraction of DEGs lie close to the perturbed gene in the graph. This supports the use of local graph context for perturbation modeling.



(a) Top-10 predicted genes (b) Top-20 predicted genes

Figure 8: DEG coverage as a function of graph hop distance for different prediction depths. (a) Top-10 predicted genes. (b) Top-20 predicted genes.

A.3 Gene Descriptions

We use GenePT (Chen & Zou, 2024) embeddings derived from NCBI and UniProt gene descriptions encoded via OpenAI's text embedding models (Dunefsky et al., 2024). The embeddings are available in two variants: Ada (1,536-dim) and Model 3 (3,072-dim), covering 93,800 and 133,736 genes respectively. Coverage for our datasets is high: 95.8% for K562 and 98.6% for RPE1 HVGs.

Table 7: Statistics and HVG coverage of GenePT-based gene embeddings.

Statistic	**Ada Embedding**	**Model 3 Embedding**
Total genes	93,800	133,736
Embedding dimension	1,536	3,072
K562 HVG coverage	4,790 / 5,000 (95.8%)	4,790 / 5,000 (95.8%)
RPE1 HVG coverage	3,304 / 3,352 (98.6%)	3,304 / 3,352 (98.6%)

B Metric Definitions

Let $\mathbf{X}^c, \mathbf{X}^p \in \mathbb{R}^N$ denote the control and perturbed expression profiles, and let $\hat{\mathbf{X}}^p$ be the predicted perturbed profile. We define the true and predicted perturbation effects as

$$\Delta\mathbf{X}^p = \mathbf{X}^p - \mathbf{X}^c, \qquad \Delta\hat{\mathbf{X}}^p = \hat{\mathbf{X}}^p - \mathbf{X}^c. \tag{22}$$

B.1 Global metrics.

Pearson-Δ. We compute the Pearson correlation between the predicted and true perturbation effects:

$$\text{Pearson-}\Delta = \text{corr}\left(\Delta\hat{\mathbf{X}}^p, \Delta\mathbf{X}^p\right). \tag{23}$$

This metric measures global agreement in perturbation-induced expression changes.

Perturbation Discrimination Score (PDS). To evaluate whether predicted perturbation effects are specific to the correct perturbation, we use the Perturbation Discrimination Score (PDS) following the Virtual Cell Challenge. For each perturbation p, we compute the distance between its predicted effect $\Delta\hat{\mathbf{X}}^p$ and the true effects of all perturbations in the test set:

$$d_{p,t} = \left\| \Delta\hat{\mathbf{X}}^p - \Delta\mathbf{X}^t \right\|_1, \qquad \forall\, t \in \mathcal{T}, \tag{24}$$

where $\mathcal{T}$ denotes the set of test perturbations.

We rank these distances in ascending order and define the rank of the correct perturbation as

$$r_p = 1 + \sum_{t \neq p} \mathbb{I}[d_{p,t} < d_{p,p}]\,. \tag{25}$$

The discrimination score for perturbation p is then

$$\mathrm{PDS}_p = 1 - \frac{r_p - 1}{|\mathcal{T}|}. \tag{26}$$

The final PDS is obtained by averaging PDS_p over all perturbations in the test set. Higher values indicate better discrimination of perturbation-specific effects.

Let $\mathcal{D}(p)$ denote the set of differentially expressed genes for perturbation p, defined using the ground-truth data.

B.2 DEG-AWARE METRICS.

Differential Expression Score (DES). Following the Virtual Cell Challenge evaluation, we assess whether a model recovers the correct set of differentially expressed genes after perturbation. For each perturbation p, let $G_{\text{true}}(p)$ denote the ground-truth set of significant DEGs and $G_{\text{pred}}(p)$ the predicted set of significant DEGs, both defined at a fixed false discovery rate threshold.

The Differential Expression Score for perturbation p is defined as the fraction of true DEGs that are recovered in the predicted set:

$$\mathrm{DES}(p) = \frac{|G_{\text{true}}(p) \cap G_{\text{pred}}(p)|}{|G_{\text{true}}(p)|}. \tag{27}$$

The overall DES is obtained by averaging $\mathrm{DES}(p)$ over all perturbations in the test set. Higher values indicate better recovery of differentially expressed genes.

DE-Spearman (significant genes). We compute the Spearman rank correlation between predicted and true effects over DEGs:

$$\text{DE-Spearman-sig} = \rho_s(\Delta\hat{\mathbf{x}}^p_{\mathcal{D}}, \Delta\mathbf{x}^p_{\mathcal{D}})\,. \tag{28}$$

DE-Spearman (LFC-weighted). To emphasize genes with larger effect sizes, we compute a weighted Spearman correlation using absolute ground-truth effects as weights:

$$\text{DE-Spearman-lfc-sig} = \rho_s^{(w)}(\Delta\hat{\mathbf{x}}^p_{\mathcal{D}}, \Delta\mathbf{x}^p_{\mathcal{D}},\, |\Delta\mathbf{x}^p_{\mathcal{D}}|)\,. \tag{29}$$

DE Direction Match. We measure the fraction of DEGs for which the predicted and true effect directions agree:

$$\text{DE-Dir} = \frac{1}{|\mathcal{D}(p)|} \sum_{i \in \mathcal{D}(p)} \mathbb{I}\Big[\mathrm{sign}(\Delta\hat{\mathbf{X}}^p_i) = \mathrm{sign}(\Delta\mathbf{X}^p_i)\Big]\,. \tag{30}$$

C ADDITIONAL RELATED WORKS

C.1 DATA-DRIVEN AND GENERAL-PURPOSE MODELING APPROACHES

A broad class of prior work models transcriptional responses to perturbations primarily through *data-driven learning*, without explicitly encoding biological mechanisms. Early generative frameworks such as (Lopez et al., 2018; Lotfollahi et al., 2019) learn latent representations of gene expression and infer perturbation effects through shifts in latent space. Subsequent methods, including

(Lotfollahi et al., 2023; Adduri et al., 2025), extend this paradigm by conditioning latent variables on perturbation identities and cellular contexts.

Related to these approaches, several models formulate perturbation prediction as a *distributional mapping problem*. Optimal-transport–based methods such as (Bunne et al., 2023; Chen et al., 2025) and causal transport models like (Dong et al., 2023) aim to align control and perturbed cell populations at the distribution level. While effective at capturing global expression shifts, these methods are not explicitly designed to recover sparse gene-level effects.

More recently, large-scale *foundation models* have been introduced for single-cell biology, including (Cui et al., 2024; Hao et al., 2024; Theodoris et al., 2023; Rosen et al., 2023; Pearce et al., 2025). These models learn transferable gene or cell representations from massive datasets and are often used as pretrained encoders for downstream tasks. However, they do not explicitly model perturbation-specific sparsity or directionality, and their predictions may still be dominated by averaged transcriptional responses.

C.2 Knowledge-driven perturbation models

To address the limitations of purely data-driven approaches, a growing line of work incorporates *biological prior knowledge* into perturbation response modeling. Methods such as (Roohani et al., 2024; Wenkel et al., 2025) leverage gene–gene interaction networks or pathway graphs to propagate perturbation signals through known biological relationships, improving generalization to unseen perturbations.

Recent studies further explore the integration of *textual and semantic biological knowledge*. Approaches including (Chen & Zou, 2024; Istrate et al., 2024; Wu et al.; Istrate et al., 2025) use pretrained language models to construct gene representations from literature, functional annotations, or structured biological descriptions. These methods demonstrate that external knowledge can complement expression data, particularly in low-data or out-of-distribution settings.

However, most existing knowledge-driven models treat biological knowledge as static and globally shared across perturbations. Dense graphs or fixed embeddings are typically reused for all perturbations, which can propagate irrelevant interactions and obscure perturbation-specific signals. This static usage of knowledge limits the ability of models to adaptively focus on the most relevant biological substructures for a given genetic intervention.

In addition, prior knowledge is often integrated uniformly, without explicit mechanisms to separate true perturbation-induced signals from background transcriptional variation.

C.3 Positioning of this work

Our work builds on the knowledge-driven paradigm by introducing *perturbation-conditioned adaptation* in the use of biological knowledge. Rather than relying on static graphs or fixed embeddings, we learn sparse, perturbation-specific subgraphs that dynamically emphasize relevant biological interactions. This design complements prior data-driven and knowledge-based approaches and enables more accurate recovery of perturbation-specific transcriptional signals.

D Baselines Details

We compare AdaPert with a set of representative baselines for single-cell genetic perturbation modeling. These baselines differ in model design, conditioning strategy, and the use of biological prior knowledge.

D.1 Baselines without biological knowledge graphs.

scVI (Lopez et al., 2018) is a variational autoencoder for single-cell RNA-seq data that learns a latent representation of gene expression without explicit conditioning on perturbations. It models the distribution of expression counts using a probabilistic decoder and serves as a purely data-driven baseline for comparing latent generative approaches.

CPA (Lotfollahi et al., 2023) (Compositional Perturbation Autoencoder) learns disentangled latent representations of control and perturbed cells. It separates a cell's basal state from perturbation effects in the latent space, enabling prediction of unseen perturbations and combinations. CPA can also learn interpretable embeddings for cells and perturbations and supports out-of-distribution predictions by recombining learned latent factors.

STATE (Adduri et al., 2025) is a deep generative model designed to predict perturbation effects on single-cell expression by transforming latent representations in a structured space. It accounts for cellular heterogeneity and aims to capture complex, nonlinear responses across conditions.

D.2 Baselines with biological knowledge graphs.

GEARS (Roohani et al., 2024) integrates protein–protein interaction information into perturbation prediction by using graph-based message passing to propagate perturbation signals over network structure. This allows the model to leverage known gene interaction topology when predicting expression changes.

TxPert (Wenkel et al., 2025) uses graph representations of biological relationships to inform prediction of transcriptional responses under out-of-distribution settings. It conditions expression prediction on graph-based embeddings that capture biochemical relationships among genes, enabling generalization to unseen perturbations and cell contexts.

MORPH (He et al., 2025) combines a discrepancy-based variational autoencoder with an attention mechanism to predict cellular responses to unseen perturbations, including unseen single genes, perturbation combinations, and cell contexts. The attention mechanism enables the model to infer gene interactions and regulatory effects while learning latent perturbation representations.

All baselines are evaluated using their recommended settings and official implementations when available. We apply the same data splits, preprocessing, and evaluation protocols across all methods to ensure fair comparison.

E Training Details

Baseline reproduction. All baseline models are reproduced using their official implementations when available. For each method, we follow the training procedures described in the original papers, including model architecture, optimization strategy, and data preprocessing. When minor implementation choices are not specified, we adopt standard defaults used in the corresponding codebases.

To ensure a fair comparison, all models are trained and evaluated using the same train/validation/test splits, the same set of highly variable genes, and the same evaluation protocol. Results are reported on the held-out test set.

Hyperparameter tuning. For each baseline, we perform hyperparameter tuning over a predefined search space (Appendix Table 8). Hyperparameters are selected based on validation performance, using Pearson-Δ as the primary selection metric. The best-performing configuration on the validation set is then used for final evaluation on the test set. The same tuning protocol is applied consistently across all baselines. No test data are used during model selection.

F Additional Results on RPE1 dataset

As shown in Table 10, we report DEG-aware evaluation results on the RPE1 test set. We use the Differential Expression Score (DES@K) to measure how well true DEGs are ranked among top predicted genes. Data-driven methods such as scVI and CPA achieve low DES values, indicating limited ability to separate DEGs from non-DEGs. GEARS performs poorly on RPE1, suggesting that static graph propagation is insufficient for this dataset. TxPert improves DEG ranking and DEG-specific metrics, but its performance remains constrained, especially on DEG correlation and direction consistency. MORPH shows moderate gains, but its DEG recovery is weaker than TxPert.

AdaPert achieves the best performance across all reported metrics. It obtains the highest DES at both $k = 50$ and $k = 100$, indicating more accurate ranking of true DEGs. AdaPert also shows

Table 8: Hyperparameter search space used for the TxPert baseline. For each dataset, the best configuration is selected based on validation Pearson-Δ.

Category	**Hyperparameter**	**Search Space**
Model Architecture	Hidden dimension	{64, 128, 256, 512}
	Latent dimension	{64, 128, 256, 512}
	Dropout rate	{0.1, 0.2}
	Batch normalization	{True, False}
GNN (Perturbation Encoder)	Layer type	{GAT, GAT-v2}
	Number of layers	{2, 3, 4}
	Hidden dimension	{64, 128}
	Attention heads	{1, 2, 4}
	Skip connection	{None, Add, Concat}
	Self-loops	{True, False}
Training	Batch size	{32, 64, 128}
	Learning rate	$\{1\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}\}$
	Weight decay	$\{0, 1\times10^{-5}, 1\times10^{-4}\}$
	Max epochs	{100, 200}
	Early stopping patience	{20, 30}
	LR scheduler	{ReduceLROnPlateau, Cosine}
LR Scheduler	Reduction factor	{0.3, 0.5}
	Patience	{5, 10}
	Monitor metric	val_pearson_delta
Loss Function	MSE weight	{1.0}
	DEG weight	{0.0}
	Non-DEG Huber weight	{0.01, 0.02, 0.05}
	Non-DEG Huber δ	{0.5, 1.0}
Graph (STRING)	Edge selection	{Top-10, Top-20}
	Normalize weights	{True}
	Reduce to perturbations	{True}

Table 10: DEG-aware evaluation on the RPE1 test set.

Method	DES@50 ↑	DES@100 ↑	DE-Spear-Sig ↑	DE-Spear-LFC ↑	DE-Dir-Match ↑
scVI	0.036	0.041	-0.016	0.472	0.647
CPA	0.208	0.337	-0.048	0.659	0.703
GEARS	0.010	0.016	0.070	-0.089	0.471
TxPert	0.242	0.314	0.210	0.714	0.860
MORPH	0.090	0.155	0.207	0.544	0.773
Ours	**0.244**	**0.320**	**0.267**	**0.729**	**0.870**

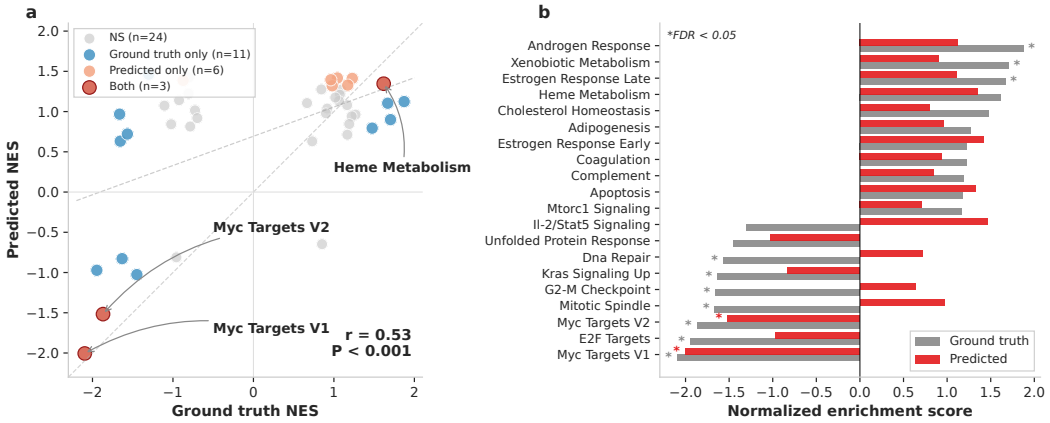


Figure 9: Correlation of pathway enrichment between predicted and ground truth responses for HIRA knockdown. Each point represents one of 44 Hallmark pathways, with the x-axis showing ground truth NES and the y-axis showing predicted NES. Colors indicate significance status (FDR < 0.25): gray, non-significant in both; blue, significant in ground truth only; coral, significant in predictions only; red, significant in both. Pathways significant in both analyses (Myc Targets V1, Myc Targets V2, and Heme Metabolism) are labeled. Dashed lines indicate the diagonal ($y = x$) and the linear regression fit. Pearson correlation $r = 0.53$, $P < 0.001$.

consistent improvements in DEG-specific Spearman correlation, log-fold change agreement, and direction matching. These results show that AdaPert produces sparse and reliable perturbation-specific responses on the RPE1 dataset.

G Correlation of Pathway Enrichment Between Predicted and Ground Truth HIRA Knockdown

To assess agreement between predicted and experimental pathway enrichment, we compare normalized enrichment scores (NES) across all 44 Hallmark gene sets. As shown in Figure 9, predicted and ground truth NES values show a significant positive correlation ($r = 0.53$, $P < 0.001$).

Among the 14 pathways significantly enriched in the ground truth analysis (FDR < 0.25), the model identifies 9 as significant. Three pathways (Myc Targets V1, Myc Targets V2, and Heme Metabolism) are significant in both analyses. Agreement is strongest for pathways with large effect sizes. For example, Myc Targets V1 shows closely matched enrichment between prediction (NES = -2.01) and ground truth (NES = -2.10), indicating accurate recovery of both direction and magnitude of pathway-level effects. You may include other additional sections here.