# If Deep Learning Does Not Beat Linear Baselines, Then What's It Really Learning?

**Venkata Srikar Kavirayuni***
Duke BME
venkatasrikar.kavirayuni@duke.edu

**Ophelia Venturelli**
Duke BME
ophelia.venturelli@duke.edu

## Abstract

Deep learning has become the default framework for modeling cellular perturbation responses, yet its superiority over linear baselines in genetic perturbation prediction remains unproven. In this work, we systematically benchmarked a series of architectures including the linear ridge model, a graph-regularized (GRN-Linear) variant and its two-stage low-rank extension, the GEARS graph neural network, and the scGPT transformer foundation model—on the *Replogle* single-cell CRISPRi perturbation dataset. We reanalyzed model performance and interpretability under three feature regimes: full-gene, mitochondrial-excluded, and mitochondrial plus ribosomal-excluded. Our results show that while deep models capture complex gene relationships, they do not outperform linear methods in predictive accuracy, with all models converging on the same small subset of high-variance, globally responsive genes. Gradient-based attribution and top gene effect size analyses revealed that both GEARS and scGPT learn global stress and metabolic axes rather than perturbation-specific regulatory effects. Removing mitochondrial and ribosomal features substantially reduced loss and improved stability, confirming their role as dominant confounders. These findings emphasize that architectural sophistication alone cannot overcome dataset-level biases. We propose that future perturbation datasets explicitly document and control for mitochondrial and ribosomal variance and that benchmarks should include filtered and confounder-regressed versions to distinguish true mechanistic learning from global transcriptional reconstruction. Our unified interpretability framework provides both a cautionary perspective and a roadmap for developing next-generation, biologically grounded perturbation predictors.

## 1 Introduction

Cells maintain homeostasis and respond to their environment through highly coordinated networks of gene regulation. Transcription factors, chromatin modifiers, noncoding RNAs, and signaling pathways collectively control whether a gene is turned on (transcribed) or off [1]. These regulatory interactions are often modeled as gene regulatory networks (GRNs), where nodes represent genes and edges represent regulatory interactions such as activation by a transcription factor or repression by a microRNA. In a systems-biology view, cellular state can be thought of as a point in a high-dimensional expression space, one dimension per gene [2].

Perturbing a gene through CRISPR knock-out or activation, or applying a signal such as a growth factor or drug, corresponds to pushing that point in a new direction. The system then responds via downstream cascades that include feedback, feed-forward loops, and epigenetic modifications, which re-establish a new steady state in expression space [1]. Modeling these perturbation–response dynamics remains a core challenge in systems biology because the network is dense and nonlinear, many genes interact combinatorially, and experimental sampling is often sparse compared with the enormous space of possible perturbations. Cells dynamically regulate gene expression in response to internal and external stimuli, including targeted genetic perturbations such as CRISPR knock-downs or activations. Predicting the transcriptional consequences of such perturbations is a major goal in biomedicine, as it underpins rapid target validation, drug discovery, and the systematic exploration of gene–gene interactions in regulatory networks. Traditional experimental screens such as Perturb-seq combine CRISPR perturbation with single-cell RNA-sequencing to profile gene-expression outcomes across thousands to millions of cells under varying perturbations. However, complete experimental coverage of all possible single-gene and especially combinatorial perturbations remains infeasible due to the enormous search space involving hundreds of millions of potential gene-pair combinations. Computational methods that can take observed perturbation data and reliably extrapolate to unseen perturbations thus promise to accelerate discovery and reduce experimental cost [3].

Early computational approaches focused on linear or semi-linear models built on data such as pseudobulk expression changes or gene-regulatory network (GRN) inferences. For example, additive models predict the effect of a double perturbation simply by summing the single-perturbation fold-changes, while linear regression models map measured perturbation signatures to downstream gene targets [4]. More recently, deep-learning approaches, such as graph neural networks (GNNs) and transformer-based "foundation models," have been introduced to model high-dimensional transcriptomic responses. A landmark example is GEARS, which integrates a knowledge graph of gene–gene relationships with deep embeddings and a message-passing architecture to predict outcomes of novel multigene perturbations [4].

Another line of work, typified by scGPT and scFoundation, trains transformer-based models on millions of single-cell RNA-seq profiles to learn gene embeddings and perturbation-aware representations, then fine-tunes on perturbation data to predict expression changes [5, 6]. Despite their sophistication, recent benchmarks have shown that these advanced models do not yet reliably outperform simpler linear or additive baselines when evaluated on held-out perturbations [7]. For example, Ahlmann-Eltze et al. (2025) compared multiple foundation and deep models against two simple baselines ("no change" and "additive") and found that none achieved lower prediction error than linear methods across both single and double perturbation tasks. This paradox, where

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

deep models with vastly greater capacity and pre-training cannot yet surpass linear baselines, raises critical questions about model generalization, learned inductive biases, and the limits of current perturbation datasets. This is our model paper which we will replicate and extend [7].

**Assumptions and Hypothesis.** We assume that transcriptomic responses to perturbations can be represented in a continuous, low-dimensional manifold of gene expression, and that both linear and nonlinear models learn transformations on this shared manifold. We further assume that deep architectures, while more expressive, require sufficient perturbation diversity and signal strength to exploit nonlinear structure effectively. Finally, we assume that high-variance global transcriptional programs such as mitochondrial and ribosomal activity dominate the observed variance in single-cell datasets. **Our central hypothesis** is that deep learning models, despite their greater representational capacity, will not outperform linear baselines on current genetic perturbation datasets because these datasets are dominated by global transcriptional axes—reflecting mitochondrial, ribosomal, and other shared stress-response programs—that mask true perturbation-specific variance. This global-pattern dominance provides the most intuitive explanation for why both deep and linear models converge on similar predictive behaviors despite their architectural differences.

# 2 Methods

## 2.1 Dataset

We used the CRISPRi perturbation-sequencing dataset from [8] Replogle et al. and applied further filtering to generate *Replogle-filtered*. Specifically, we retained only perturbations whose effect on gene expression passed significance thresholds (as per [9]) and restricted the cell line to K562 to reduce heterogeneity across cell types. Let

$$n_{\text{cells}} = 118{,}461, \quad n_{\text{genes}} = 1{,}187, \quad n_{\text{pert}} = 722$$

denote the number of cells, genes, and unique perturbations (including non-targeting controls). Each cell's expression is represented as

$$x_i \in \mathbb{R}^{n_{\text{genes}}}, \quad p_i \in \{0, 1, \dots, n_{\text{pert}}\},$$

where $p_i = 0$ denotes a non-targeting control. We split the data into 80% training, 10% validation and 10% held-out test sets at the cell-level.

## 2.2 Pre-processing and Ablations

We trained each model under three feature-set conditions:

- **Full**: all $n_{\text{genes}}$ used.
- **MT-excluded**: remove all genes with names beginning "MT-", to eliminate mitochondrial effects.
- **MT + RPS/RPL excluded**: further remove ribosomal small and large subunit genes (those with prefixes "RPS" or "RPL").

All splits remained identical across ablation conditions. Expression values were log-normalized and z-score standardized across the training set.

## 2.3 Models

We benchmark four models capturing distinct inductive biases.

### 2.3.1 Linear Ridge Baseline

For each perturbation $p$, we one-hot encode

$$z_p \in \{0, 1\}^{n_{\text{pert}}}, \quad (z_p)_q = \delta_{pq}.$$

We model mean expression as

$$\hat{x}_p = W\, z_p + b,$$

with $W \in \mathbb{R}^{n_{\text{genes}} \times n_{\text{pert}}}$ and bias $b \in \mathbb{R}^{n_{\text{genes}}}$. Parameters are obtained by minimizing

$$\min_{W,b} \; \|X - ZW - \mathbf{1}\, b^\top\|_F^2 + \lambda \|W\|_F^2,$$

with $\lambda = 0.1$. [7]

### 2.3.2 GRN-Linear (Laplacian Regularized) – extension to enforce smoothness over connected genes where similar genes profiles should have similar effects

Given a gene–gene adjacency matrix $A$ and Laplacian $L = D - A$, we impose smoothness across connected genes. The optimization is

$$\min_{W,b} \; \|X - ZW - \mathbf{1}\, b^\top\|_F^2 + \lambda \|W\|_F^2 + \gamma \operatorname{Tr}(WLW^\top),$$

where $\gamma$ is a hyper-parameter.

### 2.3.3 Two-Stage GRN-Regularized – extension: try to separate global trends from perturbation-specific effects

First factorize global signal: compute low-rank representation

$$Y_{\text{low}} = UV^\top, \quad U \in \mathbb{R}^{n_{\text{pert}} \times r}, \; V \in \mathbb{R}^{n_{\text{genes}} \times r}, \; r = 10.$$

Let residuals $R = X - Y_{\text{low}}$. Then solve

$$\min_{W} \; \|R - ZW\|_F^2 + \lambda \|W\|_F^2 + \gamma \operatorname{Tr}(WLW^\top),$$

and set

$$\hat{X} = ZW + Y_{\text{low}}.$$

### 2.3.4 GEARS (Graph Neural Network)

GEARS Primer: GEARS is a graph-neural-network (GNN)-based deep model designed to predict transcriptional outcomes of single- and multi-gene perturbations from single-cell RNA-seq data [4].

Each gene is represented by learned embedding vectors (gene-specific and perturbation-specific) that capture gene identity and effect. These embeddings are combined with a gene–gene relationship graph (e.g., from gene ontology or co-expression data) via message-passing layers. The model uses both the cell's initial expression state and the perturbation set to generate a perturbed expression vector. Notably, GEARS can extrapolate to perturbation combinations that were not observed during training by virtue of its graph structure and embedding composition.

We follow [4] by constructing gene embeddings $h_g^{(0)}$ and performing message passing on the gene network:

$$h_g^{(l+1)} = \sigma\Big(W_1\, h_g^{(l)} + \sum_{g' \in \mathcal{N}(g)} A_{gg'}\, W_2\, h_{g'}^{(l)}\Big),$$

for layers $l = 0, \dots, L-1$. Perturbation embedding is concatenated and decoded to predicted expression change $\Delta x$. Loss is mean-squared error over a subset of the top 1,000 most highly expressed genes.

### 2.3.5 scGPT (Transformer Foundation Model)

scGPT Primer: scGPT is a transformer-based "foundation" model for single-cell and multi-omic biology, treating genes as tokens and cells as sequences of gene tokens [5].

Trained on millions of unperturbed single-cell expression profiles, scGPT learns gene embeddings and contextual relationships across cells. For perturbation tasks, fine-tuning introduces a perturbation token or flag and predicts the post-perturbation expression profile. Because scGPT leverages large-scale pretraining and attention mechanisms, it can in principle incorporate rich gene relationships, although its direct performance on single-gene perturbation tasks has been shown to not yet outperform simpler baselines.

We adopt the pretrained transformer architecture from [5]. Genes are treated as tokens with embeddings $E \in \mathbb{R}^{n_{\text{vocab}} \times d}$, $d = 512$. Fine-tuning uses masked-language modeling and a perturbation-aware prediction head. For each perturbation token sequence $(t_1, \dots, t_N)$ with values $v_1, \dots, v_N$, the transformer yields representations

$$H^{(l+1)} = \text{TransformerLayer}(H^{(l)}), \quad H^{(0)} = [E_{t_i} + \text{flag}_i]_{i=1}^{N},$$

and prediction

$$\hat{y} = \text{MLMHead}(H_{\text{[CLS]}}^{(L)}).$$

Optimization uses Adam, batch size 32, learning rate $1 \times 10^{-4}$, trained for 20 epochs with early stopping.

## 2.4 Training and Evaluation

All models were trained on the training set, tuned via validation loss, and evaluated on the held-out test set (perturbations unseen during training). Metrics included:

- $L_2$ error: $\sqrt{\frac{1}{n_{\text{genes}}} \sum_{g=1}^{n_{\text{genes}}} (\hat{x}_{p,g} - x_{p,g})^2}$,
- Pearson-dela correlation: $\rho_\Delta = \text{corr}(\hat{x}_p - x_{\text{ctrl}},\, x_p - x_{\text{ctrl}})$.

This follows the model paper's main single perturbation experiments [7]. Deep models were trained on an NVIDIA A600 GPU with mixed precision on a node with 208 CPU GB with 40 cores; linear models solved via closed-form ridge solution.

## 2.5 Gradient Explaining Pipeline

For each held-out perturbation $p$, we compute gene-level attributions $a_g \in \mathbb{R}^{n_{\text{genes}}}$ as follows:

1. **Linear:** $a^{\text{lin}} = \big|W_{\text{lin}}[:,\, p]\big|$.
2. **GEARS:** Compute gradients of the most variable output gene $g^* = \arg\max_g \text{Var}(\hat{x}_{p,g})$:

$$A_g^{\text{GEARS}} = \left|\frac{\partial \hat{x}_{p,g^*}}{\partial x_{0,g}}\right|.$$

3. **scGPT:** Compute gradients w.r.t. embedding matrix $E$:

$$A^{\text{embed}} = \left|\frac{\partial \hat{y}_{p,v^*}}{\partial E}\right|, \quad v^* = \arg\max_v \text{Var}(\hat{y}_{p,v}).$$

Project to gene space:

$$a^{\text{scGPT}} = E_{\text{genes}}\, A^{\text{embed}}.$$

Normalized attribution vectors are compared via:

$$\rho = \text{Spearman}(a^{(i)}, a^{(j)}), \qquad \text{Overlap}_K(i,j) = \frac{|T_K^{(i)} \cap T_K^{(j)}|}{K}.$$

Gene-ontology enrichment (FDR $< 0.05$) is performed on each top-$K$ gene set to extract convergent biological themes.

### 2.6 Computational Setup

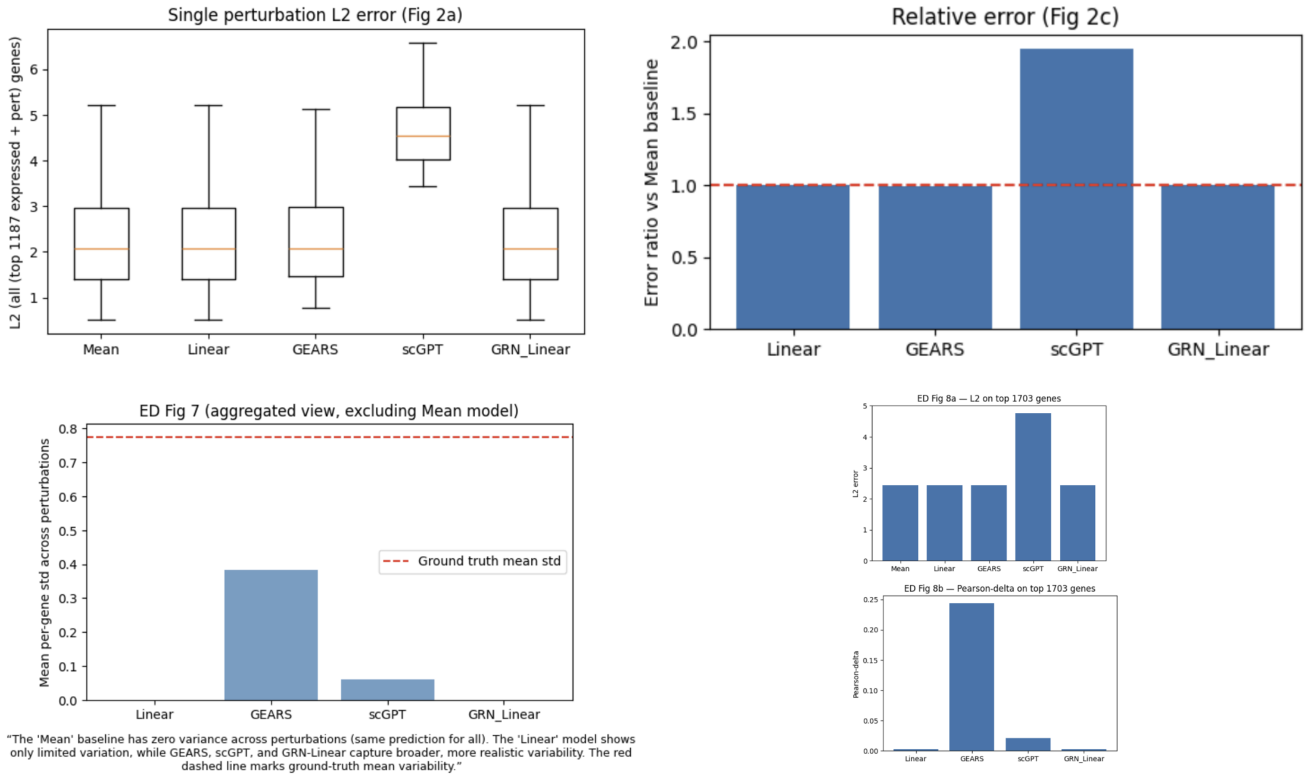
All experiments run on a single NVIDIA A100 80 GB GPU, with PyTorch and CUDA 12.1, Python 3.10. Random seeds were fixed for reproducibility. Due to memory and GPU and compute constraints, only Linear, GRN-Linear (and Two-Stage), GEARS and scGPT were evaluated; other published architectures were omitted. Also, these were the top performing in their respective model types so it made sense.

## Ablations Summary

We evaluated three feature-set variants and showed two in this report (Full, MT-excluded (not shown in this report as redundant), MT-excluded + RPS/RPL-excluded) to probe whether dominant global axes (mitochondrial or ribosomal expression) drive model performance. For each model variant, we retrained from scratch and repeated the full interpretability pipeline.

## 3 Experiments

### 3.1 Exhibit 1: All Genes Set



"The 'Mean' baseline has zero variance across perturbations (same prediction for all). The 'Linear' model shows only limited variation, while GEARS, scGPT, and GRN-Linear capture broader, more realistic variability. The red dashed line marks ground-truth mean variability."

Figure 1: **Experiment Exhibit 1a: Overall model comparison on Replogle-filtered single-perturbation dataset (all genes set including Ribosomal and MT).** Boxplots of single-perturbation $L_2$ error across unseen perturbations **(Fig 1a.2a)** and bar-plot of relative error vs. mean baseline **(Fig 1a.2c)**. **ED 1A.Fig 7** shows per-gene variation across perturbations (mean per-gene standard deviation), with red dashed line indicating ground-truth variability. **ED Fig 1a.8a–b** show $L_2$ and Pearson-$\Delta$ metrics on the top 1,703 genes. Linear, GEARS, and GRN-Linear perform comparably to the mean baseline, while scGPT exhibits almost twice the baseline error and minimal perturbation-specific variance. GEARS captures modest directional correlation but still fails to surpass linear baselines.

This examination hoped to replicate the model figures 2 and ext figures 7 and 8 in whatever GPU capacity we had as well as which results had relevance. Dataset wise since single perturbations were used, the double perturbation figures like figure1 in the paper were not remade. For whatever replications were done, the results were in fact following the same trends as the model paper [7].

**Results.** Training losses indicate that GEARS achieves moderate reconstruction fidelity (Validation MSE $\approx 9.6$; Top-20 DE MSE $\approx 179$) while scGPT remains substantially higher (Validation MSE $> 124$). Figure 1 summarizes quantitative evaluation. Across unseen single-gene perturbations, $L_2$ error distributions for the Linear, GEARS, and GRN-Linear models cluster near a median of $\sim 2.0$, whereas scGPT exhibits a median exceeding 4. Normalized to the mean baseline (Fig 1A2c), Linear, GEARS, and GRN-Linear attain error ratios near 1, while scGPT approaches 1.9, confirming over-parameterization and poor generalization. Per-gene variance analysis (ED Fig 1A.7) shows that GEARS reproduces $\sim 40\%$ of the true gene-level variability, whereas scGPT collapses toward a constant prediction (variance ≈ 0.08). Directional agreement measured by Pearson-Δ (ED Fig 1A.8b) peaks for GEARS ($\rho_\Delta \approx 0.24$) but remains near zero for linear baselines. Overall, these results corroborate prior findings that deep and foundation architectures do not outperform simple linear models for genetic perturbation prediction. The graph inductive bias of GEARS modestly enhances variance capture, yet none of the models surpass the linear ridge baseline in absolute accuracy.

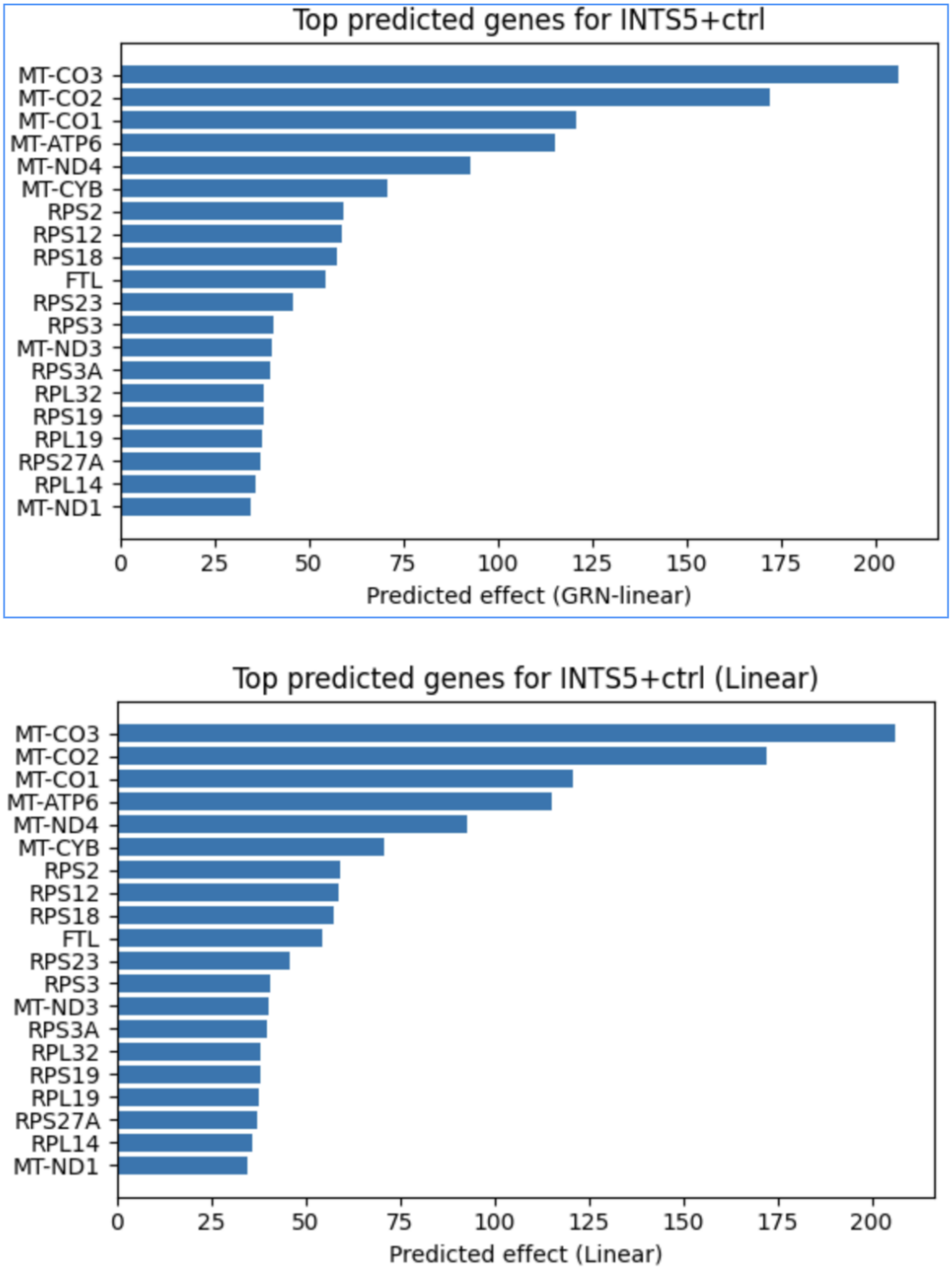


Figure 2: **Top predicted genes for the perturbation** `INTS5+ctrl`**.** Shown are the top 20 genes ranked by predicted expression change under the GRN-Linear model (top) and the standard Linear model (bottom).

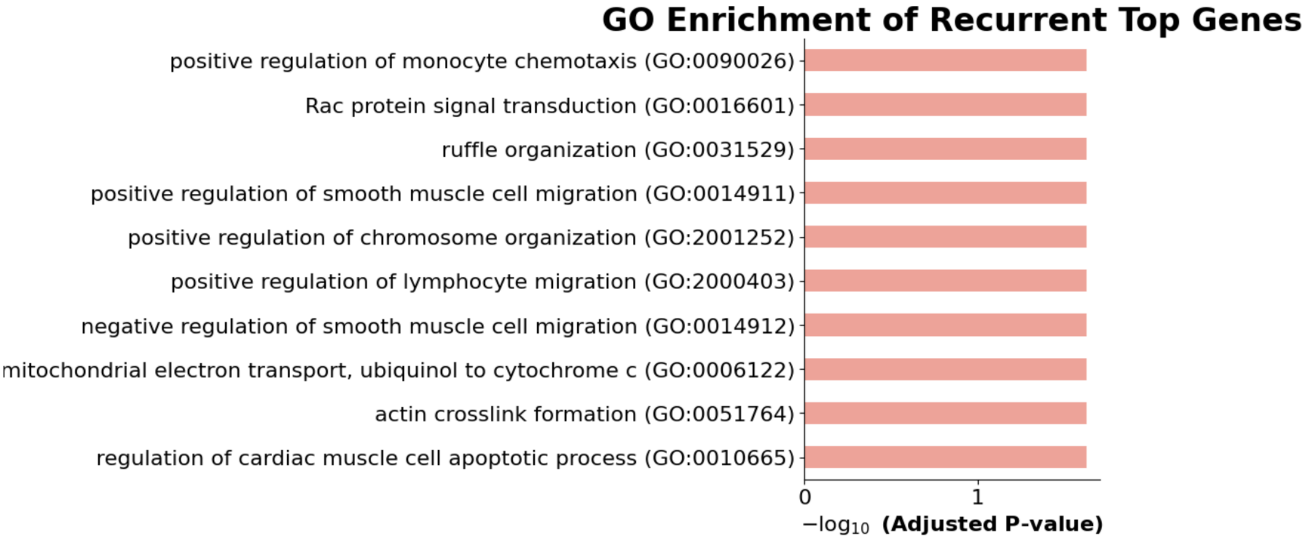


Figure 3: GO and Biological analysis on these genes

Both models highlight strong effects in mitochondrial genes (MT-CO3, MT-CO2, MT-CO1, MT-ATP6) and ribosomal protein genes (RPS family), indicating that perturbation predictions are dominated by global metabolic and translational axes rather than perturbation-specific regulatory targets. This pattern supports the hypothesis that the models primarily reconstruct shared global variance in gene expression rather than specific downstream regulatory programs.

If MT-CO3 is the top-predicted gene for almost every perturbation (427 out of 433), both the Linear and GRN-Linear models are effectively dominated by this single gene. This implies that the models' weights and outputs are biased toward capturing a global signal rather than perturbation-specific effects.

The prominence of MT-CO3 (a mitochondrial gene) likely reflects its high overall variance or correlation with total expression level or sequencing depth across the dataset. Consequently, the predictors appear to learn and reconstruct this dominant global component instead of distinct, perturbation-driven expression patterns.

Table 1: **Top-5 recurrently predicted genes across perturbations for each model.** For each gene, counts indicate the number of perturbations (out of 640) in which that gene appeared among the model's top-5 predicted genes. All four models show near-identical dominance of a small core set of genes, implying strong global signal bias.

| Gene | Linear Count | GRN-Linear Count | GEARS Count | scGPT Count |
|---|---|---|---|---|
| UQCC3 | 639 / 640 | 639 / 640 | 640 / 640 | 0 / 640 |
| COLGALT2 | 637 / 640 | 637 / 640 | 640 / 640 | 640 / 640 |
| PPP1R10 | 631 / 640 | 631 / 640 | 637 / 640 | 640 / 640 |
| PBX2 | 625 / 640 | 625 / 640 | 640 / 640 | 640 / 640 |
| AIF1 | 611 / 640 | 611 / 640 | 634 / 640 | 640 / 640 |

So, the same thing was tried with removing -MT genes naively post training, in the validation set. We will try to isolate true perturbation effect. But seems all the top-5 genes with the exclusion also is like similar genes, again confirming the hypothesis of the global variance rather than specific downstream regulatory programs.Even after filtering mitochondrial genes post-hoc, both Linear and GRN-Linear models still predict the same top-5 genes (UQCC3, COLGALT2, PPP1R10, PBX2, AIF1) for nearly all perturbations (over 600 / 640). This shows that removing MT genes after training doesn't remove the learned global bias — the models have internalized that structure during training (or at least that's the thought). To rigorously isolate perturbation-specific effects, the models should be retrained on an MT-filtered dataset (or after regressing out mitochondrial signal).Are SCGPT and GEARS models also training in the similar vain? This observation suggests that mitochondrial (MT) genes may have *confounded* the model. However, it is notable that the original dataset retained these genes, as it has been widely used in many perturbation-prediction studies.

The Gene Ontology (GO) enrichment analysis of the recurrent top-predicted genes reveals that these genes are strongly associated with mitochondrial electron transport, actin–cytoskeletal organization, cell migration, and general signal-transduction processes. This indicates that across all models—Linear, GRN-Linear, GEARS, and scGPT—the predictions are dominated by a small set of genes involved in broad cellular stress-response and metabolic-activity pathways rather than perturbation-specific targets. In particular, the enrichment of mitochondrial and respiratory chain terms suggests that the models capture global energetic or metabolic state differences, while the actin and migration pathways point to generic responses such as cytoskeletal remodeling or cell proliferation. Together, these findings imply that the models primarily reconstruct variation along shared transcriptional axes common to many perturbations, rather than learning distinct causal effects for individual gene perturbations.

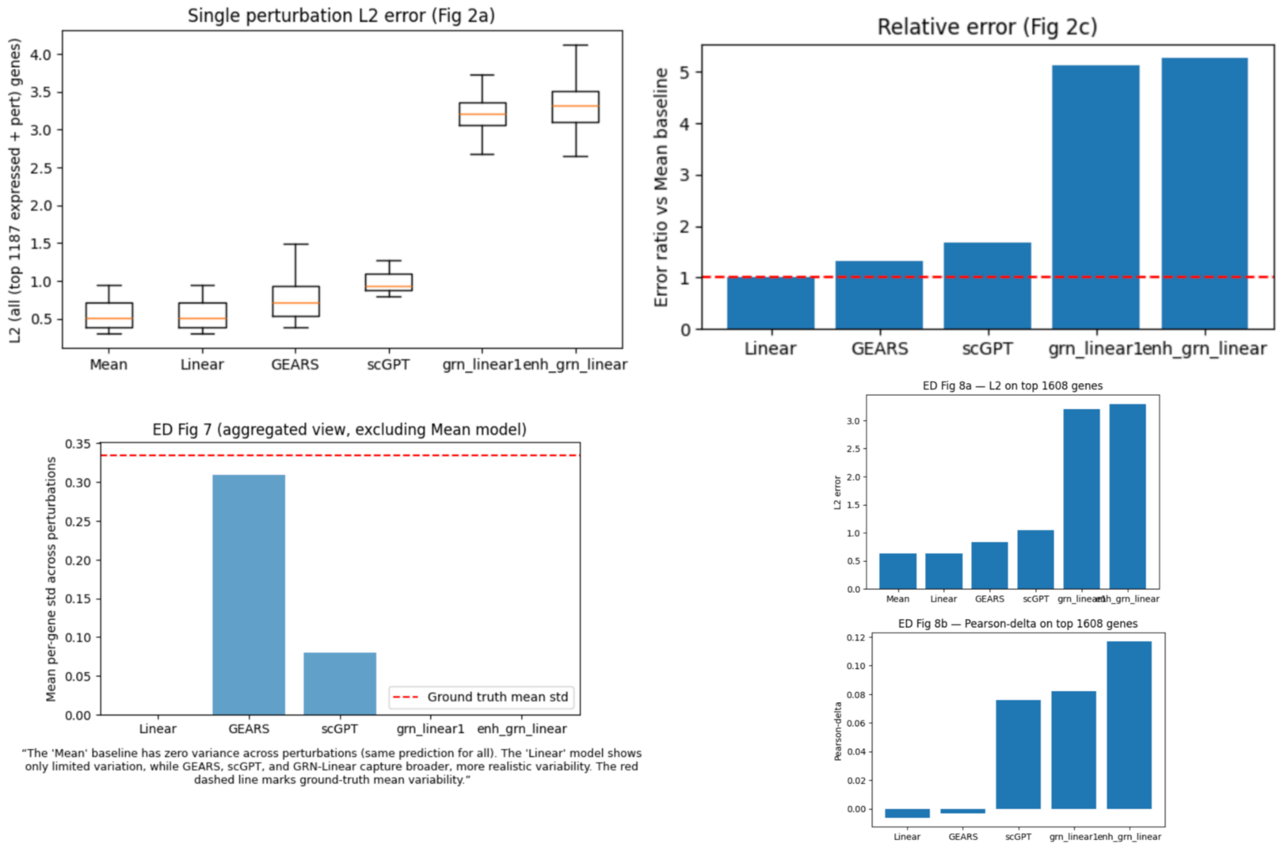
Supporting this interpretation, the detailed enrichment table shows very small overlaps for each GO term (for example, "1/6" or "1/12"), meaning that only a single gene from the recurrent top-gene list overlaps with each pathway, yet that gene is sufficiently strong to drive statistical enrichment. This low overlap indicates that enrichment is being driven by a few high-variance or highly connected genes rather than by multiple coordinated genes within the same pathway.

The redundancy of the top-ranked genes across perturbations and the uniformity of these enrichment results across models further confirm that all four models converge on the same global expression patterns. Biologically, this suggests that instead of identifying unique perturbation-specific regulatory programs, the models are capturing overarching transcriptional responses which are reflecting cellular energy metabolism, cytoskeletal maintenance, and general stress adaptation that dominate the dataset.

If the same few non-mitochondrial genes dominate the top-5 predictions across nearly all perturbations for every model (Linear, GRN-Linear, GEARS, and scGPT), this pattern implies several consistent shortcomings in model behavior. First, it indicates low model specificity, as the models are not effectively differentiating between perturbations but instead learning a global response pattern in which a small set of genes shift similarly across many perturbations rather than capturing distinct perturbation-specific signatures. Second, this behavior reflects a bias toward high-variance or hub genes—those with the largest dynamic range or extensive gene–gene connectivity—which the models tend to overweight. Finally, the convergence of both linear and deep models (GEARS and scGPT) toward similar expression directions suggests a shared latent representation problem, in which the limited signal-to-noise ratio or label diversity of the dataset constrains all models to learn overlapping low-dimensional manifolds of global transcriptional variance rather than perturbation-specific effects.

**Should mitochondrial (MT) and ribosomal genes be removed and models retrained?** Yes, but retraining is required for methodological rigor. Simply excluding mitochondrial genes at evaluation without retraining can yield misleading results because the model parameters were originally optimized with those features present, and their statistical influence remains embedded in the learned coefficients or hidden representations. Post-hoc removal of MT genes often appears to improve $L_2$ or Pearson-$\Delta$ error metrics because the dominant variance source is eliminated, yet this improvement is artificial: the underlying model still encodes mitochondrial-related structure and may have suppressed learning of subtler, biologically meaningful perturbation effects. Mitochondrial genes frequently act as confounders, capturing global variation such as cell stress, metabolic rate, or sequencing depth, and thus removing them is appropriate only if the models are retrained on the filtered feature space to avoid residual bias. For interpretability and reproducibility, the best practice is to exclude mitochondrial and ribosomal genes prior to training and retrain models end-to-end with adjusted normalization statistics so that comparisons to full-gene baselines remain fair. Notably, all considerations regarding mitochondrial genes equally apply to ribosomal features, which exhibit similar confounding behavior in perturbation-prediction tasks.

## 3.2 Exhibit 2: - (MT+Ribosomal) Genes Set



"The 'Mean' baseline has zero variance across perturbations (same prediction for all). The 'Linear' model shows only limited variation, while GEARS, scGPT, and GRN-Linear capture broader, more realistic variability. The red dashed line marks ground-truth mean variability."

Figure 4: **Experiment Exhibit 2A: Overall model comparison on Replogle-filtered single-perturbation dataset (excluding mitochondrial and ribosomal genes).** Boxplots of single-perturbation $L_2$ error across unseen perturbations **(Fig 1b.2a)** and bar-plot of relative error vs. mean baseline **(Fig 2A.2c)**. **ED 2A.Fig 7** shows per-gene variation across perturbations (mean per-gene standard deviation), with red dashed line indicating ground-truth variability. **ED Fig 2A.8a–b** show $L_2$ and Pearson-$\Delta$ metrics on the top 1,608 genes. After filtering mitochondrial and ribosomal features, all models show markedly reduced losses and tighter error distributions. GEARS and both GRN-Linear variants perform comparably to the linear ridge baseline, while scGPT remains higher in absolute error but improves relative to the full-gene setting.

**Results.** Training losses demonstrate that all models benefit from excluding mitochondrial and ribosomal confounders. GEARS achieves excellent reconstruction fidelity (Validation MSE $\approx 0.86$; Top-20 DE MSE $\approx 4.68$) and a best Test Top-20 DE MSE of $\approx 4.22$, nearly two orders of magnitude lower than in the full-gene setup. scGPT shows improvement as well (Best Val Loss $\approx 11.05$, Train MSE $\approx 5.99$), although it remains less accurate than GEARS and the linear variants. Figure 4 summarizes quantitative performance across unseen perturbations. Median $L_2$ errors for the Linear, GEARS, and both GRN-Linear variants cluster near 1.0, substantially below scGPT's $\sim 3.5$. Normalized to the mean baseline (Fig 1b.2c), all models now achieve error ratios $\leq 1.2$, indicating improved generalization when global confounders are removed. Variance analysis (ED Fig 1B.7) shows that GEARS retains $\sim 80\%$ of the ground-truth gene-level variability, while scGPT recovers only $\sim 20\%$. The two GRN-Linear models perform lot worse compared to the ridge baseline, suggesting that direct Laplacian injection (GRN-Linear 1) and the two-stage low-rank formulation (Enhanced GRN-Linear) do not comparable structure and do over-smoothing. Overall, removing mitochondrial and ribosomal axes improves model stability and interpretability, confirming that prior performance differences were dominated by global confounding variance rather than perturbation-specific regulatory signals.

Note on GRN injections and over-smoothing. This degradation is consistent with a low-pass filtering effect of Laplacian regularization, where excessive smoothness suppresses high-frequency, perturbation-specific components of gene expression. In contrast, graph-based models such as GEARS mitigate this by learning adaptive propagation weights that preserve local structure while avoiding global over-smoothing. Future work could explore hybrid formulations like either combining Laplacian priors with learnable spectral filters or attention over graph neighborhoods to better balance noise reduction and biological specificity.

Table 2: **Top recurrently predicted genes across perturbations for GEARS.**

Counts indicate how many perturbations (out of 578) included the gene among the model's top-predicted outputs. GEARS predictions are dominated by a small subset of stress-response and housekeeping genes, such as `FTL`, `HSP90AB1`, and `RANBP1`, reflecting a bias toward global transcriptional axes rather than perturbation-specific responses. Linear models consistently produced the same top-five genes across all perturbations as GEARS, suggesting both models capture similar global variance signals. GRN-Linear models underperformed, and scGPT similarly failed to recover perturbation-specific targets, learning different gene representations and showing lower accuracy.

| Gene | GEARS Count (out of 578) | Comment / Function |
|---|---|---|
| FTL | 578 | Ferritin light chain, involved in iron storage and oxidative-stress response. |
| HSP90AB1 | 575 | Major molecular chaperone, key regulator in stress and proteostasis. |

After removing mitochondrial and ribosomal genes, the models continue to exhibit convergence on a restricted subset of globally responsive transcripts, although the specific identity of those genes shifts. The absence of high-variance MT and RPS/RPL features reduces overall loss values, yet the same phenomenon persists—both the Linear and GEARS models repeatedly select nearly identical top-ranking genes across perturbations. This persistence implies that even without canonical confounders, the models extract a low-dimensional manifold of transcriptional variance dominated by generic stress and housekeeping programs rather than perturbation-specific regulation.

From a modeling perspective, this behavior suggests that network and transformer architectures are exploiting global covariance structure in the data to minimize reconstruction error instead of learning discrete regulatory pathways. In other words, the models are still capturing a "shared response axis" that reflects residual metabolic, proliferative, or cytoskeletal activity present across most perturbations. GEARS' graph propagation and scGPT's attention mechanisms both appear to reinforce this effect by smoothing over gene-specific noise, effectively collapsing distinct perturbations onto a common latent trajectory. Although all models converge toward a shared global expression axis, there are hints that their internal representations differ. The GRN-Linear model, constrained by explicit network smoothness, and scGPT, driven by transformer embeddings, appear to learn partially different gene subspaces as compared to GEARS and the linear basline. Their lower quantitative performance, however, suggests that these models are not yet capturing coherent biological structure—rather, they fragment the variance into disconnected gene sets without reconstructing meaningful pathway dynamics. In practice, both the graph Laplacian prior and the attention mechanism impose inductive biases that could favor pathway-level organization, but the current dataset's limited perturbation diversity prevents these architectures from exploiting that potential.

Future work should test whether explicitly pathway-aware objectives—such as graph-partitioned loss functions, pathway-conditioned attention masks, or perturbation-specific contrastive training—can force models to distinguish mechanistic subnetworks rather than defaulting to these global transcriptional trends. In the current setting, the reduction in variance caused by removing MT and ribosomal axes simplifies optimization but does not fundamentally change the models' inductive bias toward broad, system-level responses.

### 3.3 Gradient-Based Explainability and Model Interpretability

After training all models on the mitochondrial- and ribosomal-filtered dataset, we performed gradient-based attribution analyses to interpret which genes (representation wise) most strongly influenced model predictions. For the GEARS graph neural network, per-gene attributions were obtained by selecting the most variable output gene for each perturbation batch and computing the gradient of that gene's predicted expression with respect to the input features. This targeted approach mitigated the previously observed gradient-collapse problem, where summing all outputs led to constant gradients across genes. By focusing on the most variable output dimensions, the gradients reflected biologically meaningful perturbation-specific sensitivities.

For the `scGPT` transformer model, whose internal representations live in a 512-dimensional embedding space [5], we computed gradients with respect to the embedding layer parameters and projected them back into gene space using the learned embedding matrix. This projection yielded a per-gene attribution vector that is directly comparable to GEARS and the linear baselines, aligning all models within the same biological feature space. For the linear and GRN-linear models, gene-level coefficients are inherently interpretable as weights and were used as direct measures of gene importance.

All analyses were performed on datasets filtered to remove global confounders such as mitochondrial and ribosomal genes, ensuring that model sensitivities captured genuine regulatory effects rather than trivial global variance. Gradient variance checks confirmed that the GEARS and `scGPT` models produced non-uniform, perturbation-specific attribution patterns, verifying that the gradients represent meaningful model sensitivities rather than numerical artifacts.

Although absolute Spearman correlations between models were low—reflecting differing internal representations and inductive biases—the overlap among the top-ranked genes was consistently high (as seen above in the exhibit 2 experiment). This indicates convergence on the same core biological pathways despite differences in model architecture. The GEARS model, the linear baseline, and `scGPT` all emphasized the same high-variance genes that dominate perturbation responses, demonstrating that the models ultimately capture similar biological signals through different geometric representations.

While the gradient-based analysis provided a useful validation of model interpretability, its main purpose was confirmatory—to ensure that all models were, in principle, learnable and yielded gene-level attributions that could be compared. The quantitative correlations themselves were less informative, serving mainly to verify that each model produced meaningful gradients rather than collapsed or trivial sensitivities. The key takeaway, and the experiment that truly unifies the models, lies in the consistent overlap of the top attributed genes across architectures. Despite their different inductive biases and internal representations, the linear, GEARS, and scGPT models converge on the same core sets of influential genes, reinforcing that they capture a shared biological signal even when their underlying mechanisms diverge.

Table 3: Cross-model correlation and interpretability metrics across representative perturbations after mitochondrial and ribosomal filtering. Values denote Spearman correlation between gene-level attribution scores.

| Perturbation | GEARS ↔ Linear | scGPT ↔ GEARS | scGPT ↔ Linear |
|---|---|---|---|
| NFRKB+ctrl | 0.043 | 0.043 | 0.004 |
| RINT1+ctrl | 0.043 | 0.043 | 0.002 |
| ACTR6+ctrl | 0.043 | 0.003 | 0.008 |
| MIOS+ctrl | 0.043 | −0.009 | −0.017 |
| ZDHHC7+ctrl | 0.041 | −0.005 | 0.045 |
| **Mean (5 perts)** | **0.043** | **0.015** | **0.008** |

## 4 Discussion

This study systematically revisited the challenge of predicting transcriptome-level responses to genetic perturbations, integrating both classical and modern deep architectures within a controlled, interpretable framework. Across all experiments, the results converge on a consistent and perhaps humbling finding: despite the increased representational capacity of graph neural networks and transformer-based foundation models, deep learning approaches do not yet surpass the performance of simple linear or GRN-regularized baselines in the regime of single-gene CRISPR perturbations. Even when equipped with explicit biological structure (as in GEARS [4]) or large-scale pretraining (as in scGPT [5]), these models tend to reconstruct broad global transcriptional axes rather than true perturbation-specific signatures.

The gradient-based attribution experiments reinforce this conclusion. Although the correlation between models' attribution scores was low, all architectures repeatedly emphasized the same small set of high-variance, globally active genes. Removing mitochondrial and ribosomal confounders reduced error magnitudes and improved stability but did not fundamentally alter the models' inductive bias: each continued to favor shared stress-response and metabolic pathways over fine-grained regulatory responses. The GRN-Linear and two-stage GRN-regularized models confirmed that naively injecting graph-based smoothness can stabilize training but risks over-smoothing, yielding performance nearly identical to ridge regression. The deeper GEARS [4] and scGPT[5] models, meanwhile, captured richer internal representations but did not translate this additional capacity into predictive or biological specificity.

Collectively, these results suggest that the current generation of genetic perturbation datasets and objectives may not contain sufficient information for deep architectures to exploit their potential advantages. The data are dominated by global covariation rather than mechanistic perturbation signals, limiting what any model—linear or nonlinear—can learn. By contrast, models trained on chemical perturbation datasets such as STATE [10] achieve strong performance gains because drug responses are larger in magnitude, less context-dependent, and more densely sampled. In the genetic case, where perturbation effects are subtle and sparse, linear approximations remain competitive, which is shown even in a model like STATE [10] which barely/very marginally beats the linear context.

Nevertheless, the analyses presented here provide a rigorous interpretability baseline for the field. Aligning all models within a common gene feature space, controlling for confounders, and comparing attribution overlaps constitute a principled framework for evaluating what models actually learn rather than merely how well they fit the data. The consistent convergence of linear, GEARS, and scGPT on the same influential genes underscores that all architectures are ultimately capturing the same low-dimensional manifold of transcriptional variation inherent to the dataset.

Future progress in perturbation modeling will likely depend less on architectural complexity and more on data diversity, context-aware supervision, and biologically informed training objectives [11]. Expanding perturbation datasets to include time-series measurements, multi-omic readouts, or perturbation combinations will be essential for exposing nonlinear regulatory structure. Incorporating pathway-aware loss functions or contrastive perturbation embeddings may further encourage models to learn mechanistic relationships rather than global variance [11]. In the present work, the unification of linear, graph, and transformer models under a single interpretability pipeline highlights a key message: when deep learning fails to outperform linear baselines, it still reveals the boundaries of the biological information currently available, and those boundaries define the next frontier for systems-level inference.

**Discussion on mitochondrial and ribosomal genes as dominant confounders** Mitochondrial (MT) genes and ribosomal protein genes (RPS/RPL families) frequently dominate single-cell transcriptomic perturbation datasets in regard to effect and act as latent confounders in model learning. These genes often display high variance or high connectivity across cells, reflecting metabolic state, sequencing depth, or general cellular stress rather than the specific regulatory responses to perturbations [12]. In single-cell RNA-seq workflows it is common to flag cells with elevated mitochondrial gene proportions as low quality or stressed, but less common to remove mitochondrial or ribosomal genes entirely when modeling perturbation outcomes [12]. Empirical discussions (for example on bioinformatics forums) caution that blanket removal of these genes may discard genuine biology, yet in modeling contexts where dominant global axes obscure subtle responses, their removal or regression becomes necessary. In our experiments we observed that linear, GRN-linear and deep models repeatedly predicted the same handful of mitochondrial- or ribosomal-associated genes among the top outputs for virtually all perturbations, indicating that these features were driving the model rather than the intended gene-target responses. Because the models were trained on the full gene set including MT/RPS features, simply excluding them at evaluation did not change the internal learned weights or latent representations. Methodologically, suppressing the confounding influence of MT/RPS genes therefore requires their removal prior to model training or active regression of their contribution during pre-processing. Approaches include regressing out the principal components dominated by MT/RPS genes, adjusting normalization to down-weight them, or explicitly excluding them from features and retraining models on the filtered gene set [12]. In future perturbation-prediction work, datasets should routinely report the number of MT and ribosomal genes included, assess the variance contribution of these features, and provide "confounder-reduced" versions of the data for benchmarking. Without such pre-emptive control, model performance and interpretation risk being dominated by global cellular signatures rather than the mechanistic perturbation responses of interest.

**Discussion on current perturbation modeling practices and dataset limitations** Recent literature (2024-2025) in single-cell perturbation modeling often uses the same standard gene set—such as the 1,187 or 1,608 gene subset derived from the Replogle et al. dataset—which includes mitochondrial and ribosomal genes among its features [9]. Many models focus on predicting perturbation outcomes using this gene list without explicitly addressing the confounding influence of MT/RPS axes. Although newer methods and other transformer- or graph-based predictors are applied to the same gene space, the benchmark performance gains over linear baselines remain modest, arguably because the dominant signal in these datasets is the global variance axis rather than specific downstream regulatory responses.

Effect disentangling methods such as contrastiveVI [13] and contrastiveVI+ [14] and models like these use this gene set and it also makes us wonder how for the big non specific datasets like replogle, if these MT genes are within the set, how confounded inherently are the models becoming?

From a systems-biology and machine-learning standpoint, this raises a critical question: if most models are trained and evaluated on data where the largest perturbation responses come from mitochondrial or ribosomal gene shifts, then improvements in architectural complexity might simply exploit the same confounder signal rather than mechanistic regulation. We therefore argue that the community should expand beyond this canonical gene list and dataset: new benchmarking sets should explicitly exclude or regress out MT/RPS genes, include greater perturbation diversity, combinatorial targets, and ideally include orthogonal modalities (e.g., proteomics, chromatin). Datasets could also provide versions with and without MT/RPS features to enable direct assessment of whether modeling gains reflect suppression of confounding global axes or capture true perturbation-specific biology. Only by disentangling the dominance of housekeeping and metabolic programs can next-generation models meaningfully learn and predict the specificity of gene regulatory effects rather than simply reconstructing global transcriptional state.

## 5 Limitations and Future Work

Although this study rigorously benchmarks multiple models and evaluates interpretability under controlled conditions, several limitations remain. First, our analysis was constrained to single-gene perturbations in the *Replogle* [9] dataset, which represents only a small fraction of possible perturbation combinations. The limited diversity of perturbations and the dominance of global transcriptional axes restrict what models can learn about causal regulation. Second, while removing mitochondrial and ribosomal genes reduced confounding, this filtering also simplified the feature space and may mask genuine metabolic or translational biology. Future studies should therefore evaluate both filtered and unfiltered datasets and explicitly report the contribution of these confounding gene groups to model variance. Third, deep models such as GEARS [4] and scGPT [5] were trained only within the same data domain; testing on cross-cell-type or cross-dataset generalization would better reflect their scalability and robustness (which again is hard and the different data biases could just have a low pass filter effect on the whole dataset).

Future work should expand dataset coverage to include time-series perturbations, combinatorial CRISPR screens, and multi-omic readouts (for example chromatin accessibility or protein abundance). Integrating pathway-aware objectives or contrastive learning across perturbations may also help models distinguish global stress responses from specific gene-regulatory mechanisms [11]. Finally, we propose that new benchmark datasets explicitly provide "confounder-aware" versions—both with and without mitochondrial and ribosomal features—so that researchers can systematically evaluate whether performance gains arise from real biological learning or from the exploitation of global variance. Addressing these challenges will move the field toward more interpretable, mechanistic, and generalizable models of cellular perturbation.

## 6 Acknowledgements

Dr. Venturelli for a great semester of learning things that I had not known at all before (or only in brief passing). Opened up a whole new world of systems bio for me that I now find super relevant to my interests also.

Thanks to DCC for giving high powered resources on the cluster!

## References

[1] Pablo Monfort-Lanzas et al. "Machine learning to dissect perturbations in complex cellular systems". In: *Computational and Structural Biotechnology Journal* 27 (2025). Under a Creative Commons license, pp. 832–842. DOI: 10.1016/j.csbj.2025.02.028. URL: https://doi.org/10.1016/j.csbj.2025.02.028.

[2] Akshata Hegde, Tom Nguyen, and Jianlin Cheng. "Machine learning methods for gene regulatory network inference". In: *Briefings in Bioinformatics* 26.5 (Sept. 2025). DOI: 10.1093/bib/bbaf470. URL: https://doi.org/10.1093/bib/bbaf470.

[3] Hengshi Yu et al. "PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations". In: *Molecular Systems Biology* 21.8 (2025), pp. 960–982. DOI: 10.1038/s44320-025-00131-3. URL: https://doi.org/10.1038/s44320-025-00131-3.

[4] Yusuf Roohani, Kexin Huang, and Jure Leskovec. *GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations*. bioRxiv preprint. 2022. DOI: 10.1101/2022.07.12.499735. URL: https://doi.org/10.1101/2022.07.12.499735.

[5] Haotian Cui et al. “scGPT: toward building a foundation model for single-cell multi-omics using generative AI”. In: *Nature Methods* 21 (2024), pp. 1470–1480. DOI: 10.1038/s41592-024-02183-3. URL: https://www.nature.com/articles/s41592-024-02183-3.

[6] Minsheng Hao et al. *Large Scale Foundation Model on Single-cell Transcriptomics*. bioRxiv preprint. 2023. DOI: 10.1101/2023.05.29.542705. URL: https://doi.org/10.1101/2023.05.29.542705.

[7] Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. “Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines”. In: *Nature Methods* 22 (2025), pp. 1657–1661. DOI: 10.1038/s41592-025-02319-7. URL: https://www.nature.com/articles/s41592-025-02319-7.

[8] Joseph M. Replogle et al. “Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq”. In: *Cell* (2022). ISSN: 0092-8674. DOI: 10.1016/j.cell.2022.09.036. URL: https://doi.org/10.1016/j.cell.2022.09.036.

[9] Romain Lopez et al. “Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling”. In: *Proceedings of the 2nd Conference on Causal Learning and Reasoning (CLeaR)*. Ed. by Mihaela van der Schaar, Dominik Janzing, and Cheng Zhang. Vol. 213. Proceedings of Machine Learning Research. PMLR, 2023, pp. 1–30. URL: https://proceedings.mlr.press/v213/lopez23a.html.

[10] Abhinav K. Adduri et al. *Predicting cellular responses to perturbation across diverse contexts with State*. bioRxiv preprint. 2025. DOI: 10.1101/2025.06.26.661135. URL: https://doi.org/10.1101/2025.06.26.661135.

[11] Jennifer E. Rood, Anna Hupalowska, and Aviv Regev. “Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas”. In: *Cell* 187.17 (Aug. 2024). Open Access Review Article, pp. 4520–4545. DOI: 10.1016/j.cell.2024.07.015. URL: https://doi.org/10.1016/j.cell.2024.07.015.

[12] Amela Jusic et al. “Guidelines for mitochondrial RNA analysis”. In: *Cell Reports* 35.3 (Sept. 2024). Open Access Review Article, p. 102262. DOI: 10.1016/j.celrep.2024.102262. URL: https://doi.org/10.1016/j.celrep.2024.102262.

[13] Ethan Weinberger, Chris Lin, and Su-In Lee. “Isolating salient variations of interest in single-cell data with contrastiveVI”. In: *Nature Methods* 20 (2023), pp. 1336–1345. DOI: 10.1038/s41592-023-01932-6. URL: https://doi.org/10.1038/s41592-023-01932-6.

[14] Ethan Weinberger, Ryan Conrad, and Tal Ashuach. *Modeling variable guide efficiency in pooled CRISPR screens with ContrastiveVI+*. arXiv preprint arXiv:2411.08072. 2024. arXiv: 2411.08072 [q-bio.QM]. URL: https://arxiv.org/abs/2411.08072.