A PITFALL IN THE EVALUATION OF GENE PERTURBATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Gene perturbation models aim to predict transcriptional responses to unseen genetic interventions, but comparisons rely on heterogeneous evaluation metrics and differentially-expressed (DE) gene subsets. We investigate the consequences of these choices using eight methods across six single-gene perturbation datasets under a unified evaluation protocol. Varying the number of top DE genes reveals substantial instability in Precision-based rankings, with the winning method changing 86.7% of the time, while rankings based on the rest of the metrics remain comparatively stable. Investigating this discrepancy exposes that repeating a single predicted expression profile across cell-level rows, a standard practice for many methods, can inflate significance estimates and produce excessively large predicted significant-DE sets. Specifically we found that 67.1% of tested perturbation–gene pairs pass the significance threshold despite some of them having minuscule effects. We derive an inflation factor showing that replicated predictions can amplify the standardized Mann–Whitney statistic even after correction for tied ranks, without adding independent information. This artifact changes Precision mainly as the number of DE genes increases because for smaller cutoffs, most of the genes are already significant due to the effect-based ordering. Taking this pitfall into account in our benchmark comparison, we find that a random forest using Gene Ontology features achieves the best overall results.

1 INTRODUCTION

Large-scale single-cell perturbation screens provide a promising means of studying gene function and identifying potential therapeutic targets. However, experimentally measuring every possible genetic intervention is prohibitively expensive. Gene perturbation models (GPMs) aim to address this limitation by predicting the transcriptional response to previously unseen interventions, typically represented as the change in gene expression relative to non-targeting controls. Recent methods span linear predictors, biologically informed models, graph neural networks, and pretrained foundation models, yet there is still limited agreement on which modeling paradigms perform best.

A major obstacle to reaching consensus is the absence of a common evaluation protocol. Most studies quantify numerical reconstruction, but apply the same families of metrics to substantially different gene set sizes. Error-based metrics such as MSE, MAE, and RMSE are computed on the top 20 observed differentially expressed (DE) genes (Roohani et al., 2024; Wenteler et al., 2025), over complete profiles (Roohani et al., 2025; Adduri et al., 2025; Liu et al., 2026; Miladinovic et al., 2025), over the 1,000 most highly expressed genes (Ahlmann-Eltze et al., 2025), or over combinations of highly variable and DE genes on different thresholds N (Wenteler et al., 2025; Li et al., 2026; Wei et al., 2026). Pearson correlation is similarly reported over all genes (Roohani et al., 2024; Liu et al., 2026), over both all genes and the top 20 DE genes (Miladinovic et al., 2025), or just in top-DE genes (Csendes et al., 2025).

Accurate detection of affected genes or perturbations is equally important, but detection metrics also differ in both definition and gene scope. The 2025 Virtual Cell Challenge evaluates recovery of the complete perturbation-specific significant-DE set through its differential-expression score, while the associated Cell-Eval framework includes Precision@k, Overlap@k, and AUPRC,(Roohani et al., 2025; Adduri et al., 2025). Jaccard similarity, Precision@10, and precision–recall analysis are used in (Roohani et al., 2024), true-positive rate and false-discovery proportion in (Ahlmann-Eltze

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

et al., 2025), and AUC and false omission rate in (Miladinovic et al., 2025). PerturBench combines perturbation-rank metrics with recall of the top 20 DE genes (Wu et al., 2026). Li et al. (2026) combine reconstruction metrics computed over all genes and top-20, top-50, and top-100 subsets with top-50 DEG overlap and direction accuracy. Finally, Radig et al. (2026) uses top-20 DEG overlap and Wei et al. (2026) combine MSE and PCC-delta with a Common-DEGs metric, reporting results over all genes and the top 100 DE genes.

These divergent protocols render method rankings difficult to discern. In this work, we treat the evaluation protocol itself as the object of study. We compare eight methods and baselines spanning structured regression, Gene Ontology-derived features, frozen foundation-model representations, graph neural networks, and pretrained perturbation models across six single-gene perturbation datasets. All methods are evaluated using common perturbation-disjoint five-fold splits in unseen perturbation prediction. We systematically vary the number N of evaluated DE genes and we focus on $@N$ versions of MSE and Pearson for accuracy and Overlap and Precision for detection because they are the most representative.

We find that Precision-based rankings are particularly sensitive to the number evaluated DE genes and identify inflated predicted significance as a contributing mechanism. To be specific, ranking based on Precision through different N produces a mean absolute rank displacement of 2.58 positions and changes the top-ranked method in 86.7% of dataset–fold evaluations. In contrast, MSE, Pearson, and Overlap exhibit mean displacements below one position and winner-change rates no greater than 33.3%. Exact dataset-level sign-flip tests identify systematic reordering only for Precision after Benjamini–Hochberg correction.

Investigating this divergence reveals an artifact in DE testing. The evaluation wrapper repeats a single predicted expression profile across multiple rows to match the ground truth expression matrix, creating a predicted group with no within-condition variation. Mann–Whitney U testing against control cells can then assign very small p-values to weak predicted shifts.We further derive the replication-induced inflation factor of the standardized Mann–Whitney statistic, showing that correction for tied ranks does not account for the dependence among copied predictions. Under the assumptions of our analysis, 100 copies compared with 100 control cells multiply the standardized statistic by approximately 7.62, despite leaving the predicted expression unchanged and adding no independent information. In practice we find that across 620.9 million tested perturbation–gene pairs, 67.1% pass the nominal $\mathrm{FDR} < 0.05$ threshold, while 8.7% of these significant calls have minuscule predicted effect magnitudes of.The resulting predicted significant-DE lists are typically much larger than their ground truth counterparts: only approximately 22% of perturbations have at least 1000 observed significant genes, and fewer than 1% have at least 4000. This asymmetry interacts with the metric definitions. Overlap caps its effective cutoff at the observed significant-DE count, so the predicted p-values do not enlarge it. Precision, however, uses the predicted significant-DE count and can therefore continue evaluating inflated predicted lists. Since genes are ordered by absolute log-fold change, smaller N emphasize the strongest predicted effects which are indeed significant, hence the artifact, despite being there, has a limited impact on Precision. However as N increases, weaker effects are admitted by the inflated significance estimates, as proven by our experiments.

Taking this pitfall into account we create a benchmark comparison that aggregates results for small cutoffs of N. In this manner, we aim to account for the aforementioned variability between N but reduce the influence of the artifact on Precision. Overall the results agree with the literature that greater architectural complexity does not consistently improve perturbation prediction, but they are not in favor of simple linear models either. A Random Forest with gene ontology features Csendes et al. (2025) has the best performance overall, followed by the same model with pretrained embeddings from scFoundation Liu et al. (2026). This shows that informative biological or pretrained representations can be highly effective when combined with comparatively simple predictors.

2 Related Work

Known evaluation artifacts in GPMs. Recent studies have identified several ways in which standard evaluation protocols can distort measured performance and model rankings in GPMs. Viñas Torné et al. (2025) show that systematic transcriptional shifts shared across perturbations can dominate reference-based metrics, allowing a perturbed-mean baseline to perform competitively

without recovering perturbation-specific effects. Radig et al. (2026) show that several commonly used metrics induce highly correlated method rankings, such that aggregating redundant metrics can implicitly over-weight the corresponding aspect of performance. Nicol et al. (2026) demonstrate that computing observed and predicted perturbation effects relative to the same estimated control profile introduces spurious correlation in Pearson- and cosine-based evaluations, particularly when the number of control cells is limited. Csendes et al. (2025) show that including the perturbed gene among the evaluated genes can favor models that explicitly encode perturbation identity because that gene's effect change is obviously large, while Li et al. (2026) find that absolute-expression metrics favor weak perturbations whose post-perturbation profiles remain close to controls. Finally, the 2025 Virtual Cell Challenge exposed objective-specific incentives in composite evaluation. Because many submissions performed worse than the baseline on MAE and worse-than-baseline values were not further penalized, MAE ceased to influence the leaderboard and top teams concentrated on PDS and DES. Moreover, the L1-based PDS is scale-sensitive, such that increasing the magnitude of predicted effects can improve discrimination without improving magnitude calibration, hence one top submission explicitly selected a global scaling factor to optimize PDS (Roohani et al., 2025).

GPM benchmarks. Several recent studies have broadened the empirical evaluation of cellular perturbation models. PerturBench (Wu et al., 2026) provides standardized datasets, task definitions, model implementations, and metrics for genetic and chemical perturbations, including covariate-transfer and combinatorial-prediction settings. Foundation-model-focused evaluations include PertEval-scFM (Wenteler et al., 2025), which examines whether frozen single-cell foundation-model representations improve zero-shot genetic perturbation prediction, and Csendes et al. (2025), who compare scGPT and scFoundation with train-mean and biologically informed classical baselines across four Perturb-seq datasets. The latter study finds that simple baselines can outperform foundation models and identifies low perturbation-specific variance as a limitation of commonly used benchmark datasets. In a broader biomedical setting, El Kanbi et al. (2026) introduce a transcriptomic foundation-model benchmark comprising 35 tasks across eight immunological and inflammatory diseases, spanning target discovery, preclinical translation, treatment-response prediction, and patient stratification; perturbation prediction is one component of this wider drug-development benchmark. Broader comparisons of gene perturbation models have also rapidly expanded in scale. Wei et al. (2026) compare 27 methods across 29 datasets under perturbation- and cellular-context generalization settings while Li et al. (2026) conduct a task-spanning comparison of 13 methods on 25 datasets using 24 metrics covering expression accuracy, differential-expression recovery, and distributional similarity. Radig et al. (2026) evaluates the transfer of non-genetic perturbation responses across cell types and individuals rather than unseen genetic interventions.

3 Experimental Design

3.1 Evaluation Metrics

We evaluate model performance with metrics computed on increasingly large sets of differentially expressed (DE) genes and over the complete gene set. The finite-N analysis uses $N \in \{50, 100, 200, 1000, 2000, 4000\}$ and includes two metric families Detection $= \{\text{Overlap@}N, \text{Precision@}N\}$, and numerical accuracy Accuracy $= \{\text{MSE@}N, \text{Pearson@}N\}$ as explained below.

We use **cell-eval** (Adduri et al., 2025) to compute differential-expression statistics from observed and predicted profiles. Each model receives the name of the perturbation as input, and its output is represented as a vector $\widehat{\Delta} \in \mathbb{R}^{|G|}$ corresponding to the predicted effect. For methods producing a single profile per perturbation, this vector is repeated across the prediction rows into an effect matrix $\widehat{\mathbf{\Delta}} \in \mathbb{R}^{|C|\times|G|}$, where C is the set of observed cells receiving that perturbation, G is the set of expressed genes in the profile. The observed effect matrix $\mathbf{\Delta}$ has the same dimensions so that they can be compared.

The average control expression is added to each row of $\widehat{\mathbf{\Delta}}$ to derive the predicted post-perturbation expression matrix:

$$\overline{x}_{\text{ctrl}} = \frac{1}{|C_{\text{ctrl}}|} \sum_{j \in C_{\text{ctrl}}} x_j^{\text{ctrl}}, \qquad \widehat{\mathbf{X}} = \widehat{\mathbf{\Delta}} + \mathbf{1}_{|C|} \overline{x}_{\text{ctrl}}^{\top},$$

where C_{ctrl} is the set of non-targeting control cells and is not necessarily equal to $|C|$, and $x_j^{\text{ctrl}} \in \mathbb{R}^{|G|}$ is the expression profile of control cell j and $\mathbf{X} \in \mathbb{R}^{|C|\times|G|}$ be the observed post-perturbation expression matrix.

For each perturbation and each tested gene, DE is assessed separately by comparing (i) expression across observed perturbed cells with expression across non-targeting control cells, and (ii) expression across predicted rows with expression across the same control cells. Each comparison uses the Mann–Whitney U test (Mann & Whitney, 1947), a nonparametric test based on the ranks of expression values pooled across the two groups. It tests the null hypothesis that the groups have the same underlying expression distribution, without assuming normality. Thus, each gene receives separate observed and predicted p-values:

$$p_g^{\text{obs}} = \text{MWU}\left((X_{ig})_{i\in C}, (X_{jg}^{\text{ctrl}})_{j\in C_{\text{ctrl}}}\right),$$
$$p_g^{\text{pred}} = \text{MWU}\left((\widehat{X}_{ig})_{i\in C}, (X_{jg}^{\text{ctrl}})_{j\in C_{\text{ctrl}}}\right),$$

where MWU returns the p-value from the test. Predicted expression is not tested directly against observed perturbed expression. Within each perturbation, the p-values are adjusted across tested genes using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995), separately for the observed and predicted comparisons. Genes with adjusted $p < 0.05$, denoted $\text{FDR} < 0.05$, are retained as significant DE genes. We exclude the directly perturbed gene following the observation of (Csendes et al., 2025) and rank the remaining genes by decreasing absolute log-fold change relative to controls, using observed fold changes for the observed list and predicted fold changes for the predicted list:

$$R = \text{sort}_{|\text{LFC}_g^{\text{obs}}|\downarrow}\left(\{g : p_g^{\text{obs}} < 0.05,\ g \neq g_\star\}\right),$$
$$P = \text{sort}_{|\text{LFC}_g^{\text{pred}}|\downarrow}\left(\{g : p_g^{\text{pred}} < 0.05,\ g \neq g_\star\}\right),$$

where $g_\star$ is the directly perturbed gene and sorting is by decreasing absolute observed or predicted log-fold change relative to controls, respectively. Thus, p-values determine list membership, while fold-change magnitudes determine ordering. Let

$$R = (r_1, \ldots, r_{|R|}) \qquad \text{and} \qquad P = (p_1, \ldots, p_{|P|})$$

denote the resulting ranked lists of observed and predicted significant DE genes, respectively. Reconstruction metrics are evaluated on the strongest available observed DE genes,

$$S_N = R_{1:K_R}, \qquad K_R = \min(N, |R|).$$

For each perturbation, we compute MSE@N and Pearson@N using the signed observed and predicted perturbation effects averaged across cells and prediction rows, respectively. For each gene $g \in S_N$, these are defined as

$$\Delta_g = \frac{1}{|C|}\sum_{i\in C}(\boldsymbol{\Delta})_{ig}, \qquad \widehat{\Delta}_g = \frac{1}{|C|}\sum_{i\in C}(\widehat{\boldsymbol{\Delta}})_{ig}.$$

The metrics are then computed across the selected observed significant DE genes:

$$\text{MSE@}N = \frac{1}{K_R}\sum_{g\in S_N}\left(\widehat{\Delta}_g - \Delta_g\right)^2,$$

and

$$\text{Pearson@}N = \text{corr}\left((\widehat{\Delta}_g)_{g\in S_N}, (\Delta_g)_{g\in S_N}\right),$$

where corr denotes Pearson correlation across genes and $K_R = |S_N| = \min(N, |R|)$.

Overlap@N and Precision@N compare the identities of the observed and predicted significant DE genes, irrespective of effect direction. Overlap uses an effective cutoff determined by the number of observed significant genes,

$$\text{Overlap@}N = \frac{|R_{1:K_R} \cap P_{1:K_R}|}{K_R}, \qquad K_R = \min(N, |R|),$$

whereas Precision uses a cutoff determined by the number of predicted significant genes,

$$\text{Precision@}N = \frac{|R_{1:K_P} \cap P_{1:K_P}|}{K_P}, \qquad K_P = \min(N, |P|).$$

Thus, Precision@N uses a model-dependent effective cutoff because $|P|$ varies across models and perturbations, whereas Overlap@N uses an observed-data-dependent cutoff that is shared across models for a given perturbation. When both R and P contain at least N genes, $K_R = K_P = N$ and the two scores are identical. Perturbations with $K_P = 0$ are excluded from Precision@N, while those with $K_R = 0$ are excluded from Overlap@N. For brevity, we omit the @N suffix from metric names throughout the remainder of the text.

3.2 METHODS

Table 1: Overview of the evaluated methods. "Prior" denotes an explicit structured biological prior, while "Pretrained" indicates that the model or representation was learned from external data before training the benchmark-specific predictor.

Method	**Representation**	**Predictor**	**Prior**	**Pretrained**
Bilinear	Low-dimensional factors	Bilinear regression	None	No
RF-GO	GO-derived gene features	Random forest	GO	No
Ridge-scf	scFoundation embeddings	Ridge regression	None explicit	Yes
RF-scf	scFoundation embeddings	Random forest	None explicit	Yes
GEARS	GO node embeddings	GNN	GO	No
LPM	Perturbation embeddings	NN	None explicit	No
Average Control	Mean control profile	-	None	No
Average Train	Mean training profile	-	None	No

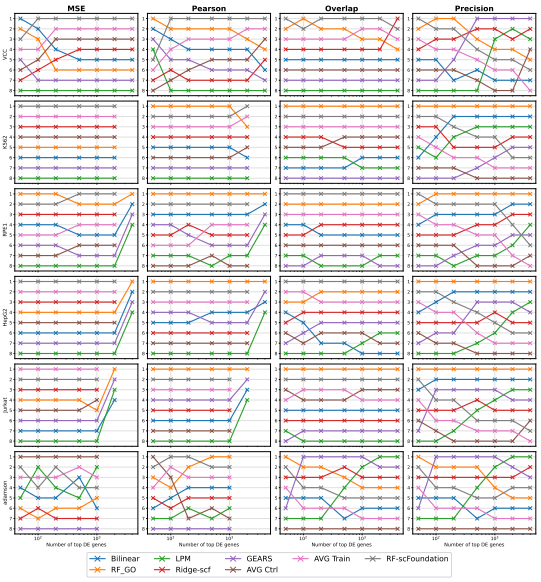
We study the prediction of transcriptional responses to previously unseen genetic perturbations. We benchmarked a diverse set of perturbation-response prediction methods, including the bilinear model of Ahlmann-Eltze et al. (2025), a random forest with Gene Ontology (GO) features (RF-GO) (Csendes et al., 2025), ridge regression using scFoundation representations (Ridge-scf) (Liu et al., 2026), random forest using pretrained scFoundation representations (RF-scf) (Liu et al., 2026), the graph neural network GEARS that also relies on GO (Roohani et al., 2024), and the Large Perturbation Model (LPM) (Miladinovic et al., 2025). We additionally include Average control (AVG Ctrl) and Average train (AVG Train) baselines (Viñas Torné et al., 2025) which are computed throughout the whole sets, not per batch. All methods are shown in Table 1. All methods are evaluated using the same perturbation-disjoint five-fold cross-validation splits, such that each test perturbation is absent from the corresponding training fold. Hyperparameters are fixed to the default configurations of the corresponding implementations, and no hyperparameter optimization is performed. A brief explanation of each method along with implementation details are given in App. A.2.

3.3 DATASETS

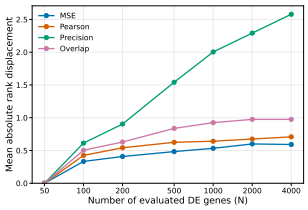
We use six single-gene perturbation datasets spanning different perturbation scales, number of cells, and experimental settings based on single-cell RNA-sequencing experiments using CRISPR interference (CRISPRi). As summarized in Table 2, the datasets contain between 86 and 2,393 perturbations, with an average of approximately 59 to 1,221 cells per perturbation. Median cell counts are below the corresponding means in every dataset, consistent with unequal sampling across perturbations. Control cells account for approximately 3–5% of all cells in Jurkat, RPE1, HepG2, and K562, compared with 17.3% in VCC and 35.4% in Adamson. The preprocessing pipeline along with a comparison between datasets are included in App. A.1.

Table 2: Dataset characteristics. Mean and median refer to cells per perturbation; perturbed and control counts include all perturbed and non-targeting control cells, respectively. "Genes" denotes the number of retained expression features.

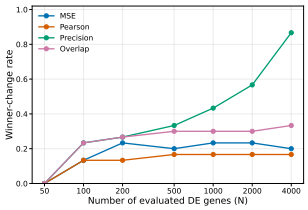
		Cells per perturbation		Cell counts		
Dataset	Perturbations	Mean	Median	Perturbed	Control	Genes
Jurkat (Nadig et al., 2025)	2,393	104.9	83.0	250,943	12,013	8,882
RPE1 (Replogle et al., 2022)	2,393	98.8	72.0	236,429	11,485	8,749
HepG2 (Nadig et al., 2025)	2,393	58.7	45.0	140,497	4,976	9,624
K562 (Replogle et al., 2022)	2,057	145.7	121.0	299,694	10,691	8,563
VCC (Youngblut et al., 2025)	150	1,220.6	1,045.0	183,097	38,176	18,080
Adamson (Adamson et al., 2016)	86	515.6	504.5	44,340	24,263	5,060



(a)



(b)



(c)

Figure 1: **(a)** Model rankings as a function of the number N of evaluated DE genes. Each row corresponds to a dataset and each column to an evaluation metric. Rank 1 denotes the best-performing method. Different datasets have different number of DE genes. **(b)** Mean absolute rank displacement relative to the ordering obtained at $N = 50$, averaged across dataset–fold blocks. **(c)** Frequency with which the top-ranked method changes relative to the ordering obtained at $N = 50$. The final cutoff corresponds to 4000 genes or the maximum available number of eligible DE genes when fewer are present.

4 Results

4.1 Inflated predicted significance distorts Precision-based evaluation

4.1.1 Precision-based rankings change with N

Figure 1a shows that the evaluated gene-set size can alter the ordering of competing methods. Rankings based on numerical accuracy are comparatively stable after the smallest values of N, whereas Precision exhibits frequent rank crossings across datasets. Overlap presents an intermediate pattern, with substantial changes in some datasets but greater stability than Precision. Mean absolute rank displacement measures how many ranking positions a method moves, on average, relative to its rank at $N = 50$, while the winner-change rate is the proportion of dataset–fold evaluations in which the top-ranked method differs from that at $N = 50$. Figures 1b and 1c show that both quantities increase substantially for Precision as N grows, reaching an average displacement of 2.58 positions

and a winner-change rate of 86.7% at the largest cutoff. In contrast, the rankings induced by MSE, Pearson, and Overlap remain substantially more stable, with mean displacements below one position and winner-change rates no greater than 33.3%. To test whether changing N systematically alters model rankings, we compare every pair (N_1, N_2) separately for each metric. After averaging model ranks across the five folds within each dataset, we define the model-wise rank change as

$$\Delta_{d,m}^{(k;N_1,N_2)} = \bar{r}_{d,m}^{(k,N_2)} - \bar{r}_{d,m}^{(k,N_1)},$$

where k denotes the metric, d the dataset, and m the model. We summarize systematic reordering as

$$T^{(k;N_1,N_2)} = \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{D} \sum_{d=1}^{D} \Delta_{d,m}^{(k;N_1,N_2)} \right)^2 .$$

For each metric and N-pair, we obtain an exact p-value by jointly sign-flipping the complete rank-change vector of each dataset over all 2^D possible assignments. The resulting p-values are adjusted across the 15 pairwise N-comparisons within each metric using the Benjamini–Hochberg procedure. Precision was the only metric for which systematic ranking changes remained significant after correction. Twelve of the 15 comparisons were significant ($p = 0.03125$, $q_{\mathrm{BH}} = 0.03906$), while the rest are still close to significance with ($p = q_{\mathrm{BH}} = 0.0625$). This instability is also reflected in the relationship between detection metrics. Agreement between Precision and Overlap falls from $\tau_b = 0.83$ at $N = 50$ to 0.03 at $N = 4000$, while agreement between MSE and Pearson decreases only from 0.77 to 0.66. This can instill bias in case of scores that aggregate through all four metrics. We have a devoted section on this metric disagreement in App. B.

4.1.2 Inflated predicted significance contributes to Precision instability

To understand why Precision becomes unstable while the rest remain comparatively robust, we will examine its definition and how the predicted significant-DE lists evolve with N. As defined in Sec. 3.1, Precision uses $K_P = \min(N, |P|)$, where P is the list of predicted significant genes. This changes based on the model, in contrast to Overlap, where it is determined by the observed significant-DE list and is therefore shared by all models for the same perturbation. In our evaluation, the real significant genes $|R| \geq N$ account for 35% of the overall cases, whereas the predicted significant genes $|P| \geq N$ are approximately 95%. The real significant gene set size $|R|$ starts at 70% for $N = 50$ and drops to 21% for $N = 1000$ and 6% for $N = 4000$ while the predicted $|P|$ go from 97% to 95% and 94% respectively. Consequently, in our experiments, while Overlap saturates due to the finite size of the observed DE set, Precision continues to evaluate longer prefixes because the predicted significant set is typically much larger than the observed one.

This excess of predicted significant genes does not necessarily imply that the predicted effects are larger. In fact, the results indicate that predicted significance can be inflated by an evaluation artifact. To be specific, of the 620.9 million tested perturbation–gene pairs, 416.5 million (67.1%) passed the predicted $FDR < 0.05$ threshold. The pattern was consistent across methods, and even the AVG Ctrl baseline, which contains no perturbation-specific information, classified 69.2% of pairs as significant. However, 36.3 million significant calls (8.7%) had a minuscule effect, i.e., $|\log_2 \mathrm{FC}| < 0.1$, including some effects arbitrarily close to zero. These results show that predicted significance is not determined by effect magnitude alone.

We hypothesize that replication of perturbation-level predictions contributes to this excess significance. During evaluation, each predicted expression profile is repeated to match the observed cell count, adding no independent information while increasing the nominal sample size supplied to the Mann–Whitney U test. We next analyze how this mismatch affects the test statistic and nominal p-values, and explain how an enlarged predicted significant-DE set can contribute to the observed sensitivity of Precision to N.

4.1.3 Theoretical analysis on the Precision sensitivity

To analyze the impact of replication of predicted expression profiles on the Precision evaluation, we revisit the mathematics underlying the Mann–Whitney U test and how it affects the DE gene lists. Recall that C, C_{ctrl} are the sets of perturbed and non-targeting control cells, respectively. For each

gene, the statistics $U_g^{\text{obs}}, U_g^{\text{pred}}$ on observed and predicted gene expression are defined as:

$$U_g^{\text{obs}} = \sum_{i \in C} \sum_{j \in C_{\text{ctrl}}} \left[\mathbb{1}\{X_{ig} > X_{jg}^{\text{ctrl}}\} + \frac{1}{2} \mathbb{1}\{X_{ig} = X_{jg}^{\text{ctrl}}\} \right],$$
$$U_g^{\text{pred}} = \sum_{i \in C} \sum_{j \in C_{\text{ctrl}}} \left[\mathbb{1}\{\widehat{X}_{ig} > X_{jg}^{\text{ctrl}}\} + \frac{1}{2} \mathbb{1}\{\widehat{X}_{ig} = X_{jg}^{\text{ctrl}}\} \right]. \tag{1}$$

To simplify the discussion, we denote $s \in \{\text{obs}, \text{pred}\}$ and omit the continuity correction. For each gene, assume that ties occur only among the replicated predictions. Let t_g^s denote the size of this single tie group, with $t_g^{\text{obs}} = 0$ for the observed comparison, which has no ties. With large samples, the standardized value $Z_g^s \xrightarrow{d} N(0,1)$, where:

$$Z_g^s = \frac{U_g^s - \mu_g^s}{\sigma_g^s}, \qquad \mu_g^s = \frac{|C| \times |C_{\text{ctrl}}|}{2},$$
$$(\sigma_g^s)^2 = \frac{|C| \times |C_{\text{ctrl}}|}{12} \left[|C| + |C_{\text{ctrl}}| + 1 - \frac{(t_g^s)^3 - t_g^s}{(|C| + |C_{\text{ctrl}}|)(|C| + |C_{\text{ctrl}}| - 1)} \right]. \tag{2}$$

The two-sided p-values are:

$$p_g^{\text{obs}} = \text{MWU}\left(((X_{ig})_{i \in C}, (X_{jg}^{\text{ctrl}})_{j \in C_{\text{ctrl}}}) \right) = 2\left(1 - \Phi(|Z_g^{\text{obs}}|)\right),$$
$$p_g^{\text{pred}} = \text{MWU}\left(\left((\widehat{X}_{ig})_{i \in C}, (X_{jg}^{\text{ctrl}})_{j \in C_{\text{ctrl}}} \right) \right) = 2\left(1 - \Phi(|Z_g^{\text{pred}}|)\right),$$

where Φ is the cumulative distribution function of the standard normal distribution.

Deterministic models and models that ignore cell heterogeneity can only generate a single prediction, because their prediction is purely based on the perturbation condition and target gene, without considering the variability among individual cells' gene expression profiles. During Mann–Whitney U test, Z_g^s is asymptotically normal because we assume that the sample sizes $|C|$ and $|C_{\text{ctrl}}|$ are sufficiently large. For models with single predictions, the prediction group contains only one profile and the assumption is violated.

Replicating a prediction can increase the nominal sample size. However, this replication strategy can also increase the number of predicted significant DE genes. Suppose we generate c replicates for each prediction, then $|C| = c$ and $\widehat{X}$ has c replicated rows. The single tie group then has size $t_g^{\text{pred}} = c$. Define

$$B(c, |C_{\text{ctrl}}|) = c + |C_{\text{ctrl}}| + 1 - \frac{c^3 - c}{(c + |C_{\text{ctrl}}|)(c + |C_{\text{ctrl}}| - 1)}. \tag{3}$$

Writing $U_g^{\text{pred}}(c)$ and $Z_g^{\text{pred}}(c)$ for the statistics with c copies, the tie-corrected variance used by the test is $(\sigma_g^{\text{pred}}(c))^2 = c|C_{\text{ctrl}}|B(c, |C_{\text{ctrl}}|)/12$. Thus,

$$U_g^{\text{pred}}(c) = c\, U_g^{\text{pred}}(1), \quad Z_g^{\text{pred}}(c) = Z_g^{\text{pred}}(1) \sqrt{\frac{c(|C_{\text{ctrl}}| + 2)}{B(c, |C_{\text{ctrl}}|)}}. \tag{4}$$

For $c = |C_{\text{ctrl}}| = 100$, the inflation factor for Z_g^{pred} relative to the single-prediction standardized statistic is approximately 7.62. As c increases, the inflation factor for Z_g^{pred} also increases, leading to an overestimation of statistical significance. The observed perturbed X remains unchanged with $|C|$ original samples, whereas the prediction side is inflated by the replication strategy. Thus, the predicted DE gene list P is longer, $|P|$ is larger, and the inflation propagates to K_P and the calculation of Precision.

Although the significance-estimation artifact is present at every cutoff N, smaller cutoffs yield comparatively more stable rankings and stronger agreement between Precision and Overlap. Since the predicted significant-DE list is ordered by absolute predicted log-fold change, smaller cutoffs emphasize the strongest predicted effects which have a higher chance of really being significant. Moreover, Precision and Overlap are identical whenever both the observed and predicted significant-DE lists contain at least N genes, a condition that is more readily satisfied at smaller cutoffs. As N increases beyond the observed significant-DE count, Precision can continue evaluating longer predicted prefixes while Overlap stops expanding, exposing the asymmetry between the two lists, as indicated also by the percentages in Sec. 4.1.1.

4.2 BENCHMARK

For the benchmark comparison, we restrict the evaluation to $N \in \{50, 100, 200\}$ as a pragmatic choice based on the observed ranking stability. This restriction does not correct the underlying significance estimates but limits evaluation to the strongest retained effects and hence reduces the influence of the artifact to Precision. We aggregate scores separately for each method and metric in three stages. First, within each dataset, we compute a mean score for each cutoff by averaging over the folds in which that cutoff was evaluated. Missing scores are excluded, rather than replaced by scores from smaller cutoffs. Second, we average these cutoff-specific means with equal weight to obtain one dataset-level score, excluding cutoffs with no available results. Thus, cutoffs evaluated in more folds do not receive greater weight in the dataset-level average. Finally, we average the dataset-level scores across datasets with available results, giving each dataset equal weight.

Overall the benchmark shows that greater architectural complexity does not consistently translate into better performance. RF-GO achieves the highest Pearson, Precision, and Overlap, while RF-scf obtains the lowest MSE and ranks second on the remaining metrics (Table 3). Notably, AVG Train ranks third in MSE, Pearson, and Overlap, demonstrating that even a simple averaging baseline remains competitive as indicated in the literature. These findings agree with part of the literature on simpler methods (Csendes et al., 2025; Viñas Torné et al., 2025); however, they disagree that simple regression models such as Bilinear are highly competitive (Ahlmann-Eltze et al., 2025). We hypothesize that this is because that work's conclusions rely partly on top 1000 expressed, not DE, genes, which is a very specific evaluation not used extensively in the literature. Finally, more complex models such as GEARS and LPM also underperformed compared to RF. Results disaggregated by dataset and cutoff are provided in App. C.

Table 3: Model performance across evaluation metrics. All values $\times 100$.

	MSE	Pearson	Precision	Overlap
Bilinear	14.31	47.70	6.15	6.09
LPM	25.14	26.70	4.21	4.42
Ridge-scf	12.32	45.96	6.06	6.37
GEARS	16.53	45.21	6.29	6.33
AVG Ctrl	14.59	33.23	4.15	4.69
AVG Train	12.40	48.92	6.15	6.65
RF-scf	**11.77**	51.92	6.61	7.10
RF-GO	12.68	**54.77**	**7.20**	**7.19**

5 CONCLUSION

We benchmarked eight methods and baselines across six single-gene perturbation datasets in order to discern the role of different cutoffs for DE genes in the results of current methods. We found that Precision changes significantly as N increases and we identified an evaluation artifact that contributes to this instability. Repeating predictions across cell-level rows to match the ground truth expression matrix can inflate predicted DE significance, which leads to admitting weak effects into the predicted significant-DE lists. Since Precision determines its effective cutoff from the predicted rather than observed significant-DE count, this effect increasingly influences evaluation as N grows. Restricting the benchmark comparison to smaller cutoffs does not alleviate the effect, but since the majority of genes at these cutoffs are significant, it diminishes it. To this end, we take an aggregate of the results for $N = \{50, 100, 200\}$. Under this regime, random forests using gene ontology features or scFoundation representations achieve the strongest results, while a training-mean baseline remains competitive.

Future work will develop and validate evaluation procedures suited to perturbation-level point predictions, rather than treating duplicated profiles as independent cellular observations. For small numbers of independent cell-level predictions, particularly for deterministic models, the exact Mann–Whitney U test should be used, so as to avoid artificially replicating predictions to increase the nominal sample size. Finally, we plan to extend the benchmark to other advanced methods (Cui et al., 2024; Adduri et al., 2025; Wenkel et al., 2026), and add finetuning versions of the large models.

STATEMENTS

AI USE STATEMENT

We used AI to identify assist with literature searches, typographical errors and improve language and LaTeX formatting. The authors reviewed the material incorporated into the manuscript and take full responsibility for its content.

REPRODUCIBILITY STATEMENT

Data preprocessing is described in Appendix A.1. Section 3 describes the evaluation protocol and hyperparameter settings, while Section 4.1.1 details the statistical tests for ranking changes. The supplementary code includes a README with instructions for reproducing the results and documentation of model-specific data handling and implementation details.

ETHICS STATEMENT

This work analyzes existing single-cell perturbation datasets to assess the reliability of model evaluation. Benchmark performance should not be interpreted as evidence of therapeutic efficacy, and downstream biological applications require independent experimental validation.

REFERENCES

Brian Adamson, Thomas M. Norman, Marco Jost, Michael Y. Cho, Kevin J. Nuñez, Yining Chen, Jacqueline E. Villalta, Luke A. Gilbert, Max A. Horlbeck, Marco Y. Hein, Ryan A. Pak, Andrew N. Gray, Carol A. Gross, Abhishek Dixit, Oren Parnas, Aviv Regev, and Jonathan S. Weissman. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016. doi: 10.1016/j.cell.2016.11.048.

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with state. *Cell*, 189(19):5914–5931, 2025. doi: 10.1016/j.cell.2026.07.052.

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

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

G. Csendes, G. Sanz, K. Z. Szalay, and B. Szalai. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. *BMC Genomics*, 26:393, 2025. doi: 10.1186/s12864-025-11600-2.

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.

Karim El Kanbi, Yannis Cattan, Pierre Marschall, Matthew Corney, Aziz Fouché, Vincent Bouget, and Julien Duquesne. A transcriptomic benchmark for foundation models in immunology and inflammation drug development. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2026*, 2026.

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

Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. Evaluating the utilities of foundation models in single-cell data analysis. *Advanced Science*, 13(27):e14490, 2026. doi: https://doi.org/10.1002/advs.202514490.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.

Djordje Miladinovic, Tobias Höppe, Mathieu Chevalley, Andreas Georgiou, Lachlan Stuart, Arash Mehrjou, Marcus Bantscheff, Bernhard Schölkopf, and Patrick Schwab. In silico biological discovery with large perturbation models. *Nature Computational Science*, 5(11):1029–1040, 2025.

Anjali Nadig, Joseph M. Replogle, Angela N. Pogson, Mohan Murthy, Steven A. McCarroll, Jonathan S. Weissman, Eric B. Robinson, and Luke J. O'Connor. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, 2025. doi: 10.1038/s41588-025-02169-3.

Phillip B Nicol, Shriya Shivakumar, and Rafael A Irizarry. Spurious correlation inflates performance in single-cell perturbation prediction. *Biorxiv: the Preprint Server for Biology*, 2026.

Jean Radig, Robin Droit, Daria Doncevic, Albert Li, Duc Thien Bui, Luis Herfurth, Thaddeus Kühn, and Carl Herrmann. scArchon: a scalable benchmarking framework for assessing single-cell perturbation models. *Genome Biology*, 27(1):162, 2026. doi: 10.1186/s13059-026-04104-z.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gilad Lithwick-Yanai, Nino Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Yusuf. Roohani, A. Dobin, N. Teyssier, et al. Behind the data of the virtual cell challenge. Arc Institute, August 2025. URL `https://arcinstitute.org/news/behind-the-data-virtual-cell-challenge`. Accessed: 2026-09-21.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, pp. 1–10, 2025. doi: 10.1038/s41587-025-02777-8.

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, Xingbo Yang, Chen Tang, Shaliu Fu, Xiaohan Chen, Wannian Li, Yuzhou You, Chen Zhang, Aibin Liang, Guohui Chuai, and Qi Liu. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23:451–464, 2026. doi: 10.1038/s41592-025-02980-0.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T Whitfield, Craig Russell, Yassir El Mesbahi, et al. Txpert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, pp. 1–8, 2026.

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. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR/OpenReview.net, 2025.

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

Nicholas D. Youngblut, Christopher Carpenter, Arshia Nayebnazar, Abhinav Adduri, Rohan Shah, Chiara Ricci-Tam, Jaanak Prashar, Rajesh Ilango, Noam Teyssier, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. scbasecount: an ai agent-curated, uniformly processed, and autonomously updated single cell data repository. *bioRxiv*, 2025. doi: 10.1101/2025.02.27.640494.

A Appendices

A.1 Data processing

For each dataset we load the provided `AnnData` object and harmonize it to a common schema. Gene identifiers are mapped to gene symbols and a `target_gene` field is added to every cell, recording either the perturbed gene or the control label `non-targeting`; control-cell labels from heterogeneous source naming conventions are unified to this single label. Duplicated gene names are removed, and any perturbation whose target gene is not itself among the measured/profiled genes is dropped, since such perturbations cannot be scored against a matching expression readout. For the Adamson dataset specifically, the provided expression matrix is supplied in natural-log1p-normalized form rather than raw UMI counts; we invert this transform ($\exp(x) - 1$) to recover linear-normalized expression values, consistent with the scale used by the other five datasets, while preserving matrix sparsity. Non-targeting control cells are separated from perturbed cells and held out as a dedicated control set, used both for baseline predictors and as a reference profile during evaluation. An overview of the dataset characteristics and comparison can be seen in Fig. 2.

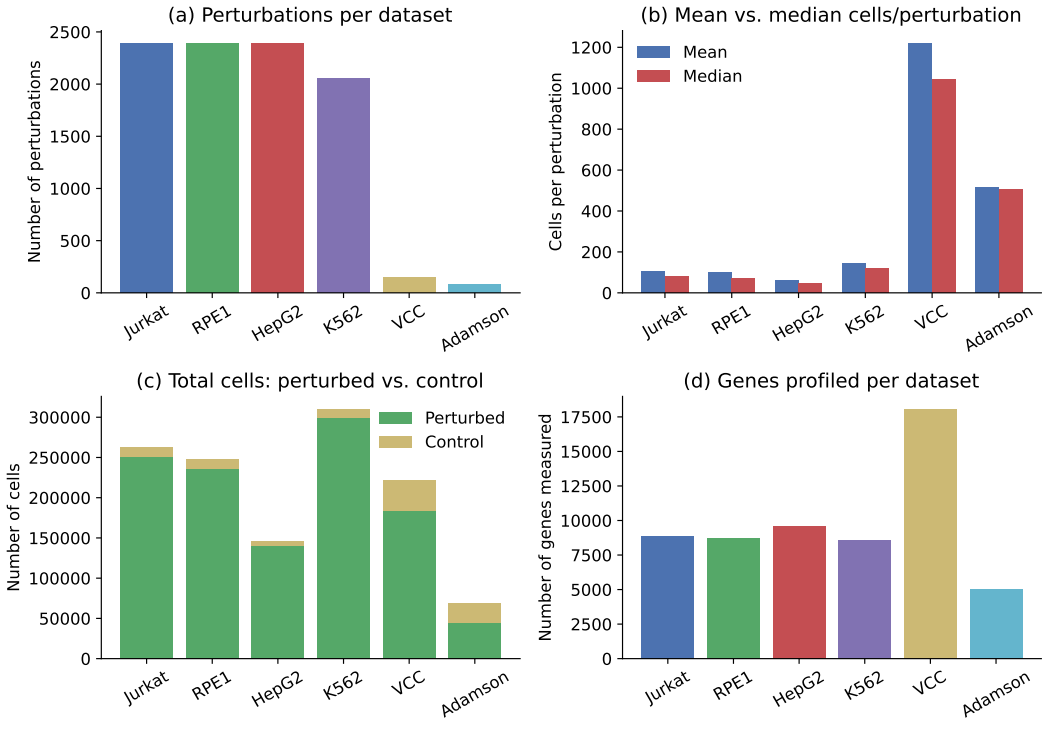


Figure 2: Overview of dataset characteristics. (a) Number of perturbations per dataset. (b) Mean vs. median cells per perturbation, illustrating the right-skewed coverage typical of single-cell perturbation screens. (c) Total cells per dataset, split into perturbed and non-targeting control cells. (d) Number of genes profiled per dataset.

A.2 Methods

In detail the methods we use are:

GEARS (Roohani et al., 2024): The neural network predicts an expression change and adds it to an input control cell's expression. At inference, it uses sampled control cells and averages their predicted post-perturbation profiles to obtain one expression profile per perturbation.

Ridge/RF-scf (Liu et al., 2026): Contextual scFoundation embeddings describe the perturbed gene. A ridge regression or a random forest model learns the difference between mean perturbed expression and mean control expression. Its predicted change is added to the control mean to reconstruct post-perturbation expression.

RF-GO (Csendes et al., 2025): A random forest maps GO features of the perturbed gene directly to mean post-perturbation expression. Its training targets are expression profiles, so there is no explicit control-subtraction and control-addition step.

Bilinear (Ahlmann-Eltze et al., 2025): The model learns expression changes relative to a baseline b, defined as the per-gene mean across training perturbation profiles. It fits the model to baseline-subtracted expression and reconstructs post-perturbation expression.

LPM (Miladinovic et al., 2025): The Large Perturbation Model is trained on mean expression profiles grouped by perturbation and context. The benchmark obtains predicted expression for each perturbation–gene pair and assembles these values into a post-perturbation profile, without an explicit control-addition step in the wrapper.

Average control (Viñas Torné et al., 2025): Every perturbation receives the mean expression profile of all the non-targeting control cells (not batch based). This represents an assumption of no perturbation-induced change before output transformation.

Average train (Viñas Torné et al., 2025): Every perturbation receives the mean expression profile across all perturbed training cells (not batch based). The prediction is identical for every test perturbation.

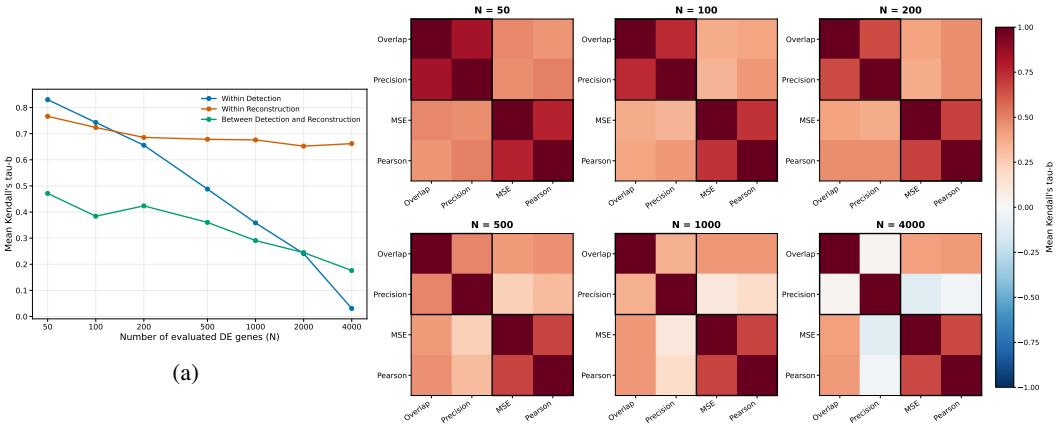
Before evaluation, the wrappers clip negative predicted expression to zero and apply log1p. Methods producing one profile per perturbation repeat that profile across the required prediction rows. For methods evaluation we use 5-fold cross-validation performed at the *perturbation* level rather than the cell level: unique target genes are randomly partitioned into five folds, so that all cells belonging to a given perturbation fall entirely within either the training or the test split. This tests each model's ability to generalize to genes it has not seen perturbed during training, rather than merely interpolating between cells of an already-observed perturbation. Depending on configuration, non-targeting control cells are either excluded from both splits and reserved solely for control-based baselines and evaluation, or retained in the training split only, never appearing in the test set. All splits use a fixed random seed for reproducibility. Beyond this common pipeline, a subset of models require additional, model-specific dataset adaptations applied at run time (e.g., GEARS and LPM expect the control label `ctrl` rather than `non-targeting`, among other format requirements); these adaptations are not part of the shared preprocessing described above.

B Changes in Metric Agreement

Changes in N affect not only model scores and rankings, but also the agreement among the metrics. We therefore rank the methods independently under each metric within every dataset–fold block and compute Kendall's τ_b between the resulting rankings. As shown in Figure 3a, agreement between Overlap and Precision decreases monotonically from 0.83 at $N = 50$ to 0.03 at $N = 4000$, whereas agreement between MSE and Pearson remains comparatively stable, decreasing only from 0.77 to 0.66. Agreement between detection and reconstruction also weakens as N increases. Figure 3b localizes this divergence to the detection metrics. In accordance with section 4.1.1, precision and overlap agree at $N = 50$ but, due to significant changes in precision, they gradually discord. Consequently, composite evaluation schemes that average heterogeneous metrics are sensitive to the choice of N and may instill bias towards one metric family over the other.

C Benchmark Heatmap

Fig. 4 summarizes the benchmark by showing, for every non-baseline model and evaluation metric, the rank achieved across the range of N values considered.



(a) (b)

Figure 3: (a) Agreement among metric-induced model rankings as a function of the number N of evaluated DE genes (mean Kendall's τ_b, averaged across folds within each dataset and then across datasets with equal weight). Detection= {Overlap,Precision}, reconstruction= { MSE and Pearson}. (b) Pairwise Kendall's τ_b between metric-induced model rankings at representative values of N. Black rectangles indicate metric families. Values are averaged across folds within each dataset and then across datasets with equal weight.

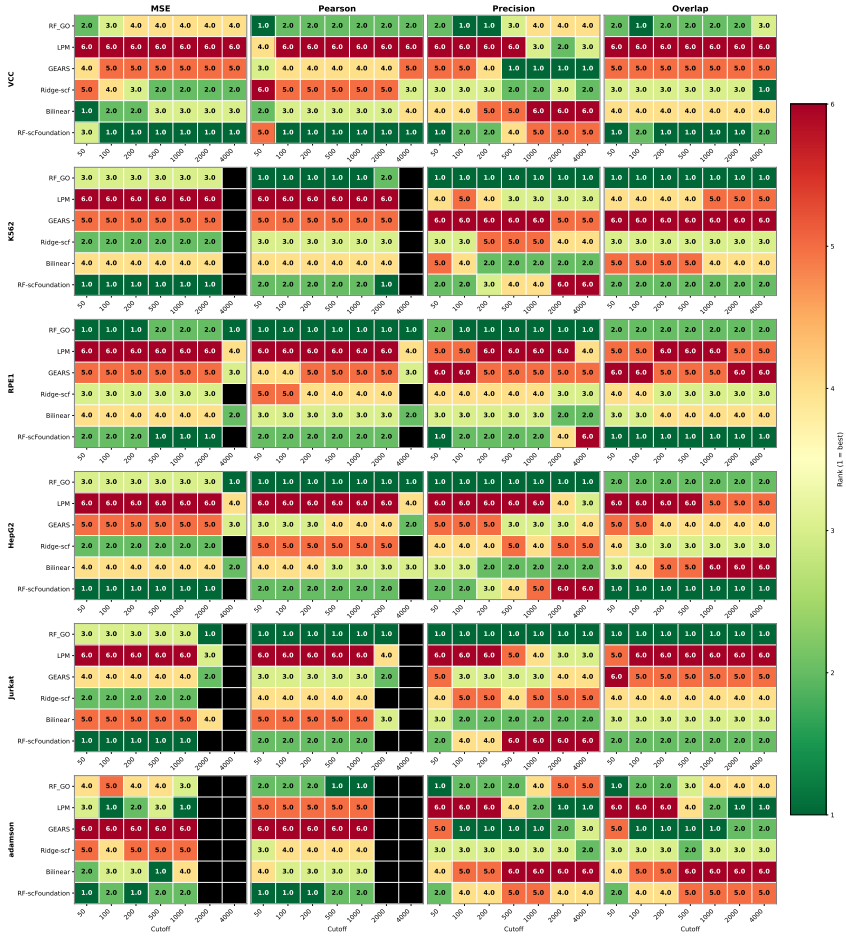


Figure 4: Heatmap showing the rank of each evaluated model. Lower ranks indicate better performance. Black cells indicate that no DE genes were available for the corresponding value of N.