WHY DO SOPHISTICATED PERTURBATION PREDICTION MODELS NOT OUTPERFORM LEARNING-FREE BASELINES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Why are modern perturbation-prediction models often not much better than simple learning-free baselines? This phenomenon has been widely observed in genetic perturbation benchmarks. However, its interpretation in chemical perturbation prediction remains unclear. Using sci-Plex3 as a primary case study, we show that simple pseudo-bulk-level baselines, together with two new single-cell-level learning-free baselines, remain highly competitive, even against modern diffusion-based models. We then propose a three-axis diagnostic framework comprising Input Use, Response Scaling, and Stratified Performance, which evaluates whether perturbation information is used, whether predicted responses scale with the strength of the observed cellular responses, and where along the perturbation effect distribution models succeed or fail. Applying the framework to ChemCPA, CondOT, Doloris, and PerturbDiff reveals that failure to outperform learning-free baselines can arise from two distinct causes: limited perturbation signal in the data or limited utilization of that signal by the model. Independent evaluation on Tahoe-100M further confirms that the competitiveness of learning-free baselines is dataset- and metric-dependent: ChemCPA remains below the perturbation reference, while PerturbDiff exceeds it on all pseudo-bulk metrics but not consistently on distributional metrics. Our results suggest that comparing performance against learning-free baselines alone is not a reliable indicator of model quality. Instead, our diagnostic analyses provide mechanistic insights into why models succeed or fail, help distinguish data limitations from model limitations, and guide the development of more robust perturbation-prediction models.

1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) perturbation screens have emerged as a powerful tool for studying how cells respond to genetic and chemical interventions (Ahlmann-Eltze et al., 2025; Wang et al., 2026; Srivatsan et al., 2020). By measuring transcriptomic changes at cellular resolution, these assays enable systematic characterization of drug mechanisms, target engagement, and cellular heterogeneity, while supporting large-scale *in silico* screening to reduce experimental costs and accelerating therapeutic discovery.

Driven by the increasing availability of large-scale Perturb-seq (Wang et al., 2026; Csendes et al., 2025) and chemical screening datasets Zhang et al. (2025); Srivatsan et al. (2020), recent years have witnessed rapid advances in single-cell perturbation prediction models. Existing methods span a wide range of modeling paradigms, including latent-variable approaches such as scGen (Lotfollahi et al., 2019), CPA (Lotfollahi et al., 2023), BioLord (Piran et al., 2024), graph-based methods such as GEARS (Roohani et al., 2023), and more recent foundation-model (Cui et al., 2024; Adduri et al., 2025) and diffusion-based architectures (Chi et al., 2026; Yuan et al., 2026). Despite increasing model complexity, recent benchmarking studies suggest that sophisticated models do not necessarily improve predictive performance. Counter-intuitively, **sophisticatedly designed state-of-the-art methods often fail to match or outperform simple learning-free and linear baselines**, especially in genetic perturbation settings (Vinas Torne et al., 2025; Ahlmann-Eltze et al., 2025).

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

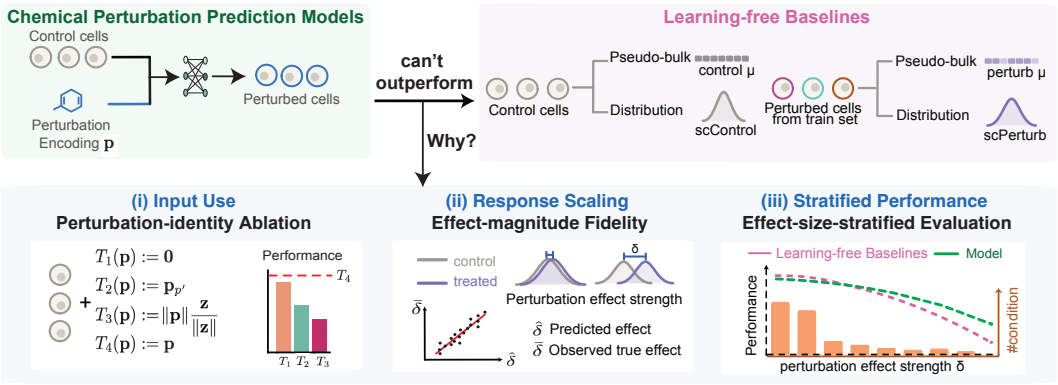


Figure 1: **Overview.** Chemical perturbation prediction models take control cells and a queried drug as input and predict perturbed gene expression profiles. Despite their sophistication, they tend to fail to outperform simple learning-free baselines. We propose three diagnostics to explain why: (i) **Input Use**: does the model use perturbation identity? (ii) **Response Scaling**: does it adapt to the strength of the observed cellular responses? (iii) **Stratified Performance**: when and where does perturbation utilization translate into predictive gains? Together, these complementary diagnostics disentangle perturbation utilization, effect-magnitude fidelity, and effect-strength-dependent performance.

The strength of simple baselines has motivated a parallel line of work on evaluation, with several studies documenting the sensitivity of model rankings to metric choice, feature selection, and benchmark design, thus calling for more principled protocols (Wu et al., 2024; Kernfeld et al., 2025; Schäfer et al., 2026; Vinas Torne et al., 2025; Mejia et al., 2026). **However, existing benchmarks leave an important ambiguity unresolved: *why and where* does this failure occur?** When a model fails to outperform simple learning-free baselines, it is still unclear whether this reflects (i) a genuinely weak perturbation signal in the data itself such that learning-free baselines already suffice, (ii) a true modeling failure, in which the signal is present but the model cannot exploit it, or (iii) aggregated metrics that, by summarizing performance across genes, cells, and perturbation conditions, are too coarse to reveal the biological information the model has captured. Distinguishing between these possibilities is essential for interpreting benchmark outcomes and guiding model development.

This ambiguity matters especially in chemical perturbation. Unlike genetic perturbations, which directly alter specific genes, chemical perturbations may act post-transcriptionally, operate below an effective concentration, or affect pathways that scRNA-seq resolves weakly (Srivatsan et al., 2020), so it may induce little detectable transcriptomic response despite having a genuine biological effect. Consequently, where that response is small, the perturbed profile stays close to the control cells, and a control-like prediction will be nearly perfect. Failing to outperform a learning-free baseline here is not necessarily an evidence of a failed model, since the data may leave little room to outperform.

To address this limitation, (i) we first show that strong learning-free pseudo-bulk baselines, together with two new single-cell baselines SCCONTROL and SCPERTURB remain highly competitive compared to modern models including ChemCPA (Hetzel et al., 2022), CondOT (Bunne et al., 2022), Doloris (Chi et al., 2026), and PerturbDiff (Yuan et al., 2026) using sci-Plex3 as a primary case study. (ii) We introduce a three-axis diagnostic framework comprising *Input Use*, *Response Scaling*, and *Stratified Performance* that moves beyond asking whether a model beats a baseline. We further reveal whether perturbation information is used, whether its prediction adapts to the strength of the observed cellular response, and under which perturbation regimes model failures occur. Together, these diagnostics disentangle two distinct causes of comparable performance relative to learning-free baselines: *limited perturbation signal in the data* and *limited utilization of that signal by the model*. We further confirm the need to interpret baseline comparisons in their dataset and evaluation context using an independent evaluation on Tahoe-100M. (iii) we show that simple interventions driven by the diagnostics can lead to substantial performance improvement on ChemCPA. Together, our results move perturbation prediction evaluation beyond aggregate benchmark scores and toward a more mechanistic understanding of model behavior.

Contributions.

- **Establish distribution-level strong learning-free baselines for chemical perturbation prediction.** Beyond pseudo-bulk baselines, we introduce two single-cell counterparts, SCCONTROL and SCPERTURB, enabling distribution-level evaluation of cell population. These learning-free baselines remain highly competitive against modern models on sci-Plex3 and Tahoe-100M.
- **Introduce a three-axis diagnostic framework.** We propose Input Use, Response Scaling, and Stratified Performance diagnostics to determine whether perturbation information is used, whether model predictions adapt to the strength of the observed cellular responses, and where along the perturbation effect distribution models succeed or fail.
- **Provide a mechanistic explanation for strong baseline performance.** Our diagnostic framework locates that baseline competitiveness depends on the dataset, model, and evaluation metric. A failure to outperform learning-free baselines can arise from a data limitation and/or a model limitation. **Accordingly, a model may appear only marginally better in aggregate while substantially outperforming baselines on perturbations with strong effect when it effectively utilizes available perturbation signal.**
- **Demonstrate that diagnostics can guide model improvement.** Using ChemCPA as an example, we show that diagnostic-motivated interventions substantially improve predictive performance.

Related Work Benchmarking studies have repeatedly found that simple linear models, control-based references, and other perturbation-independent predictors remain competitive with far more complex architectures (Ahlmann-Eltze et al., 2025; Wu et al., 2024; Li et al., 2024; Wei et al., 2026; Csendes et al., 2025; Bendidi et al., 2024), and a parallel line of work has traced how sensitive model rankings are to metric choice, feature selection, and benchmark design (Kernfeld et al., 2025; Schäfer et al., 2026; Vinas Torne et al., 2025; Mejia et al., 2026). Two limitations run through this literature. First, it is descriptive: it establishes that a model *does or does not* beat a baseline, but not necessarily a mechanistic understanding of *why*. Second, most of its evidence comes from genetic perturbation, where an intervention is expected to produce a measurable transcriptomic response; chemical screens violate that assumption often enough that the same comparison carries a different meaning. We address both limitations in this work. A more in-depth discussion of related work is given in Appendix A.

2 COMPETITIVE LEARNING-FREE BASELINES IN CHEMICAL PERTURBATION PREDICTION

We focus on single-cell chemical perturbation prediction. Given a chemical perturbation dataset, let $\mathbf{x}_i^{(c_i,p_i,s_i)} \in \mathbb{R}^n$ denote the observed expression profile of cell i, where $c_i \in \mathcal{C}$ is its cell line, $p_i \in \{\text{ctrl}\} \cup \mathcal{D}$ is the perturbation, and $s_i \in \mathbb{R}_{\geq 0}$ is the dosage. Control cells satisfy $p_i = \text{ctrl}$ and $s_i = 0$. A perturbation prediction model is trained to map a control cell together with a target condition (c, p, s) to a predicted perturbed profile $\hat{\mathbf{x}} = f_\theta\big(\mathbf{x}^{(c,\text{ctrl},0)}, c, p, s\big)$ so that it can predict the transcriptional state that the control cell would attain under an unseen perturbation condition shift.

Perturbation prediction models are commonly evaluated against two learning-free baselines derived from training-set population statistics (Kernfeld et al., 2025; Adduri et al., 2025; Vinas Torne et al., 2025; Miller et al., 2025): cell–line specific control mean $\boldsymbol{\mu}_{c,\text{ctrl},0}$ (average control cell expression from training set) and cell line–specific perturbation mean $\boldsymbol{\mu}_{c,\mathcal{D},S}$ (average over all perturbed cells from the training set, ignoring the query perturbation). Both baselines are meaningful because control mean implies that all perturbations have no effect, and perturbation mean implies that all perturbations have the same effect (Miller et al., 2025).

However, these baselines only characterize average responses and cannot capture cell population distribution. We therefore further propose single-cell counterparts: SCCONTROL and SCPERTURB, which sample control or perturbed cells from the training data of the target cell line and assume similar biological effect as cell line–specific control and perturb mean. By operating directly on observed cells, they preserve the underlying distributional structure and enable metrics evaluation with E-distance (Peidli et al., 2024) and EMD (Chi et al., 2026). We study the robustness of SCCONTROL and SCPERTURB with respect to sample size (Appendix F). The point-estimate and single-cell variants are the same baselines viewed at two resolutions; we use whichever form a metric admits.

Table 1: **Unseen-drug generalization on sci-Plex3.** Results cover 360 conditions across 30 unseen drugs; mean ± standard deviation across conditions. Pseudo-bulk metrics evaluate point estimates, whereas single-cell metrics evaluate predicted distributions. **Bold** indicates the best mean per column.

	Pseudo-bulk ↑				*Single-cell* ↓			
Method	Full r	Full R^2	DEG50 r	DEG50 R^2	Full ED	Full EMD	DEG50 ED	DEG50 EMD
Learning-based models								
ChemCPA	$0.953_{\pm 0.077}$	$0.896_{\pm 0.164}$	$0.894_{\pm 0.207}$	$0.739_{\pm 0.529}$	$2.084_{\pm 0.220}$	$0.065_{\pm 0.010}$	$1.169_{\pm 0.313}$	$0.212_{\pm 0.101}$
CondOT	$0.962_{\pm 0.071}$	$0.916_{\pm 0.161}$	$0.901_{\pm 0.205}$	$0.739_{\pm 0.604}$	$1.241_{\pm 0.308}$	$0.060_{\pm 0.010}$	$0.678_{\pm 0.353}$	$0.166_{\pm 0.091}$
Doloris	$0.938_{\pm 0.068}$	$0.874_{\pm 0.129}$	$0.893_{\pm 0.171}$	$0.769_{\pm 0.337}$	$1.094_{\pm 0.514}$	$0.027_{\pm 0.006}$	$0.908_{\pm 0.663}$	$0.099_{\pm 0.073}$
PerturbDiff	$0.947_{\pm 0.061}$	$0.866_{\pm 0.145}$	$\mathbf{0.926}_{\pm 0.111}$	$\mathbf{0.808}_{\pm 0.288}$	$0.537_{\pm 0.354}$	$0.051_{\pm 0.007}$	$\mathbf{0.255}_{\pm 0.277}$	$0.091_{\pm 0.052}$
Learning-free baselines								
Control mean	$0.961_{\pm 0.080}$	$0.915_{\pm 0.181}$	$0.894_{\pm 0.220}$	$0.719_{\pm 0.628}$	$0.410_{\pm 0.286}$	$\mathbf{0.016}_{\pm 0.008}$	$0.408_{\pm 0.387}$	$0.091_{\pm 0.102}$
Perturbation mean	$\mathbf{0.963}_{\pm 0.077}$	$\mathbf{0.920}_{\pm 0.170}$	$0.901_{\pm 0.213}$	$0.747_{\pm 0.580}$	$\mathbf{0.403}_{\pm 0.275}$	$\mathbf{0.016}_{\pm 0.008}$	$0.386_{\pm 0.375}$	$\mathbf{0.086}_{\pm 0.098}$

A Case Study on sci-Plex3. To validate whether learning-free baselines can outperform complicated models in chemical perturbation prediction, we use sci-Plex3 (Srivatsan et al., 2020), which is one of the most widely used chemical perturbation prediction dataset (Hetzel et al., 2022; Chi et al., 2026; Piran et al., 2024; Yu et al., 2025). We compare four well-established representative models: ChemCPA (Hetzel et al., 2022) predicts single-cell gene expression responses by learning disentangled representations of cellular states and perturbations via auto-encoders. PerturbDiff (Yuan et al., 2026) and Doloris (Chi et al., 2026) represent more recent diffusion-based approaches that explicitly model the distribution of perturbed cells and have shown strong performance on chemical perturbation benchmarks. CondOT Bunne et al. (2022) learns context-conditioned optimal transport maps via partially input convex networks, predicting cell responses to unseen treatment combinations. We do not include foundation-model-based methods in this comparison, as recent task-specific models have often been shown to outperform them (Chi et al., 2026; Yuan et al., 2026; Adduri et al., 2025).

Across pseudo-bulk and distributional evaluations, Table. 1 shows that the conclusions from recent genetic–perturbation studies largely carry over to the chemical setting. While PerturbDiff consistently achieves the strongest performance among the evaluated models, improvements over learning-free baselines are often modest and depend strongly on the evaluation metric and the evaluated gene panel. However, this observation alone is difficult to interpret. Small gains may result from weak perturbation-specific signal, substantial intrinsic cellular heterogeneity, or genuine limitations of predictive models. Since these factors are conflated in aggregate performance metrics, we next explain the performance gap using a three-axis diagnostic framework.

3 Reasons Behind the Gap: A Diagnostic Framework

We propose a three-axis diagnostic framework: *Input Use*, *Response Scaling*, and *Stratified Performance*. These diagnostics are complementary and they answer that: given a model, (i) whether perturbation information is used (Input Use), (ii) whether its predicted responses properly scale with the strength of the observed cellular responses (Response Scaling), and (iii) where along the biological response spectrum its resulting predictions succeed or fail (Stratified Performance).

3.1 Input Use: Is the perturbation identity being used?

The first diagnostic asks whether perturbation information contributes to the prediction at all. This is essential for perturbation responses modeling: a model whose predictions are invariant to the queried perturbation cannot outperform a perturbation-independent baseline except through incidental dataset structure. We therefore evaluate input use by corrupting perturbation information at inference time as Eq. 1 and measuring the resulting performance degradation.

$$
\begin{aligned}
&T_1(\mathbf{p}) := \mathbf{0} && \text{no signal, no scale,}\\
&T_2(\mathbf{p}) := \mathbf{p}_{p'}, \quad p' \neq p && \text{wrong identity, real embedding,}\\
&T_3(\mathbf{p}) := \|\mathbf{p}\| \frac{\mathbf{z}}{\|\mathbf{z}\|}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, I) && \text{random direction with matched norm,}\\
&T_4(\mathbf{p}) := \mathbf{p} && \text{uncorrupted reference.}
\end{aligned}
\tag{1}
$$

Among the corruptions, T_2 is the primary diagnostic because it preserves a valid compound embedding while changing only the compound identity. By contrast, T_1 and T_3 move the input off the embedding manifold and are therefore interpreted as robustness checks. If performance is nearly unchanged under corruption relative to T_4, the model behaves largely independently of the queried perturbation.

3.2 Response Scaling: Do Predictions Scale with Observed Responses?

Using perturbation information is a necessary condition for perturbation prediction, but it is not sufficient. A model may be highly sensitive to perturbation identity yet still generate responses that fail to reflect the underlying biological signal. We therefore introduce a second diagnostic that measures whether predicted drug effect scale with the true drug effect induced by perturbations. We first formally define the drug effect in Def. 1:

Definition 1 (Pseudo-bulk perturbation effect)**.** *For a condition (c, p, s) with pseudo-bulk mean $\boldsymbol{\mu}_{c,p,s}$, the pseudo-bulk perturbation effect $\bar{\delta}_{c,p,s} \in \mathbb{R}$ is the scalar magnitude of the observed gene-expression shift induced by perturbation p and dose s relative to $\mathrm{r}_{c,s}$,*

$$\bar{\delta}_{c,p,s} \;:=\; \left\|\boldsymbol{\mu}_{c,p,s} - \mathrm{r}_{c,s}\right\|_2 . \tag{2}$$

In our experiments, $\mathrm{r}_{c,s} = \boldsymbol{\mu}_{c,\mathrm{ctrl},0}$. Biologically, $\bar{\delta}_{c,p,s}$ quantifies the perturbation induced expression shift relative to control cells. Equivalently, it is the pseudo-bulk error incurred by predicting the control mean for condition (c, p, s), capturing the information lost when the perturbation is ignored. By construction, it is zero for control condition. Small values indicate that a constant control-mean predictor can suffice, whereas large values highlight conditions where perturbation information is essential.

Definition 2 (Response scaling)**.** *Let $\hat{\boldsymbol{\mu}}_{c,p,s}$ be the predicted mean profile for condition (c, p, s), and $\hat{\delta}$ be the predicted pseudo-bulk perturbation effect against the same reference as Def. 1, and let β be the slope of the predicted effect regressed on the observed effect over a set of evaluated conditions $\mathcal{Q}$,*

$$\beta \;:=\; \frac{\mathrm{Cov}_{\mathcal{Q}}\big(\bar{\delta},\, \hat{\delta}\big)}{\mathrm{Var}_{\mathcal{Q}}\big(\bar{\delta}\big)} . \tag{3}$$

The coefficient β measures how predicted perturbation magnitudes scale with observed perturbation magnitudes across conditions. A perfect prediction yields $\beta = 1$, although $\beta = 1$ does not necessarily guarantee a perfect prediction. Values $0 < \beta < 1$ indicate response shrinkage, whereas $\beta > 1$ indicates amplification.$\beta = 0$ indicates no linear association with the observed effects, including but is not limited to collapse to the reference, while $\beta < 0$ indicates an inverse association.

Response scaling provides a global summary of how predicted and observed effect magnitudes covary across conditions. Accordingly, response shrinkage ($0 < \beta < 1$) suggests a weaker agreement between predicted and observed perturbation magnitudes (effect). However, it does not reveal where along the effect-size distribution these errors occur. We therefore turn to a third diagnostic that stratifies performance by perturbation effect.

3.3 Stratified Performance: where along the effect-size distribution dose the model succeed?

Input Use and Response Scaling evaluate whether perturbation information is used and scaled appropriately, but they do not reveal where performance is concentrated across the effect-size distribution. This is particularly important in chemical perturbation prediction, where many compounds can induce only weak responses. As a result, models can achieve competitive aggregate performance while still failing on large-effect perturbations.

To expose such behavior, we stratify evaluation by the observed pseudo-bulk perturbation effect $\bar{\delta}_{c,p,s}$ defined in Definition 1. Specifically, we partition evaluated conditions into bins according to $\bar{\delta}_{c,p,s}$ and report performance separately within each bin. A useful perturbation model should derive increasing benefit from perturbation information as $\bar{\delta}_{c,p,s}$ increases, since these are precisely the conditions on which perturbation-independent predictors incur the largest error.

4 EXPERIMENTAL SETUP

We use sci-Plex3 dataset (Srivatsan et al., 2020),which comprises single-cell gene expression profiles from three human cancer cell lines (A549, MCF7, K562) treated under $2,247$ conditions involving 188 distinct single compounds. Data preprocessing follows ChemCPA (Hetzel et al., 2022); details are provided in Appendix B.1. We report both pseudo-bulk and distribution-level performance including: R^2, RMSE, and Pearson correlation between predicted and observed gene expression profile mean (Hetzel et al., 2022; Adduri et al., 2025), and E-distance and EMD between predicted and observed single-cell populations, on both the full $2,000$-gene panel and the compound-specific top 50 differentially expressed genes (DEG50), selected by comparing observed treated and control cells.We use four representative models: ChemCPA (Hetzel et al., 2022), CondOT Bunne et al. (2022), Doloris (Chi et al., 2026), and PerturbDiff (Yuan et al., 2026). Unless otherwise stated, we use the official hyperparameter settings for sci-Plex3; implementation details are provided in Appendix B.2. We consider a **drug OOD** setting in which all test compounds are unseen during training. Data are split by drug identity (80/20) using 5-fold stratified cross-validation. Results from one random fold are shown in the main text, while complete five-fold results are provided in Appendix E.2 E.3.

5 RESULTS

Overall, our results show that **a small aggregate performance gap does not necessarily imply a weak model.** On sci-Plex3, the highly right-skewed distribution of perturbation effects means that most perturbations exhibit weak responses, causing aggregate metrics to be dominated by low-signal conditions and potentially underestimating model quality. This explains why models such as PerturbDiff and Doloris show only modest improvements on average despite achieving substantially larger gains on strong-response perturbations. At the same time, our diagnostics reveal genuine model-specific shortcomings: ChemCPA and CondOT exhibit weak corruption sensitivity, near-zero response scaling, and poor performance even on strong-effect perturbations, indicating limited utilization of available perturbation information. Importantly, these findings are **actionable**. Using ChemCPA as an example, we show that simple interventions targeting the identified data and model limitations lead to improved performance on sci-Plex3, demonstrating that the diagnostic framework not only explains performance bottlenecks but also guides practical model improvements.

5.1 INPUT USE: DO MODELS RELY ON PERTURBATION IDENTITY?

Table 2: **Inference-time perturbation corruption.** DEG50 R^2 under the uncorrupted embedding and the performance drop after replacing the drug embedding with zeros, another drug, or norm-matched noise. Larger negative drops indicate stronger dependence on perturbation identity. **Bold** indicates the highest uncorrupted mean and the largest drop in each corruption column.

	Reference	*Corruption*: ΔR^2 vs. uncorrupted		
Model	Uncorrupted (T_4) $R^2 \uparrow$	Zero (T_1)	Shuffled (T_2)	Noise (T_3)
ChemCPA	$0.739_{\pm 0.529}$	-0.023	-0.027	$+0.003$
CondOT	$0.739_{\pm 0.604}$	-0.003	0.000	-0.003
Doloris	$0.769_{\pm 0.337}$	-0.037	-0.047	-0.037
PerturbDiff	$\mathbf{0.808}_{\pm 0.288}$	$\mathbf{-0.273}$	$\mathbf{-0.143}$	$\mathbf{-0.343}$

We first evaluate **Input Use** through inference-time perturbation corruption (Table 2) to assess whether the perturbation information is *being used*. Since perturbation-specific predictions should depend on the queried compound, corrupting its embedding should degrade performance. If it remains unchanged, the model may under-use perturbation information, and/or naively rely on dataset-level and cell-line specific priors. Under this diagnostic, we observe substantial differences across models.

ChemCPA and CondOT show almost no performance degradation under corruption, indicating weak reliance on perturbation identity. The performance is even comparable when we replace the true perturbation embedding with random noise (T_3), suggesting that the model makes limited use of perturbation information. For ChemCPA specifically, this interpretation is further supported by the representation-norm analysis in Appendix C.1. ChemCPA encodes each cell into a basal state latent representation before combining it with the perturbation and covariate embeddings in the decoder. This basal-state latent has a norm approximately $25\times$ larger than the perturbation embedding, suggesting that perturbation information contributes only weakly to ChemCPA's decoder.

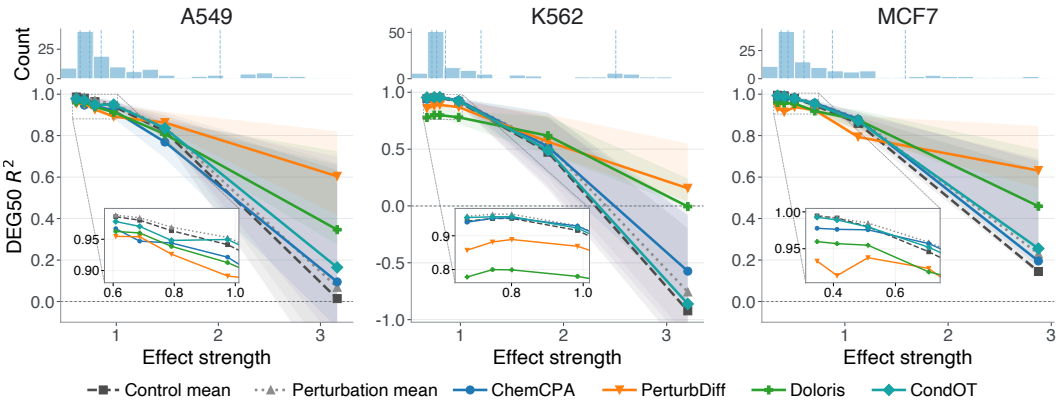


Figure 2: sci-Plex3 stratification analysis of the drug effect with respect to R^2 on held-out drugs ($n = 360$ conditions; 120 per cell line). Columns correspond to the three cell lines (A549, K562, MCF7). Conditions are split into six equal-frequency bins by pseudo-bulk effect size δ; lines show the mean DEG50 R^2 per bin, shaded regions the interquartile range, and the top row the number of conditions along the effect-size axis. The zoom-in window enlarges the weak-effect region per panel. CondOT, ChemCPA, Doloris and PerturbDiff are shown alongside the learning-free baselines.

Having established whether perturbation identity is used, we next ask whether this information is translated into appropriately scaled responses.

5.2 Response Scaling: Do predicted perturbation effect scale with induced effect?

Table 3 reports the response-scaling coefficient β, which can quantify whether the dependence on the perturbation identity can translate into appropriately scaled responses. A model can exhibit strong Input Use yet still attain a small β if its predictions are compressed toward the control state.

Table 3: **Response-shrinkage analysis**: Perfect agreement between predicted and observed perturbation effect magnitudes yields $\beta = 1$, while $0 < \beta < 1$ indicates response shrinkage. Values closer to 1 indicate better agreement.

Model	β_{Full}	β_{DEG50}
ChemCPA	0.031	0.078
CondOT	0.043	0.071
Doloris	0.341	0.344
PerturbDiff	**0.615**	**0.721**

We observe the same pattern as above. ChemCPA and CondOT attain a β value close to zero, indicating that predicted response magnitudes are largely independent of the observed perturbation effect. This is consistent with the corruption results, where ChemCPA and CondOT showed only minimal sensitivity to perturbation-embedding corruption. Moreover, Doloris occupies an intermediate regime. PerturbDiff attains the largest β, suggesting that it not only uses perturbation information, but also modulates prediction magnitude according to the strength of the induced response. Nevertheless, even the largest β remains below 1, suggesting that response shrinkage is prevalent across all evaluated models. Taken together, the Input Use and Response Scaling diagnostics paint a consistent picture of increasing perturbation utilization from CondOT and ChemCPA to Doloris and finally PerturbDiff.

Moreover, β is consistently higher on DEG50 than on the full gene panel, indicating stronger agreement between predicted and observed effect magnitudes among perturbation-responsive genes. This may reflect reduced signal dilution from weakly affected genes (Mejia et al., 2026). We next examine where along the effect-size distribution these gains arise.

5.3 Stratified Performance: Where do models succeed and fail?

We evaluate Stratified Performance by reporting DEG50 R^2 as a function of perturbation effect size $\bar{\delta}^{c,p,s}$ per cell line (Fig. 2). For all test conditions ($n = 360$), we first compute $\bar{\delta}^{c,p,s}$ and examine its distribution (Fig. 2, top histograms). **The results reveal a strong right-skewed effect distribution dominated by perturbations with weak effect.** Conditions with large transcriptomic effect constitute only a small fraction of the benchmark. To determine whether the observed distribution is specific to

Table 4: **Unseen-drug generalization on Tahoe-100M.** Results cover 76 unseen drugs across A549, RKO, and Hs766T. Pseudo-bulk metrics evaluate point estimates, whereas single-cell metrics evaluate predicted distributions. **Bold** indicates the best mean in each column; standard deviations are shown as subscripts.

	Pseudo-bulk ↑				*Single-cell* ↓			
Method	Full r	Full R^2	DEG50 r	DEG50 R^2	Full ED	Full EMD	DEG50 ED	DEG50 EMD
Learning-based models								
ChemCPA	$0.976_{\pm 0.036}$	$0.939_{\pm 0.077}$	$0.946_{\pm 0.108}$	$0.843_{\pm 0.339}$	$4.307_{\pm 0.296}$	$0.258_{\pm 0.046}$	$2.719_{\pm 0.191}$	$1.443_{\pm 0.214}$
PerturbDiff	$\mathbf{0.985}_{\pm 0.032}$	$\mathbf{0.961}_{\pm 0.070}$	$\mathbf{0.962}_{\pm 0.091}$	$\mathbf{0.894}_{\pm 0.287}$	$\mathbf{0.724}_{\pm 0.446}$	$0.096_{\pm 0.023}$	$0.677_{\pm 0.291}$	$0.623_{\pm 0.139}$
Learning-free baselines								
Control mean	$0.979_{\pm 0.037}$	$0.943_{\pm 0.106}$	$0.942_{\pm 0.113}$	$0.823_{\pm 0.554}$	$0.810_{\pm 0.462}$	$0.045_{\pm 0.024}$	$0.705_{\pm 0.369}$	$0.325_{\pm 0.196}$
Perturbation mean	$0.982_{\pm 0.037}$	$0.953_{\pm 0.094}$	$0.952_{\pm 0.109}$	$0.859_{\pm 0.470}$	$0.766_{\pm 0.460}$	$\mathbf{0.043}_{\pm 0.024}$	$\mathbf{0.629}_{\pm 0.359}$	$\mathbf{0.282}_{\pm 0.187}$

the test set, we repeat the same analysis on all conditions in the dataset and observe a similarly heavy right-skewed distribution (Appendix D.2), indicating that this pattern is not a test-set artifact.

We next stratify conditions by perturbation effect (Def. 1) size to examine how model performance varies across response regimes (Fig. 2, Appendix D.1). As expected, the control-mean and perturbation-mean baselines are competitive for weak-effect perturbations but degrade rapidly as effect size grows. More importantly, model differences become increasingly pronounced in the high-effect regime. Although aggregate metrics suggest only modest gains, PerturbDiff and Doloris consistently outperform learning-free baselines on strong-response perturbations. Given the highly right-skewed distribution of perturbation effects, where most perturbations induce relatively weak responses, the small gap in PerturbDiff and Doloris against the learning-free baselines is more indicative of limited signal across much of the dataset than of limitations of the models themselves. In contrast, ChemCPA and CondOT perform poorly even on strong-response perturbations. Together with their weak corruption sensitivity and near-zero response scaling, this pattern points to limited utilization of available perturbation information rather than dataset limitations alone.

5.4 Tahoe-100M: a less skewed effect distribution

On sci-Plex3, PerturbDiff makes the strongest use of perturbation information among the evaluated models (Tables 2 and 3), yet cannot outperform learning-free baselines on every metric. Our stratified analysis (Section 3.3) attributes this largely to the right-skewed effect distribution: its gains are concentrated in the few strong-effect conditions, which aggregate metrics underweight. If this holds, PerturbDiff's advantage should become more visible on a dataset in which drug-effect is more uniformly distributed. We test this on Tahoe-100M (Zhang et al., 2025), whose effect distribution is less right-skewed than that of sci-Plex3 (Appendix G), evaluating ChemCPA and PerturbDiff on 76 unseen drugs across 3 cell-lines (A549, Hs766T, RKO) that cover 3 organs (Table 4). ChemCPA remains below the perturbation-mean baseline on every metric, as on sci-Plex3. PerturbDiff exceeds it on all four pseudo-bulk metrics and on full-panel E-distance, with the largest gain in DEG50 R^2, but not on the remaining three distributional metrics. Whether a model outperforms learning-free baselines therefore depends on both the dataset and the model: on Tahoe-100M, the model that makes use of perturbation information on sci-Plex3 is the one that outperforms them on pseudo-bulk metrics. Together, these results reinforce the need to interpret baseline comparisons alongside diagnostics of how models use perturbation information

5.5 Diagnostic-Guided Interventions

Following the two key reasons our proposed diagnostic framework revealed, we advocate for two design principles that future models should consider. **First, model design should account for potential heavy tail in the perturbation-effect distribution.** In sci-Plex3, strong-effect perturbations lie in the tail of the distribution and are often represented by fewer cells due to cytotoxicity (Srivatsan et al., 2020). As a result, perturbations carrying the strongest signal may contribute disproportionately little to training. More generally, optimization procedures should avoid allocating gradient updates solely according to cell frequency when perturbation effects are highly imbalanced.

Second, perturbation-specific signals should be explicitly encouraged during training. In many perturbation-prediction architectures, the perturbation-induced response constitutes only a

Table 5: **Effect of the proposed interventions on perturbation utilization in ChemCPA.** Corruption sensitivity is measured by the change in DEG50 R^2 after replacing the correct drug embedding with zeros (T_1), another drug's embedding (T_2), or norm-matched noise (T_3); more negative drops indicate stronger dependence on perturbation identity. β measures how predicted effect magnitudes scale with observed effects. The last two columns report DEG50 R^2 in the weakest and strongest of six equal-count effect-size bins, pooled over the three cell lines. **Bold** indicates the best mean in each column; standard deviations are shown as subscripts.

	Input Use				*Response Scaling*		*Effect-size stratification*	
Model	Correct $R^2 \uparrow$	$\Delta R^2_{T_1}$	$\Delta R^2_{T_2}$	$\Delta R^2_{T_3}$	β_{Full}	β_{DEG50}	Weakest bin ↑	Strongest bin ↑
Vanilla ChemCPA	$0.739_{\pm 0.527}$	-0.023	-0.027	$+0.003$	0.031	0.078	**0.962**	-0.095
ChemCPA (condition-balanced)	$\mathbf{0.810}_{\pm 0.344}$	-0.078	-0.083	-0.050	0.292	0.346	0.960	0.248
ChemCPA (condition-bal. + dir. loss)	$0.772_{\pm 0.280}$	$\mathbf{-0.160}$	$\mathbf{-0.126}$	$\mathbf{-0.236}$	**0.467**	**0.513**	0.892	**0.356**

small residual on top of much larger shared cellular and contextual components (Jiang et al., 2026). Reconstruction objectives alone may therefore be dominated by these shared components, allowing models to achieve good reconstruction accuracy while under-utilizing perturbation information. Explicit supervision of perturbation-specific responses can help mitigate this failure mode.

To illustrate these principles, we implement two simple modifications to ChemCPA. The first replaces cell-level sampling with condition-balanced sampling. The second augments the reconstruction objective with an auxiliary loss that directly supervises the perturbation-induced residual. These modifications yield **ChemCPA (condition-balanced)** and **ChemCPA (condition-bal. + dir. loss)**, respectively. Additional implementation details are provided in Appendix B.3.1 and Appendix B.3.2 with hyperparameter sensitivity study in Appendix B.3.3.

Consistent with the proposed principles, both variants exhibit substantially stronger perturbation utilization than the vanilla model (Table 5). Corruption sensitivity increases across all three perturbation-corruption tests, while response scaling improves from $\beta_{\text{DEG50}} = 0.078$ to 0.346 and 0.513, respectively. The largest gains occur in the strongest-effect perturbation regime, where performance increases from -0.095 to 0.248 and 0.356. Notably, condition-balanced sampling alone already produces substantial improvements, suggesting that perturbation-specific signal, though subtle for the majority of perturbations, is nonetheless present and learnable. While these interventions are implemented in ChemCPA, their purpose is not to establish a new state-of-the-art model. Rather, they serve as a positive control demonstrating that the proposed diagnostics can identify actionable failure modes and guide improvements in perturbation utilization. Full results in Appendix E.1.

Additional Results. Our appendix provides experimental details and preprocessing (Appendix B); additional evidence for perturbation input use, showing a substantial scale imbalance between basal-state and perturbation representations in ChemCPA (Appendix C); extended response-scaling and stratified analyses, including perturbation effect distributions (Appendix D); complete benchmarks and ChemCPA intervention studies supporting the main conclusions (Appendix E); and robustness analyses of learning-free baselines under varying population sizes (Appendix F).

6 Conclusion

Learning-free baselines are often difficult to beat in perturbation prediction, but this alone does not imply that perturbation models are ineffective. While similar observations have been extensively documented in genetic perturbation benchmarks, their interpretation in chemical perturbation settings remains less clear due to the prevalence of weak and highly heterogeneous transcriptional responses. We proposed a three-axis diagnostic framework and apply it to ChemCPA, CondOT, Doloris and PerturbDiff to show that failure to outperform perturbation-independent baselines can arise from two distinct causes: limited perturbation signal in the data and/or limited utilization of that signal by the model. Results on Tahoe-100M further show that strong learning-free baselines can be dataset- and metric-specific. These findings suggest that aggregate benchmark scores alone are insufficient for interpreting model behavior; understanding whether perturbation information is used, used proportionally, and useful where it matters is essential for evaluating perturbation-prediction models. We believe such diagnostic analyses should become a routine component of perturbation-prediction evaluation, before drawing conclusions from comparisons to learning-free baselines.

AI USE STATEMENT

In this work, we used generative AI tools for grammar and writing polishing. All equations, analysis, and research contributions are our own.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with State. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135. URL https://www.biorxiv.org/content/early/2025/06/27/2025.06.26.661135.

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, Aug 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02772-6. URL https://doi.org/10.1038/s41592-025-02772-6.

Ihab Bendidi, Shawn Whitfield, Kian Kenyon-Dean, Hanene Ben Yedder, Yassir El Mesbahi, Emmanuel Noutahi, and Alisandra K. Denton. Benchmarking transcriptomics foundation models for perturbation analysis : one PCA still rules them all, 2024. URL https://arxiv.org/abs/2410.13956.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. *Advances in Neural Information Processing Systems*, 35:6859–6872, 2022.

Changxi Chi, Jun Xia, Yufei Huang, Zhuoli Ouyang, Cheng Tan, Yunfan Liu, Jingbo Zhou, Chang Yu, Liangyu Yuan, Siyuan Li, Zelin Zang, and Stan Z. Li. Doloris: Dual conditional diffusion implicit bridges with sparsity masking strategy for unpaired single-cell perturbation estimation, 2026. URL https://arxiv.org/abs/2506.21107.

Gerold Csendes, Gema Sanz, Kristóf Z Szalay, and Bence Szalai. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. *BMC Genomics*, 26(1):393, April 2025.

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, Aug 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *NeurIPS 2022*, 2022.

Wenkang Jiang, Yuhang Liu, Yichao Cai, Erdun Gao, Jiayi Dong, Ehsan Abbasnejad, Lina Yao, and Javen Qinfeng Shi. What makes a representation good for single-cell perturbation prediction?, 2026. URL https://arxiv.org/abs/2605.19343.

Eric Kernfeld, Yunxiao Yang, Joshua S. Weinstock, Alexis Battle, and Patrick Cahan. A comparison of computational methods for expression forecasting. *Genome Biology*, 26(1):388, Nov 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03840-y. URL https://doi.org/10.1186/s13059-025-03840-y.

Lanxiang Li, Yue You, Wenyu Liao, Xueying Fan, Shihong Lu, Ye Cao, Bo Li, Wenle Ren, Yunlin Fu, Jiaming Kong, Shuangjia Zheng, Jizheng Chen, Xiaodong Liu, and Luyi Tian. A systematic comparison of single-cell perturbation response prediction models. *bioRxiv*, 2024. doi: 10.1101/2024.12.23.630036. URL https://www.biorxiv.org/content/early/2024/12/23/2024.12.23.630036.

Yiming Li, Jun Zhu, Linjing Liu, Fang Wang, Min Zeng, Longkai Huang, Fan Yang, Min Li, and Jianhua Yao. Unifying genetic and chemical perturbagen representation through a hybrid deep learning framework. *bioRxiv*, 2025. doi: 10.1101/2025.02.02.635055. URL https://www.biorxiv.org/content/early/2025/02/07/2025.02.02.635055.

Yiming Li, Min Zeng, Jun Zhu, Linjing Liu, Fang Wang, Longkai Huang, Fan Yang, Min Li, and Jianhua Yao. UniPert-G2CP bridges genetic and chemical screens from molecular representation to phenotype modeling. *Cell*, 2026.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, Aug 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0494-8. URL https://doi.org/10.1038/s41592-019-0494-8.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.*, 19(6):e11517, June 2023.

Gabriel Mateo Mejia, Henry E Miller, Francis J.A. Leblanc, BO WANG, Brendan Swain, and Lucas Paulo de Lima Camillo. Needles in the haystack: Addressing signal dilution improves scRNA-seq perturbation response modeling and evaluation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XsrXLPxBJw.

Henry E. Miller, Gabriel M. Mejia, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Deep learning-based genetic perturbation models do outperform uninformative baselines on well-calibrated metrics. *bioRxiv*, 2025. doi: 10.1101/2025.10.20.683304. URL https://www.biorxiv.org/content/early/2025/10/21/2025.10.20.683304.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nat. Methods*, 21(3):531–540, March 2024.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with Biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 2023.

Philipp S. L. Schäfer, Kendall A. Reid, Zach Boldyga, Ekin D. Aksu, Hugo Hakem, and Julio Saez-Rodriguez. Towards principled evaluation of single-cell perturbation prediction models. *bioRxiv*, 2026. doi: 10.64898/2026.07.23.740433. URL https://www.biorxiv.org/content/early/2026/07/27/2026.07.23.740433.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234. URL https://www.science.org/doi/abs/10.1126/science.aax6234.

Ramon Vinas Torne, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbic. Systema: A framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 2025.

Chloe Wang, Mehran Karimzadeh, Neal G. Ravindra, Lexi R. Bounds, Nader Alerasool, Ann C. Huang, Shihao Ma, Daniel R. Gulbranson, Haotian Cui, Yongju Lee, Anusuya Arjavalingam, Elliot J. MacKrell, Matthew S. Wilken, Jieming Chen, Benjamin W. Herken, Jesse A. Weber, Massimo M. Onesto, Barbara Gonzalez-Teran, Nicole F. Leung, Sally Yu Shi, Byron J. Smith, Sharon K. Lam, Adam Barner, Philip Wright, Elizabeth M. Rumsey, Soohong Kim, Rene V. Sit, Adam J. Litterman, Ci Chu, and Bo Wang. X-cell: Scaling causal perturbation prediction across diverse cellular contexts via diffusion language models. *bioRxiv*, 2026. doi: 10.64898/2026.03.18.712807. URL https://www.biorxiv.org/content/early/2026/03/20/2026.03.18.712807.

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(2):451–464, Feb 2026. ISSN 1548-7105. doi: 10.1038/s41592-025-02980-0. URL https://doi.org/10.1038/s41592-025-02980-0.

Yan Wu, Esther Wershof, Sebastian M Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024. URL https://openreview.net/forum?id=1EJXtK3AWr.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D. Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 21(8):3, Jul 2025. ISSN 1744-4292. doi: 10.1038/s44320-025-00131-3. URL https://doi.org/10.1038/s44320-025-00131-3.

Xinyu Yuan, Xixian Liu, Ya Shi Zhang, Zuobai Zhang, Hongyu Guo, and Jian Tang. PerturbDiff: Functional diffusion for single-cell perturbation modeling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=yBAqonxCdp.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, Bryan Hariadi, Simone Marrujo, Crina Curca Alec Salvino, Guillermo Gallareta Olivares, Ryan Koehler, Gary Geiss, Alexander Rosenberg, Charles Roco, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398. URL https://www.biorxiv.org/content/early/2025/02/24/2025.02.20.639398.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3D molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

Appendix	Content
A	**Related Work**
B	**Experimental Details**
B.1	Dataset Preprocessing and Train/Test Splits
B.2	Model Implementation Details
B.3	Variant Implementation Details
C	**Additional Evidence for Input Use**
C.1	Latent-space Norm Analysis
C.2	Perturbation Embedding Analysis
D	**Additional Evidence for Response Scaling and Stratified Performance**
D.1	Residual-level Analysis
D.2	Perturbation Effect Size versus Cell Count
E	**Additional Pseudo-bulk and Distribution-level Results**
E.1	Full Intervention Results
E.2	Five-fold Complete Pseudo-bulk Results
E.3	Complete Distribution-level Results
F	**Learning-Free Baselines**
G	**Perturbation-effect Distributions in sci-Plex3 and Tahoe-100M**

A Related Work

Related Work Large-scale Perturb-seq Wang et al. (2026); Csendes et al. (2025) and chemical screening resources Zhang et al. (2025); Srivatsan et al. (2020) have made it practical to train models that predict a cell's response to a compound it has never encountered, and so to screen interventions computationally. The resulting methods differ chiefly in how a perturbation enters the model: as a shift in a learned latent space, as in scGen Lotfollahi et al. (2019), CPA Lotfollahi et al. (2023) and BioLord Piran et al. (2024); as a node in a biological graph, as in GEARS Roohani et al. (2023); or as a conditioning signal on a pretrained Cui et al. (2024); Adduri et al. (2025) or generative backbone Chi et al. (2026); Yuan et al. (2026). Assessments of this progress divide along a revealing line: papers introducing a method generally report gains over their predecessors, whereas independent evaluations that place many methods on common footing repeatedly find linear models, control-based references and perturbation-independent predictors holding their own (Ahlmann-Eltze et al., 2025; Wu et al., 2024; Li et al., 2024; Wei et al., 2026; Csendes et al., 2025; Bendidi et al., 2024).

That discrepancy has turned scrutiny on the evaluation protocols themselves. Reported rankings prove sensitive to decisions that are easy to overlook, among them the choice of metric, the gene panel retained, and how the benchmark is assembled, which has prompted calls for protocols fixed in advance (Wu et al., 2024; Kernfeld et al., 2025; Schäfer et al., 2026; Vinas Torne et al., 2025; Mejia et al., 2026). What none of this supplies is a way to read an individual result: a benchmark can order methods, but not say what produced the order. The evidence is also skewed toward genetic screens, where an effective intervention is expected to leave a measurable transcriptomic mark. Compounds meet that expectation far less reliably, and it is exactly the expectation a comparison against a control-derived baseline rests on. We therefore treat the comparison itself as the object of study, asking of each result which of its possible causes is responsible.

B Experimental Details

B.1 Dataset Preprocessing and Train/Test Splits

Dataset. sci-Plex3 Srivatsan et al. (2020) profiles three human cancer cell lines (A549, MCF7, K562) exposed to 188 single compounds at four dosages (10 nM, 100 nM, 1 μM, 10 μM), spanning 581,777 single cells and 2,247 unique conditions, where a condition is a triplet (c, p, s) of cell line c, perturbation identity p and dosage s. We follow the preprocessing pipeline of ChemCPA Hetzel et al. (2022) and use the preprocessed dataset released with the ChemCPA repository. The full gene panel comprises the 2,000 highly variable genes retained by that pipeline. DEG50 denotes the 50

most differentially expressed genes of a condition, ranked per drug; the same panel is used for every method.

Tahoe-100M. We use Tahoe-100M Zhang et al. (2025) with the fixed held-out-drug evaluation shared by all methods. The evaluation covers 76 unseen drugs and 683 drug–dose–cell-line conditions across A549, RKO, and Hs766T, which are three randomly selected cell lines across three different organs. We use the same 2,000-gene panel from PerturbDiff (Yuan et al., 2026). Held-out drugs are excluded from the training pools. For control-based references, DMSO cells are pooled across plates within each cell line; for differential-expression analysis, treated cells are compared with plate-matched DMSO controls.

Table 6: Unseen drugs in the test set grouped by pathway for one example fold. The test set contains 30 held-out drugs. Each drug is evaluated across three cell lines and four dosages, resulting in 360 test conditions.

Pathway group	**Drugs**
Cell cycle regulation	Danusertib, ENMD-2076, MK-5108, Patupilone, SNS-314
DNA damage & repair	Cyclocytidine, Mercaptopurine, Rucaparib
Epigenetic regulation	(+)-JQ1, Abexinostat, Dacinostat, Divalproex, M344, SRT3025, Sodium, Tacedinaline, Tazemetostat, Tubastatin
HIF signaling	Roxadustat
JAK/STAT signaling	Fedratinib, Filgotinib, Ruxolitinib, TG101209
Metabolic regulation	Alendronate
Nuclear receptor signaling	Fulvestrant
Other	Ofloxacin
Protein folding/degradation	Alvespimycin
Tyrosine kinase signaling	Dasatinib, PD98059, Trametinib

Data split. We split the data by *drug identity* at a ratio of $80/20$ under five-fold cross-validation. Because the split is on drug identity, every perturbation encountered at inference corresponds to a previously *unseen* compound, so the evaluation measures generalization to novel chemistry rather than the easier task of predicting held-out cells of drugs seen in training. Each fold holds out its own drugs: the test partition contains 30 held-out drugs evaluated across three cell lines and four dosages, giving $n = 360$ conditions per fold (348 for fold 3, which holds out 29 drugs), against 1,887 conditions in the corresponding training partition. We report held-out-drug results for fold 0 in the main text and give five-fold stability results in Appendix E.2 and Appendix E.3. Table 6 lists fold 0's held-out compounds grouped by pathway, and Table 7 gives the held-out compounds of all five folds. The folds partition the held-out compounds: no compound is held out in more than one fold.

B.2 Model Implementation Details

ChemCPA. ChemCPA Hetzel et al. (2022) is an autoencoder that encodes an observed expression profile into a basal-state representation and decodes it conditioned on a drug embedding, a cell-line embedding and a dosage. At inference it encodes a control cell and decodes it under the queried compound's embedding. We use the default hyperparameter configuration recommended in the official repository and hold it fixed across all arms; the learning rates come from a Bayesian sweep on the base architecture. ChemCPA and its variants are each trained for 41,000 steps at a batch size of 2,048 cells, so that every arm receives an identical computational budget and the arms differ only in the sampler and the auxiliary loss.

Doloris. Doloris Chi et al. (2026) is a conditional diffusion model trained to match the perturbed distribution, and we run it with the default hyperparameters it provides for sci-Plex3. At evaluation it is scored in its deployed form behind its mask model, generating 100 cells per condition.

Table 7: Held-out (OOD) compounds of each cross-validation fold. The folds partition the held-out compounds: no compound is held out in more than one fold.

Fold	Held-out compounds
0	(+)-JQ1, Abexinostat, Alendronate, Alvespimycin, Cyclocytidine, Dacinostat, Danusertib, Dasatinib, Divalproex, ENMD-2076, Fedratinib, Filgotinib, Fulvestrant, M344, MK-5108, Mercaptopurine, Ofloxacin, PD98059, Patupilone, Roxadustat, Rucaparib, Ruxolitinib, SNS-314, SRT3025, Sodium, TG101209, Tacedinaline, Tazemetostat, Trametinib, Tubastatin
1	A-366, AMG-900, AR-42, AZ, Alisertib, Aurora, Avagacestat, BRD4770, Decitabine, Droxinostat, Enzastaurin, G007-LK, GSK, Givinostat, IOX2, KW-2449, Lapatinib, Lomustine, Mesna, Motesanib, PCI-34051, PFI-1, PHA-680632, Resminostat, Resveratrol, S-Ruxolitinib, Sirtinol, Tanespimycin, Triamcinolone, WP1066
2	2-Methoxyestradiol, AC480, Azacitidine, BMS-265246, BMS-911543, Bisindolylmaleimide, Busulfan, CYC116, Cerdulatinib, EED226, Entacapone, GSK-LSD1, GSK1070916, Luminespib, MC1568, Navitoclax, PD173074, PJ34, Panobinostat, Pelitinib, Pirarubicin, Selisistat, Thalidomide, Tofacitinib, Trichostatin, Tucidinostat, UNC0379, UNC0631, WHI-P154, ZM
3	ABT-737, AG-14361, Altretamine, BMS-754807, Barasertib, Capecitabine, Ellagic, FLLL32, Flavopiridol, Gandotinib, Hesperadin, ITSA-1, MK-0752, Mocetinostat, NVP-BSK805, Nilotinib, Nintedanib, Pracinostat, Prednisone, Quisinostat, Raltitrexed, Roscovitine, SRT2104, TAK-901, Thiotepa, UNC1999, Valproic, Vandetanib, Veliparib
4	AG-490, AT9283, AZD1480, Baricitinib, Belinostat, CEP-33779, CUDC-101, CUDC-907, Carmofur, Cediranib, Clevudine, Entinostat, Epothilone, Fluorouracil, Glesatinib (MGCD265), INO-1001, Iniparib, JNJ-26854165, JNJ-7706621, MLN8054, Meprednisone, Momelotinib, Obatoclax, RG108, SL-327, SRT1720, Streptozotocin, TMP195, Toremifene, Tozasertib

PerturbDiff. PerturbDiff Yuan et al. (2026) is likewise a conditional diffusion model, but its public implementation does not support generalization to unseen drugs. We therefore adapt it: the compound is represented by the same RDKit descriptor as ChemCPA, encoded by an MLP and combined with dosage through the amortized dosage fusion of ChemCPA Hetzel et al. (2022). To determine a suitable training budget, we hold out 10% of the training conditions of a single fold as a validation set, drawn as unseen *conditions* of drugs seen during training. We select a budget at which validation performance has converged and use it for all PerturbDiff runs.

CondOT. CondOT Bunne et al. (2022) is a conditional optimal-transport model that learns a map from control cells to treated cells. Because each evaluation row is a different cell-line, drug, and dosage condition, we use the same latent-space amortized drug-dose fusion as ChemCPA to parameterize the condition input. We use the default CondOT implementation for sci-Plex3 with PCA dimension 50. At evaluation, CondOT transports control cells for each queried condition and is scored on the same held-out-drug sci-Plex3 conditions and DEG50 panel as the other models.

Common training setup. All ChemCPA arms use the same optimizer settings and a single seed. All models are evaluated on the same held-out-drug conditions with the same preprocessing and the same DEG50 panel.

B.3 Variant Implementation Details

B.3.1 Balanced Sampling

Under the standard cell-level sampling of ChemCPA Hetzel et al. (2022), each condition contributes to the gradient in proportion to the number of cells profiled for it. This interacts badly with a property of the data itself: strong-effect drugs sit in the long tail of the effect-size distribution and are profiled with fewer surviving cells Srivatsan et al. (2020), as shown in Appendix D.2. Standard sampling

therefore under-represents exactly the conditions on which perturbation information matters most, and lets common weak-effect conditions dominate optimization.

We therefore sample uniformly over conditions rather than over cells. In *condition-balanced* sampling, each draw first selects a condition (c, p, s) uniformly at random from the training conditions and then a cell uniformly at random from within it, so every drug–cell-line–dose condition contributes about equally to each training epoch regardless of how many cells were profiled for it. In practice each batch draws 64 conditions and 32 cells per condition, sampling with replacement when a condition holds fewer cells.

Condition-balanced sampling still favours compounds that are profiled in many conditions. We therefore also consider *drug-balanced* sampling, which applies the balancing one level up: each batch selects 16 compound–cell-line pairs uniformly at random among pairs profiled at three or more dosages, includes all dosages of each selected pair, and draws 32 cells per resulting condition, so that each compound–cell-line pair rather than each condition is equally represented. Comparing the two separates imbalance in the number of observations per compound from imbalance across conditions.

Both samplers replace ChemCPA's cell-level sampling and nothing else. Neither alters the loss, the architecture, or the test distribution, and all variants train for an identical number of gradient steps, so the arms differ only in the frequency with which conditions are visited during training.

B.3.2 Directional-loss Objective

For each condition (c, p, s) present in a mini-batch $\mathcal{B}$, let $\mathbf{m}_{c,s}$ denote the *context mean*, the pseudo-bulk mean over all treated training cells of line c at dosage s, precomputed once on the training split over strata containing at least 10 cells. The observed and predicted residuals are $\mathbf{d}^{(c,p,s)} := \boldsymbol{\mu}^{\mathcal{B}}_{c,p,s} - \mathbf{m}_{c,s}$ and $\hat{\mathbf{d}}^{(c,p,s)} := \hat{\boldsymbol{\mu}}^{\mathcal{B}}_{c,p,s} - \mathbf{m}_{c,s}$. We use the context mean rather than the control mean because it is the stronger perturbation-agnostic reference at the pseudo-bulk level; the residual should accordingly be read as perturbation-specific variability relative to that reference, and not as the drug effect of Definition 1, which is defined against the control mean. A cosine term aligns the predicted residual with the direction of the observed response, and a magnitude term calibrates its effect size,

$$\mathcal{L}_{\cos} = 1 - \frac{\hat{\mathbf{d}}^{\top}\mathbf{d}}{\|\hat{\mathbf{d}}\|_2\|\mathbf{d}\|_2 + \varepsilon}, \qquad \mathcal{L}_{\mathrm{mag}} = \Big(\|\hat{\mathbf{d}}\|_2 - \|\mathbf{d}\|_2\Big)^2, \tag{4}$$

where the condition indices are suppressed and $\varepsilon > 0$ ensures numerical stability. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \underbrace{\lambda_{\cos}\mathcal{L}_{\cos} + \lambda_{\mathrm{mag}}\mathcal{L}_{\mathrm{mag}}}_{\mathcal{L}_{\mathrm{dir}}} + \Omega, \tag{5}$$

where $\mathcal{L}_{\mathrm{rec}}$ is the original reconstruction objective and Ω the model-specific disentanglement regularizer. Since $\mathcal{L}_{\mathrm{dir}}$ is estimated from within-batch pseudo-bulk means, the quality of that estimate depends on how many cells of a condition a batch contains, which is precisely what the next intervention controls.

B.3.3 Training and Optimization Details

Because the directional loss is estimated from within-batch pseudo-bulk means, it is exposed to a trade-off that the reconstruction objective is not. With a fixed batch size of $2{,}048$, the number of cells drawn per condition determines both how noisy each within-batch pseudo-bulk estimate is and how many distinct conditions the batch spans. Table 8 varies the two quantities that control this: the cells drawn per condition, and the weight λ on the auxiliary term.

Two patterns emerge, and they pull in opposite directions. First, perturbation utilization is maximized at a *small* number of cells per condition. Under condition-balanced sampling, β_{DEG50} rises from 0.369 at 16 cells to 0.513 at 32 and then collapses to 0.076 at 64 and 0.044 at 128; DEG50 R^2 peaks at the same setting. Since $2{,}048/32 = 64$ conditions per batch against only 16 at 128 cells, the ordering says that the directional loss benefits more from seeing many conditions per step than from a low-variance estimate of any one of them. Second, the full-panel E-distance moves the other way, improving monotonically as cells per condition grows ($1.899 \to 1.662$), because a better-resolved within-batch population is what the distributional metric rewards. Drug-balanced sampling shows the same shape with a uniformly lower β, consistent with its coarser balancing unit.

The weight sweep behaves similarly. Both DEG50 R^2 and β_{DEG50} peak at $\lambda = 1$ (0.772 and 0.513), while E-distance keeps improving out to $\lambda = 10$ (1.519) as β falls back to 0.105. Pushing the auxiliary term harder therefore buys distributional fit at the cost of the response scaling it was introduced to repair. We use 32 cells per condition and $\lambda_{\cos} = \lambda_{\mathrm{mag}} = 1$ throughout the paper, which is the setting that maximizes perturbation utilization; we note that it is not the setting that minimizes E-distance, and that no setting we tried optimizes both.

Table 8: Sensitivity of the auxiliary loss to batch composition and to its weight on the validation partition. *Cells per condition* is the number of cells of a condition in a balanced batch; with a batch size of 2,048, it also determines how many conditions the batch spans. The setting used throughout the paper is 32 cells per condition with both weights set to 1.

	Condition-balanced			Drug-balanced		
Cells per condition	DEG50 R^2	β_{DEG50}	Full E-dist	DEG50 R^2	β_{DEG50}	Full E-dist
16	0.712	0.369	1.899	0.785	0.322	1.817
32	0.772	0.513	1.882	0.770	0.220	1.769
64	0.701	0.076	1.708	0.748	0.074	1.693
128	0.720	0.044	1.662	0.695	0.022	1.807
Weight λ	Condition-balanced			Drug-balanced		
0.001	0.621	0.190	2.179	0.724	0.086	2.053
0.01	0.745	0.212	2.028	0.758	0.114	1.995
0.1	0.709	0.445	1.923	0.784	0.279	2.053
1	0.772	0.513	1.882	0.770	0.220	1.769
5	0.703	0.209	1.713	0.697	0.061	1.594
10	0.657	0.105	1.519	0.684	0.054	1.426

C Additional Evidence for Input Use

C.1 Latent-space Norm Analysis

For architectures that combine basal, contextual, and perturbation components via an addition operator, we additionally measure the relative scale of each contribution at the point of fusion. This analysis compares the norms of the basal representation, context contribution, and dose-modulated perturbation contribution.

Since ChemCPA is trained solely through full-gene reconstruction and, as a conventional autoencoder architecture, may be prone to shortcut learning, it may achieve good reconstruction performance while largely ignoring perturbation information.

Following this idea, we pay specific attention to ChemCPA's additive latent fusion design, in which the cell basal-state representation, covariate encoding (*i.e.*, cell line), and perturbation embedding are combined through *simple vector addition*. Under this design choice, perturbation information may be ignored or canceled if its contribution is substantially smaller in magnitude than the other latent components. To investigate this hypothesis, we inspect the norms of the latent representations and find that the basal-state representation is approximately $25\times$ larger than the perturbation embedding. This substantial scale imbalance suggests that the prediction is dominated by the cell basal-state and covariate encodings, providing a possible explanation for the model's weak dependence on perturbation identity.

Norms alone do not prove functional importance, so the analysis should be interpreted together with ablation or corruption sensitivity. A small perturbation component that causes a large output change may still be important, while a large component may be suppressed by later layers.

C.2 Perturbation Embedding Analysis

To assess whether the perturbation embedding contains useful biological information, we evaluate pathway prediction from the embedding alone. Performance above a shuffled-label reference indicates that the embedding contains information about mechanism of action (MoA). We leverage three chemical embeddings: the default RDKit and Uni-Mol Zhou et al. (2023) embeddings used by ChemCPA and Doloris , and one additional embedding, UniPert Li et al. (2026). As an independent check of embedding quality, we perform multi-class MoA classification on the sci-Plex3 compounds.

Table 9: Latent-space norm analysis on basal state $\mathbf{z}$, perturbation encoding $\mathbf{p}$ and covariate encoding $\mathbf{c}$. Norms are reported across cells from OOD evaluation conditions ($n = 360$ conditions). Drug-to-basal and drug-to-covariate ratios quantify the relative scale of perturbation information at the additive fusion.

Model	$\|\mathbf{z}\|$	$\|\mathbf{c}\|$	$\|\mathbf{p}\|$	$\|\mathbf{p}\|$ / $\|\mathbf{z}\|$	$\|\mathbf{p}\|$ / $\|\mathbf{c}\|$
Vanilla ChemCPA	48.93 ± 3.22	7.50 ± 1.66	2.02 ± 2.28	0.041	0.269
Condition-balanced	23.15 ± 20.13	9.03 ± 1.34	7.15 ± 8.14	0.309	0.792
+ Directional loss	36.38 ± 21.64	8.36 ± 1.27	13.48 ± 11.68	0.370	1.612

Table 10: Effect of molecular representation on perturbation prediction and mechanism-of-action classification. DEG50 expression R^2 is reported as mean $\pm$ standard deviation across five random seeds on the held-out-drug test set. MoA classification uses a multinomial logistic regression (balanced class weights, $C = 1$) on z-scored embeddings under 5-fold stratified cross-validation at pathway level 1; values are mean $\pm$ standard deviation across folds. The shuffled-label reference comes from 50 random label permutations with the same classifier in the same filtered setting; its mean of about 0.10 matches the chance level for ten balanced classes.

Embedding	DEG50 Expression R^2	MoA Bal. Acc.	Shuffled MoA Bal. Acc.
RDKit	0.773 ± 0.025	0.302 ± 0.068	0.062 ± 0.020
UniMol	0.746 ± 0.017	0.282 ± 0.059	0.060 ± 0.015
UniPert	0.751 ± 0.008	$\mathbf{0.540} \pm 0.120$	0.060 ± 0.014

Stronger perturbation embeddings do not improve prediction. Despite substantial differences among RDKit, Uni-Mol Zhou et al. (2023), and UniPert Li et al. (2025), replacing the perturbation representation has little effect on prediction performance. As shown in Table 10, these embeddings differ markedly in their ability to capture compound MoA information, suggesting that representation quality is unlikely to be the primary bottleneck. In particular, replacing the original RDKit representation with a more biologically informative embedding such as UniPert leaves DEG50 expression R^2 in a similar range, around 0.75. RDKit achieves the highest mean performance across five runs, but also exhibits the largest run-to-run variability (0.025). Thus, its apparent advantage should be interpreted cautiously, as the differences between embeddings are modest relative to the observed variation across runs.

D Additional Evidence for Response Scaling and Stratified Performance

D.1 Residual-level Analysis

The stratified analysis of the main text reports raw DEG50 R^2 per effect-size bin, where a ceiling effect compresses the weak-effect regime that dominates the dataset. We therefore re-express the same quantity as a residual against the drug-blind reference (control mean), plotting each method's DEG50 R^2 *minus* the control mean's DEG50 R^2 within the same bin. Zero is then the line a model must clear to have used the compound at all, and the sign carries the interpretation directly: above zero the model has extracted something from the perturbation identity, below zero it would have done better ignoring it.

More specifically, the weak-effect regime separates the methods that the raw R^2 curves cannot. In the four weakest bins of all three cell lines, PerturbDiff sits *below* the drug-blind baseline, by between 0.023 and 0.086. However, its advantage is more obvious in the strongest bin, where it stands $+0.590$ above the control mean in A549, $+1.078$ in K562 and $+0.487$ in MCF7. Similarly, in the strongest bin Doloris reaches $+0.332$, $+0.917$ and $+0.315$ across the three lines. Strong perturbation utilization, in other words, does not necessarily lead to uniform improvement; it gives a large gain confined to the tail of the effect-size distribution, and sometimes costs a small degradation across the weak-effect bulk of drugs. Since that bulk dominates the condition count, the two effects very nearly cancel in any aggregate score, which is the mechanism behind the modest differences of Table 11.

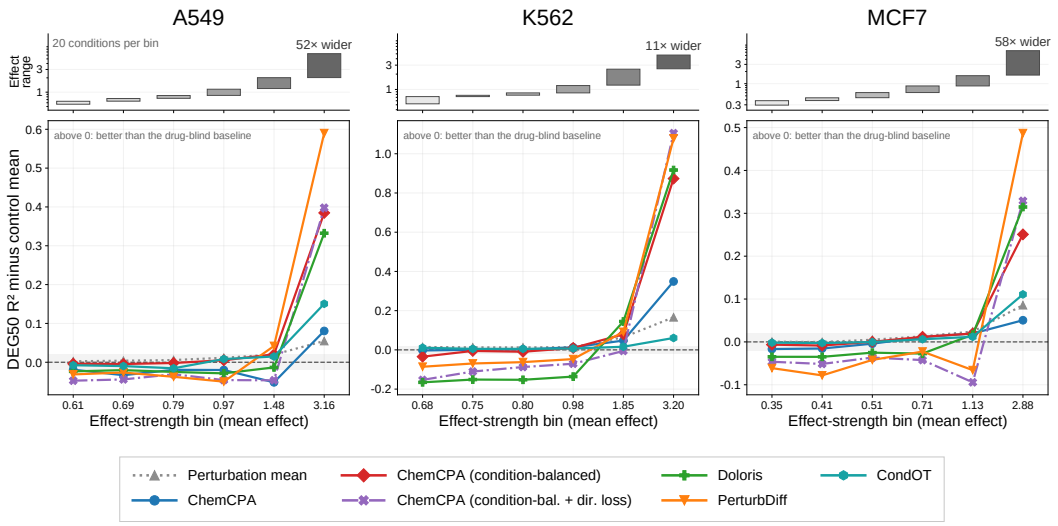


Figure 3: sci-Plex3effect-strength stratification on held-out drugs ($n = 360$ conditions; 120 per cell line), plotted as DEG50 R^2 minus the control mean's DEG50 R^2 in the same bin. Above zero a model has used the compound; below zero it has done worse than ignoring it, and the shaded band marks ± 0.02 around the baseline. Conditions are split into six equal-count bins of 20 conditions by observed pseudo-bulk effect size δ, drawn evenly spaced, so the horizontal axis is ordinal rather than linear in δ; tick labels give the mean effect of each bin. Equal-count binning conceals a long tail: most held-out conditions do very little, and the few that act strongly are spread across a range that a single bin has to absorb.

D.2 Perturbation Effect Size versus Cell Count

The imbalance that motivates balanced sampling is a property of sci-Plex3, not of the model. Cytotoxic compounds kill the cells they act on most strongly, so the conditions carrying the largest transcriptomic response are also the ones profiled with the fewest surviving cells. Figure 4 shows this directly: across all sci-Plex3conditions, perturbation effect size and cell count are strongly negatively associated in rank.

The consequence for training is direct. Under cell-level sampling, a condition's influence on the gradient is proportional to its cell count, so the conditions that carry the strongest perturbation signal are systematically the least visited. Balancing over conditions rather than cells implicitly upweights these high-effect, low-cell-count conditions, which is why the intervention of Section 5.5 improves performance precisely in the strongest effect-size bins rather than uniformly.

E Additional Pseudo-bulk and Distribution-level Results

E.1 Full Intervention Results

The main text reports the interventions on a single metric to keep the diagnostic argument legible. This subsection gives the complete picture, at both resolutions and on both gene panels, so that the trade-off between them is visible. Table 11 reports pseudo-bulk accuracy and Table 12 the single-cell distributional metrics, over the same 360 held-out-drug conditions. We give the full stratified results of ChemCPA's variates in Figure 5

E.2 Five-fold Complete Pseudo-bulk Results

Table 13 reports the pseudo-bulk metrics on the held-out-drug partition of every fold, aggregated as the mean $\pm$ standard deviation across the five folds. The picture of the main text survives the resampling. On the full gene panel no trained variant reaches the control mean (R^2 of 0.922 for the best variant against 0.927), and vanilla ChemCPA sits well below it (0.884). On the top differentially

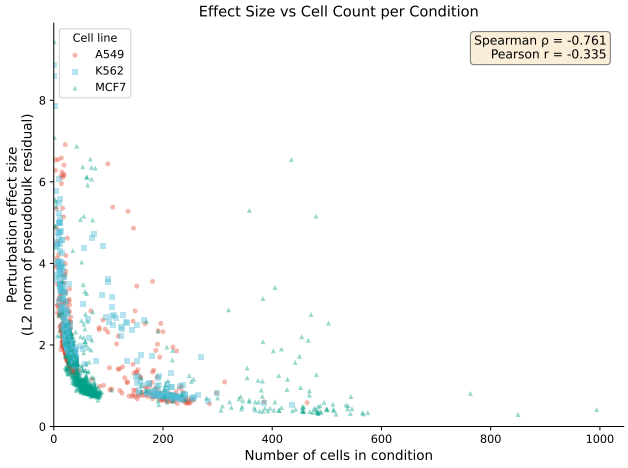


Figure 4: Perturbation effect size versus number of profiled cells per condition in sci-Plex3. Each point represents one perturbation condition, colored by cell line. Perturbation effect size is defined as the L2 norm of the pseudobulk residual, where the control mean is computed separately for each cell line. Strong perturbations are associated with fewer surviving cells (Spearman $\rho = -0.761$, Pearson $r = -0.335$). Because strong-effect conditions are profiled with fewer cells, standard cell-level sampling under-represents exactly the conditions on which perturbation information matters most; condition- or drug-balanced sampling therefore implicitly upweights these high-effect, low-cell-count conditions during training.

Table 11: Pseudobulk expression-level performance of ChemCPA under different sampling strategies and with or without directional loss over the test set ($n = 360$ conditions). Values are mean $\pm$ standard deviation across conditions. Shaded cells denote the best-mean method and methods not significantly different from it by a two-sided Wilcoxon signed-rank test with Bonferroni correction ($\alpha = 0.05$). Bold indicates the best raw mean.

	Test					
	Full panel			DEG50		
Model	Pearson ↑	R^2 ↑	RMSE ↓	Pearson ↑	R^2 ↑	RMSE ↓
Vanilla ChemCPA	0.953 ± 0.077	0.896 ± 0.164	$0.035 \pm <0.022$	0.894 ± 0.207	0.739 ± 0.529	0.113 ± 0.118
Condition-balanced	$\mathbf{0.964 \pm 0.045}$	$\mathbf{0.922 \pm 0.105}$	$0.031 \pm <0.022$	$\mathbf{0.919 \pm 0.141}$	$\mathbf{0.806 \pm 0.592}$	$\mathbf{0.099 \pm 0.095}$
+ Directional loss	0.936 ± 0.057	0.855 ± 0.110	0.046 ± 0.015	0.904 ± 0.125	0.772 ± 0.280	0.120 ± 0.089
CondOT	0.962 ± 0.073	0.916 ± 0.162	0.029 ± 0.021	0.901 ± 0.206	0.739 ± 0.604	0.108 ± 0.118
Doloris	0.938 ± 0.068	0.874 ± 0.129	0.041 ± 0.017	0.893 ± 0.171	0.769 ± 0.337	0.117 ± 0.094
PerturbDiff	0.947 ± 0.061	0.866 ± 0.145	0.043 ± 0.015	0.926 ± 0.111	0.808 ± 0.288	0.107 ± 0.073
Control mean	0.961 ± 0.080	0.915 ± 0.181	0.027 ± 0.023	0.894 ± 0.220	0.719 ± 0.628	0.111 ± 0.126
Perturbation mean	0.963 ± 0.077	0.920 ± 0.170	0.027 ± 0.022	0.901 ± 0.213	0.747 ± 0.580	0.104 ± 0.122

expressed genes the balanced variants do pull ahead of the control mean, by 0.043 for condition-balanced sampling and 0.022 for drug-balanced sampling with the directional loss, while vanilla ChemCPA remains below it (0.732 against 0.743).

E.3 Complete Distribution-level Results

Table 14 reports the distribution-level metrics on the held-out-drug partition of every fold (each fold holds out its own drugs, with the single-cell baselines rebuilt from each fold's training split.

F Distributional Baseline Robustness

A learning-free baseline that draws a fixed number of cells could in principle owe its strong distributional scores to that sample size rather than to the reference population being right. Figure 6 checks

Table 12: Single-cell distribution-level performance on the test set ($n = 360$ held-out-drug conditions, fold 0), as mean $\pm$ standard deviation across conditions; lower is better throughout. Shaded: best per metric, and significantly better than every other method (Wilcoxon signed-rank, Bonferroni, $\alpha = 0.05$). Bold: best overall. Italic: best trainable model. The single-cell baselines draw 1,000 cells.

	Test			
	Full panel		DEG50	
Model	E-Distance ↓	EMD ↓	E-Distance ↓	EMD ↓
Vanilla ChemCPA	2.08 ± 0.22	0.065 ± 0.010	1.17 ± 0.31	0.212 ± 0.101
Condition-balanced	1.96 ± 0.20	0.063 ± 0.009	1.08 ± 0.25	0.197 ± 0.086
+ Directional loss	1.88 ± 0.18	0.051 ± 0.008	1.02 ± 0.22	0.166 ± 0.072
CondOT	1.24 ± 0.31	0.060 ± 0.010	0.68 ± 0.35	0.166 ± 0.091
Doloris	1.09 ± 0.51	*0.027 ± 0.006*	0.91 ± 0.66	0.099 ± 0.073
PerturbDiff	*0.54 ± 0.35*	0.051 ± 0.007	**0.26 ± 0.28**	*0.091 ± 0.052*
scControl	0.41 ± 0.29	**0.016 ± 0.008**	0.41 ± 0.39	0.091 ± 0.102
scPerturb	**0.40 ± 0.28**	0.016 ± 0.008	0.39 ± 0.38	**0.086 ± 0.098**

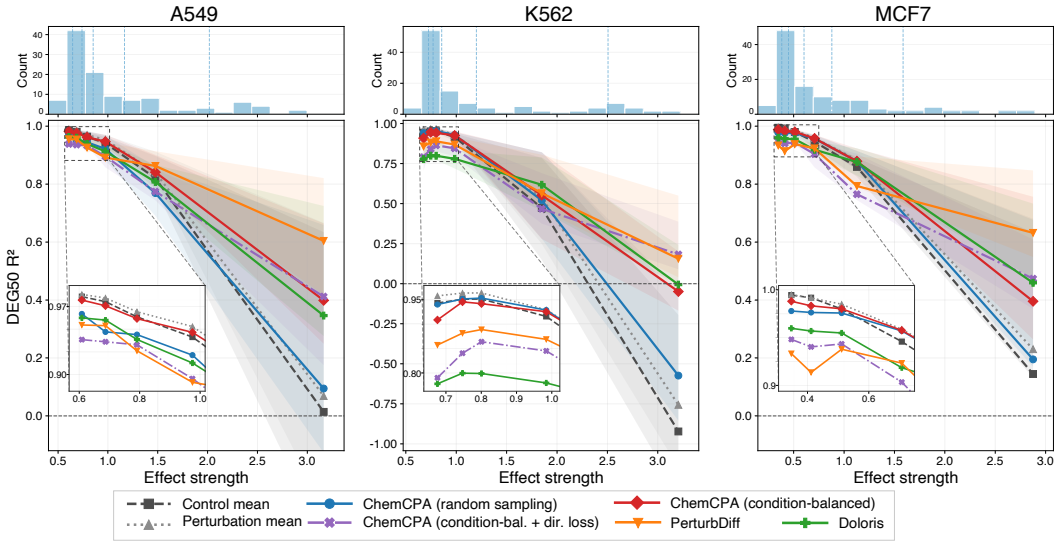


Figure 5: sci-Plex3 stratification analysis of the drug effect with respect to R^2 on held-out drugs ($n = 360$ conditions; 120 per cell line). Columns correspond to the three cell lines (A549, K562, MCF7). Conditions are split into six equal-frequency bins by pseudo-bulk effect size δ; lines show the mean DEG50 R^2 per bin, shaded regions the interquartile range, and the top row the number of conditions along the effect-size axis. The zoom-in window enlarges the weak-effect region per panel.

this by varying the number of sampled cells for scControl and scPerturb and re-measuring both metrics. Energy distance and full-gene EMD both fall steeply over the first hundred cells and then flatten, with EMD largely converged by roughly 500 cells and energy distance reaching a plateau by around 1,000. The 1,000-cell setting used throughout the paper therefore sits on the plateau rather than on the steep part of the curve, so the baselines' advantage is not an artifact of the sample size we granted them.

Table 13: Pseudo-bulk performance across five folds (held-out-drug partition per fold). Values are the mean $\pm$ standard deviation of the five per-fold means.

	Full panel			DEG50		
Model	Pearson	R^2	RMSE	Pearson	R^2	RMSE
Vanilla ChemCPA	0.946 ± 0.015	0.884 ± 0.028	0.037 ± 0.006	0.902 ± 0.018	0.732 ± 0.076	0.102 ± 0.013
ChemCPA (Condition-balanced)	0.965 ± 0.007	0.922 ± 0.018	0.029 ± 0.002	0.926 ± 0.017	0.786 ± 0.082	0.086 ± 0.011
ChemCPA (Condition-bal. + dir. loss)	0.963 ± 0.007	0.911 ± 0.019	0.030 ± 0.003	0.922 ± 0.018	0.765 ± 0.084	0.093 ± 0.013

Table 14: Distribution-level performance across five folds (held-out-drug partition per fold). Values are the mean ± standard deviation of the five per-fold means.

Model	Full E-dist	Full EMD	DEG50 E-dist	DEG50 EMD
Vanilla ChemCPA (Random)	2.018 ± 0.078	0.063 ± 0.001	1.091 ± 0.060	0.184 ± 0.015
ChemCPA (Condition-balanced)	1.948 ± 0.020	0.063 ± 0.001	1.036 ± 0.034	0.181 ± 0.010
ChemCPA (Condition-bal. + dir. loss)	1.427 ± 0.086	0.054 ± 0.002	0.737 ± 0.068	0.141 ± 0.014

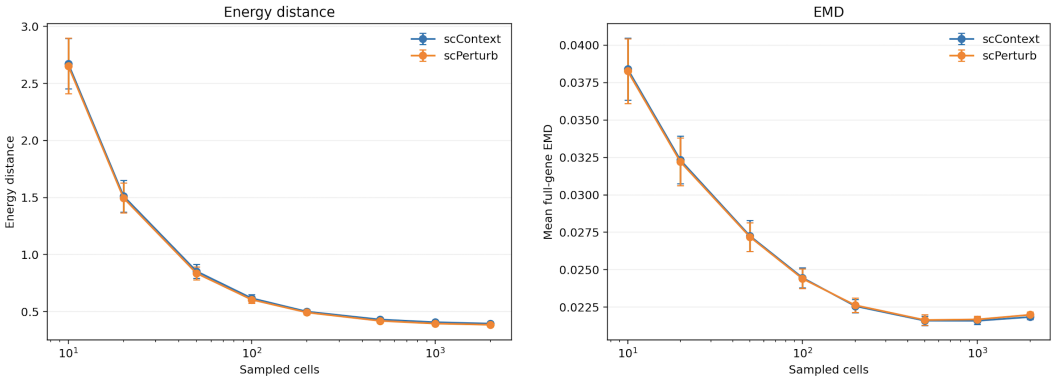


Figure 6: Distributional baseline robustness to cell subsampling for SCCONTROL and SCPERTURB. Energy distance (left) and full-gene Earth Mover's Distance (EMD; right) are evaluated as the number of sampled cells increases. Results are averaged over 50 randomly sampled conditions which are then fixed for each run, and 5 random seeds per sample size. Both metrics stabilize as the number of sampled cells increases, with EMD largely converging by approximately 500 cells and energy distance approaching a plateau by around $1,000$ cells.

G PERTURBATION-EFFECT DISTRIBUTIONS IN SCI-PLEX3 AND TAHOE

We compare observed perturbation-effect distributions in the held-out evaluation subsets of sci-Plex3 and Tahoe. The sci-Plex3 subset contains 360 conditions comprising 30 held-out drugs at four doses across A549, K562, and MCF7. The Tahoe subset contains 683 conditions comprising 76 unseen drugs at three doses across A549, Hs766T, and RKO. Each condition corresponds to a drug–dose–cell-line combination.

To reduce sensitivity to isolated extreme conditions, we summarize each cell line's effect distribution using Bowley skewness, $B = (Q_{75} + Q_{25} - 2Q_{50})/(Q_{75} - Q_{25})$, where Q_p denotes the pth percentile of observed effect strength. Positive values indicate right asymmetry within the middle 50% of the distribution. We additionally report Q_{90}/Q_{50}, which measures upper-tail spread relative to the median rather than skewness itself. Both summaries are invariant to multiplication of effect strength by a positive constant. As shown in Table 15, sci-Plex3 has greater quartile asymmetry and greater upper-tail spread relative to Tahoe-100M.

Table 15: Robust summaries of observed perturbation-effect distributions, computed separately for each cell line. n is the number of evaluated conditions. Bowley skewness measures quartile asymmetry, while Q_{90}/Q_{50} measures relative upper-tail spread.

Dataset	Cell line	n	Bowley skewness	Q_{90}/Q_{50}
sci-Plex3	A549	120	0.56	4.72
sci-Plex3	K562	120	0.68	5.80
sci-Plex3	MCF7	120	0.47	4.87
Tahoe	A549	228	0.26	1.51
Tahoe	Hs766T	227	0.05	1.55
Tahoe	RKO	228	0.31	1.62