Where Simple Baselines Fail: Mapping the Modeling Frontier of Perturbation Prediction

Anonymous Authors[1]

Abstract

Whether deep learning models for perturbation prediction outperform simple baselines remains contested. We argue this debate is ill-posed: standard benchmarks mix perturbations already well-predicted by simple rules with others whose reproducible signal remains unexplained. We introduce Baseline Saturation, a per-perturbation measure of how much of the control-bounded dynamic range a matched simple baseline already captures. Across nine CRISPR perturbation datasets and six generalization regimes, we find that within-regime heterogeneity dominates: even within a single scenario, perturbation-level saturation spans the full range, with approximately 40% of evaluations near-solved by baselines and 30% remaining resistant. Deep learning models (scGPT, GEARS, PRESAGE) outperform baselines specifically on resistant perturbations, recovering which genes respond and in which direction, but not response magnitudes. Because saturated perturbations dominate most test sets, these gains are masked by aggregate reporting. Biologically, resistance is driven by the uniqueness of a perturbation's response relative to the dataset mean and is cell-type-dependent rather than gene-intrinsic. Our results reframe evaluation: simple baselines define the solved regime, and their failures map the frontier where expressive models add value.

1. Introduction

Single-cell perturbation screens measure how genetic or chemical interventions reshape cellular states, enabling systematic studies of gene function, regulatory programs, and drug response (Adamson et al., 2016; Norman et al., 2019; Frangieh et al., 2021; Replogle et al., 2022). These data have motivated models that predict transcriptional responses to unobserved perturbations, cellular contexts, doses, or combinations (Squires et al., 2024; Roohani et al., 2024; Cui et al., 2024; Wu et al., 2024; Adduri et al., 2025). Yet the field remains divided on a basic question: do deep learning models outperform simple baselines? Recent evaluations give conflicting answers, with some supporting expressive architectures and others showing that mean-based, additive, or linear predictors are difficult to beat (Miller et al., 2025; Ahlmann-Eltze et al., 2025; Kernfeld et al., 2024; Vollenweider & Bühlmann, 2026). We argue that this aggregate question is ill-posed. Perturbation benchmarks mix regimes with very different modeling demand. Some responses are noise-limited, leaving little signal for any model. Others are reproducible but baseline-saturated, because simple structure such as shared mean effects, cross-context averages, or additivity already captures most of the signal. The relevant frontier is the remaining unsaturated subset: perturbations whose responses rise above noise but are not explained by linear rules. This perspective also clarifies why train/test split labels can be misleading. A split may be out-of-distribution yet easy if the held-out response follows a linear structure; conversely, a seemingly standard unseen-perturbation split can be difficult when held-out effects deviate from the training-set average. Split type alone, therefore, does not define benchmark difficulty. What matters is how much reproducible signal remains beyond matched baselines.

In this work, we introduce *Baseline Saturation*, a regime-aware evaluation framework that quantifies, for each perturbation and metric, what fraction of the control-bounded dynamic range is already captured by a matched simple baseline. Strong baselines are therefore not only competitors; they are diagnostic tools. They define what is already explainable under a given evaluated perturbation regime, and their failures identify where the heterogeneous signal remains. Building on recent work on empirical ceilings, simple baselines, and perturbation complexity (Miller et al., 2025; Vollenweider & Bühlmann, 2026; Liang & Singh, 2026), we use baseline performance to stratify benchmark difficulty rather than only to rank models globally.

Across nine public single-cell CRISPR perturbation datasets,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

six prediction regimes, calibrated metrics, rule-based and learned baselines, and three representative deep learning models, scGPT, GEARS, and PRESAGE (Cui et al., 2024; Roohani et al., 2024; Wu et al., 2024), we find that benchmark difficulty is strongly compositional. Prediction regimes differ in average saturation, but perturbation-level heterogeneity within a regime is often larger than the regime-level effect. Deep learning gains concentrate on baseline-resistant perturbations, particularly for metrics that capture directionality and gene-program recovery, but these gains can be hidden by aggregate metric reporting.

Our results shift perturbation-response evaluation from asking whether deep learning models win globally to asking where simple baselines fail, why they fail, and whether expressive models recover the missing signal. Baseline Saturation provides a practical diagnostic for locating this frontier and for designing benchmarks that better reflect the modeling demand of perturbation prediction.

2. Preliminaries

Prediction regimes. A central source of disagreement across perturbation-prediction benchmarks is that models are evaluated under different notions of generalization. Some splits test interpolation across observed perturbation–context structure, whereas others require extrapolation to unseen perturbations, cellular contexts, doses, or combinations. Without making these regimes explicit, aggregated comparisons can conflate qualitatively different prediction problems. We formalize six prediction regimes that partition the perturbation-effect tensor along distinct generalization axes, applicable to datasets with single or multiple cellular contexts (Figure 1A). UNSEENPERT — test perturbations are entirely absent from training; cell types are shared. UNSEENCELL — an entire cell type is held out; all perturbations are seen in other cell types. UNSEENBOTH — both the perturbation and cell type are unseen; a zero-shot setting. UNSEENPAIR — a specific combination of cell type and perturbation is held out, while both appear elsewhere in training. UNSEENDOSE — a perturbation at a given dosage is held out; the same perturbation at other dosages is observed. UNSEENCOMBO — individual single-gene perturbations are observed; the combinatorial (double-perturbation) effect is held out. UNSEENCELL, UNSEENBOTH, and UNSEENPAIR require datasets with multiple cell types. These regimes differ not only in what information is available at test time but also in how much modeling capacity they demand. Comparing raw metrics across such regimes can therefore obscure whether performance differences reflect model capacity, split difficulty, or the amount of signal left beyond simple predictors.

Controls and metric calibration. All predictors operate on pseudobulk mean expression vectors $\mu_{c,p,g}$, obtained by averaging single-cell profiles across cells within each (cell type c, perturbation p) condition for gene g. Perturbation deltas are then $\Delta_{c,p,g} = \mu^{\text{pert}}_{c,p,g} - \mu^{\text{ctrl}}_{c,g}$. Raw prediction metrics are difficult to compare across perturbations, datasets, and generalization regimes because they mix three components in unknown proportions: measurement noise, linear structure, and heterogeneous signal. If most of the observed delta is noise, even an optimal predictor has limited signal to recover. If the reproducible component is dominated by linear structure, such as mean shifts or approximately additive effects, simple baselines may approach the empirical ceiling. Only the remaining component— heterogeneous signal that is reproducible but not captured by simple predictors—defines the regime where more expressive models have meaningful headroom.

To make metrics comparable, we anchor each one between two reference points. The **negative control** (Zero) predicts no perturbation effect ($\hat{\Delta} = 0$), representing the performance floor when no predictor is used. For the **positive control**, we use an Interpolated duplicate (Miller et al., 2025), which predicts from a split-half replicate of the ground truth on differentially expressed genes and falls back to the mean baseline elsewhere. This reduces sampling noise on the majority of unaffected genes while preserving biological signal where it matters.

The Dynamic Range Fraction (DRF; Miller et al., 2025) then quantifies whether the gap between these controls is large enough for a metric to discriminate meaningful predictions from linear ones. When the perturbation signal is weak or the signal-to-noise ratio is low, the resulting delta is close to zero and predicting "no effect" is nearly as accurate as the interpolated duplicate. Our analysis shows that DRF is largely a function of how much heterogeneous signal is in the dataset and how capable the metric is in capturing this signal. We improve calibration by concentrating evaluation on signal-bearing genes and perturbations, using DEG weighting, biologically relevant masks, and high-noise perturbation filtering. This allows subsequent comparisons to distinguish noise-limited regimes, linear-baseline-saturated regimes, and regimes with residual headroom for more expressive models.

Metric selection. No single metric captures all failure modes of a perturbation predictor: a model may have low discriminative power, collapse variance across genes, or fail to recover the direction of change. From over sixty candidate metrics surveyed in the literature (Appendix F.6), we retain eight that pass DRF calibration and span six complementary categories: point-wise error (MSE_{wt}), correlation ($\text{Pearson}_{\text{wt}}$), concordance ($\text{CCC}_{\text{wt}}$), direction (fraction correct direction), discrimination (cosine rank), differentially

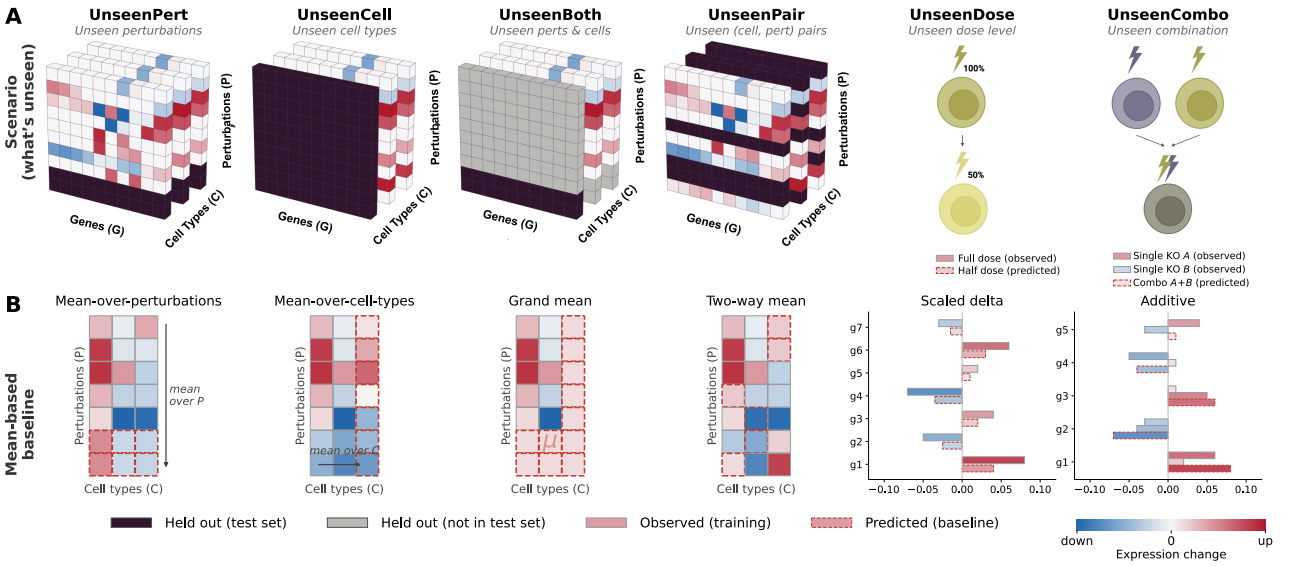

Figure 1. **Prediction regimes and non-learned baselines. A.** Schematic overview of the six prediction regimes used to evaluate perturbation-response prediction and their train/test composition. Each regime holds out a different axis of the perturbation-effect tensor: unseen perturbations (UNSEENPERT), unseen cellular contexts (UNSEENCELL), both unseen perturbations and contexts (UNSEENBOTH), unseen perturbation–context pairs (UNSEENPAIR), unseen dose levels (UNSEENDOSE), and unseen combinatorial perturbations (UNSEENCOMBO). Colored entries denote observed training responses, dark entries denote held-out test responses, and grey entries denote held-out responses that are not part of the test set. **B.** Scenario-matched non-learned baselines quantify simple recoverable structure under each regime. For each held-out setting, the baseline predicts the test response using only the structure available in the training tensor: mean-over-perturbations for UNSEENPERT, mean-over-cell-types for UNSEENCELL, the grand mean for UNSEENBOTH, a two-way mean for UNSEENPAIR, scaled deltas for UNSEENDOSE, and additive single-perturbation effects for UNSEENCOMBO. Blue and red indicate down- and up-regulated expression changes, respectively; dashed outlines indicate baseline-predicted held-out responses.

expressed gene recall (GSEA_{up}, $\text{GSEA}_{\text{down}}$), and distributional fit (energy distance in PCA space). Evaluating across these calibrated metrics yields a summary that is more robust to metric-specific artifacts and better reflects the recovery of reproducible signal beyond noise and simple structure.

Baselines. Perturbation-prediction benchmarks compare models against baselines that play distinct roles in the evaluation. We distinguish *non-learned baselines*, which require no parameter fitting and predict by averaging or algebraically combining observed responses along axes available in the training tensor, from *learned baselines*, which fit parameters while imposing simple structural assumptions, such as linearity, shrinkage, target scaling, or additivity (Adduri et al., 2025; Kernfeld et al., 2024; Ahlmann-Eltze et al., 2025; Vollenweider & Bühlmann, 2026). Non-learned baselines quantify what is recoverable from the train/test geometry alone while learned baselines test whether high-capacity models improve over simpler predictors.

For each prediction regime, we define a matched non-learned baseline that captures the simplest structure available under that split: Mean-over-Perturbations (MoP), Mean-over-Cell-Types (MoCT), Grand Mean, Two-way Mean, Scaled Delta, and Additive (Figure 1B; Table 13). *Learned baselines* fit parameters on the training split but remain far simpler than deep learning models: Ridge, Linear Additive, Correlation, and Latent Additive. These baselines provide anchors for quantifying how much reproducible perturbation signal is already explained by simple structure.

3. Regime-aware evaluation by Baseline Saturation

Perturbation-response datasets contain different mixtures of measurement noise, linear structure, and heterogeneous signal. In noise-limited regimes, little reproducible signal is available for any predictor. In baseline-saturated regimes, reproducible signal exists but is already captured by the linear component. The relevant modeling frontier is the remaining signal that is both reproducible and not explained by the matched non-learned baseline. We quantify this frontier using *Baseline Saturation*.

Baseline Saturation. For perturbation p and a calibrated higher-is-better metric, Baseline Saturation is the fraction of the control-bounded dynamic range captured by the scenario-matched baseline:

$$\text{Baseline Saturation}_p = \frac{\text{baseline}_p - \text{neg}_p}{\text{pos}_p - \text{neg}_p + \varepsilon}, \qquad (1)$$

where neg_p and pos_p are the negative- and positive-control metric values, and baseline_p is the value for the scenario-matched baseline (see Table 13).

A value $\text{BaselineSaturation}_p \approx 1$ indicates that the baseline approaches the empirical reproducibility ceiling, leaving little measurable headroom for more expressive models. A value $\text{BaselineSaturation}_p \approx 0$ indicates a *baseline-resistant* perturbation: the response is measurable relative to noise, but not explained by the linear structure available under the evaluation-regime split. Since Baseline Saturation is normalized by the control gap, it is comparable across metrics, perturbations, datasets, and prediction regimes. We compute Baseline Saturation per perturbation and calibrated metric. For dataset-level summaries, values are clipped to $[0, 1]$ and aggregated by the median across perturbations and metrics, yielding one Baseline Saturation score per dataset–evaluation-regime pair.

4. Experimental setup

Our evaluation is designed to test how model performance depends on the composition of each prediction regime: measurement noise, linear structure, and heterogeneous signal.

Datasets. We evaluate nine publicly available single-cell CRISPR perturbation datasets: six single-cell-type datasets, including Adamson16 (Adamson et al., 2016), Frangieh21 (Frangieh et al., 2021), Norman19 (Norman et al., 2019), Sunshine23 (Sunshine & Bhatt, 2023), Wessels23 (Wessels et al., 2023), and Replogle20 (Replogle et al., 2020), and three multi-cell-type datasets, including Replogle22 (Replogle et al., 2022), McFaline23 (McFaline-Figueroa & Trapnell, 2019), and Jiang24 (Jiang et al., 2024). For Replogle22, we combine the K562 and RPE1 genome-scale Perturb-seq experiments into a single multi-cell-type dataset.

Each dataset is evaluated under all applicable prediction regimes. We use cross-validation within each regime so that every held-out unit—perturbation, cellular context, perturbation–context pair, dose, or combination—appears in the test set at least once across folds. Results are reported aggregated across folds, so that Baseline Saturation and model performance reflect the full diversity of perturbations in each dataset rather than a single arbitrary split. The UNSEENDOSE regime is not evaluated in this study because it requires datasets that report knockdown strength per perturbation; dual-guide CRISPRi screens such as X-Atlas (Huang et al., 2025) provide this information and are a natural target for future investigation.

Models and Baselines. We evaluate three published deep learning architectures: scGPT (Cui et al., 2024), GEARS (Roohani et al., 2024), and PRESAGE (Wu et al., 2024). All are trained on each applicable dataset following their respective protocols and evaluated using our unified metric pipeline. We compare these models against two baseline families. First, scenario-matched non-learned baselines require no parameter fitting and capture the simplest structure available under each split: Zero, Mean-over-Perturbations (MoP), Mean-over-Cell-Types (MoCT), Grand Mean, Two-way Mean, Scaled Delta, and Additive (Appendix F.4). These baselines define the reference used for Baseline Saturation. Second, learned baselines fit parameters on the training split while imposing simple linear assumptions: Ridge, Linear Additive, Correlation, and Latent Additive. Together, these baselines distinguish what is recoverable from train/test geometry alone from what can be learned by linear models.

5. Results

5.1. Prediction regimes differ dramatically in baseline saturation

Baseline Saturation reveals that perturbation-prediction benchmarks are not uniformly difficult. Instead, each dataset–regime pair contains a different mixture of noise-limited responses, simple linear structure, and heterogeneous baseline-resistant signal. Across nine datasets and five prediction regimes, the fraction of signal captured by simple baselines varies widely across regimes and, more importantly, across perturbations within the same regime (Figure 2).

Evaluation regimes span distinct saturation levels. At the regime level, UNSEENCOMBO is the easiest (pooled median Baseline Saturation $= 0.92$): the additive baseline captures most combinatorial effects, and 72% of perturbations are already saturated (Baseline Saturation > 0.66). This quantifies prior observations that deep models often fail to improve over additive baselines in combinatorial prediction (Vollenweider & Bühlmann, 2026). At the opposite extreme, UNSEENBOTH has the lowest saturation (pooled median $= 0.47$), with 47% of perturbations baseline-resistant, consistent with a setting in which neither perturbation-level nor cellular-context-level averages provide relevant information. UNSEENPERT, UNSEENPAIR, and UNSEENCELL lie between these extremes (pooled median Baseline Saturation 0.69, 0.54, and 0.57 respectively), but these aggregate values obscure substantial internal heterogeneity.

Within-regime heterogeneity is the dominant source of variation. The main finding is not only that prediction regimes differ in average saturation, but that each regime contains a heterogeneous mixture of perturbation types (Appendix Figure 5). Within a fixed dataset–regime combination, perturbation-level Baseline Saturation often spans

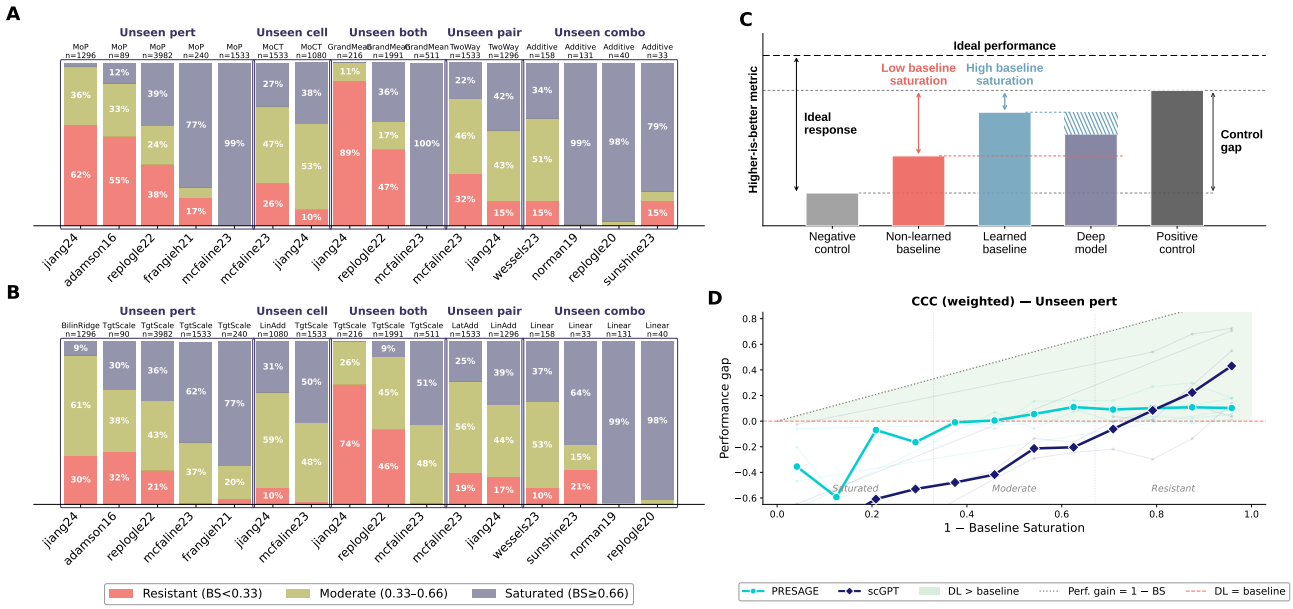

Figure 2. **Baseline Saturation reveals heterogeneous prediction difficulty and identifies where deep learning models add value. A.** Regime composition under non-learned baselines. Each bar shows the fraction of perturbations classified as baseline-resistant (BS < 0.33, red), moderate (0.33–0.66, olive), or saturated (BS > 0.66, purple) for each dataset, grouped by evaluation scenario. Baseline Saturation is computed as the 7-metric aggregate (MSE weighted, Pearson weighted, CCC weighted, frac. correct direction, cosine rank, GSEA up/down) using the scenario-matched non-learned baseline. Baseline names and sample sizes are annotated above each bar. **B.** Same as **A.** but using the best learned baseline per dataset–scenario pair. Learned baselines (e.g., TargetScaling, LinearAdditive) substantially increase saturation, narrowing the frontier where expressive models can add value. **C.** Schematic illustration of the Baseline Saturation framework. Performance is normalized between the negative control (zero prediction) and the positive control (Interpolated duplicate). Baseline Saturation measures how much of this control gap is captured by the scenario-matched baseline. The hatched region above the deep model bar indicates the portion of learned-baseline performance not yet recovered by the deep model. **D.** Performance gap of PRESAGE and scGPT over the non-learned baseline on CCC weighted in the UNSEENPERT scenario, as a function of perturbation difficulty (1 − Baseline Saturation). Bold lines show the cross-dataset aggregate (median per difficulty bin, averaged across datasets); faint traces show individual datasets. The green-shaded region marks where deep learning models outperform the baseline. DL models consistently underperform on saturated perturbations (left) but recover perturbation-specific signal on resistant perturbations (right).

much of the unit interval, with typical interquartile ranges of 0.3–0.9 and frequently skewed or bimodal distributions. For example, in Jiang24/UNSEENPERT, 65% of perturbations are unsaturated (Baseline Saturation < 0.33), whereas only 14% are saturated (Baseline Saturation > 0.66). The same prediction regime can have a very different composition in other datasets. McFaline23/UNSEENPAIR, for instance, is explicitly bimodal, with 52% unsaturated and 29% saturated perturbations, and little mass between these regimes.

Across all 15,663 perturbation evaluations, the evaluation set contains both perturbations that are largely explained by simple baselines and perturbations for which substantial signal remains unexplained. Thus, a single aggregate benchmark score averages over qualitatively different prediction problems: saturated perturbations with little remaining headroom, intermediate perturbations partially captured by simple structure, and unsaturated perturbations that define the residual modeling frontier.

UNSEENPERT mixes solved and resistant perturbations. This issue is especially important for UNSEENPERT, the regime most commonly used to evaluate perturbation-response models. Its aggregate saturation is moderate, but its composition is highly mixed: 50.9% of perturbations are saturated (Baseline Saturation ≥ 0.66), while 37.7% are baseline-resistant (Baseline Saturation < 0.33). This bimodality suggests that some unseen perturbations produce responses close to the average perturbation effect, whereas others induce idiosyncratic, reproducible responses not captured by the mean baseline. Aggregate comparisons in UNSEENPERT can therefore understate model value on the resistant subset or overstate it by rewarding performance on already-saturated perturbations. Stratifying by saturation is necessary to determine where a model is actually learning beyond linear structure.

Learned baselines narrow the modeling frontier. Baseline Saturation depends on the baseline used to define headroom and the prediction-regime evaluated. When we replace non-learned averages with learned baselines such as Ridge

and Linear Additive models, saturation increases substantially. Thus, many perturbations that appear resistant to mean-based baselines are explained by simple linear structure. The frontier for high-capacity models is, therefore, the heterogeneous signal that remains beyond both non-learned and linear learned baselines.

Results are stable across cross-validation folds (median CV $= 5\%$), with approximately $5\times$ higher variability in scenarios that rotate cell types across folds (UNSEENCELL, UNSEENBOTH), reflecting that cell types differ in how well cross-cell-type averages approximate their perturbation responses (Appendix B). Biological characterization of baseline-resistant perturbations is provided in Section 5.3.

5.2. Expressive models add value primarily beyond the baseline frontier

The heterogeneity in Baseline Saturation has a direct consequence for model evaluation. A test set mixes easy perturbations, where baselines are near-optimal, with hard ones, where baselines are uninformative. Aggregate metrics weight all perturbations equally, so the result is dominated by whichever regime is more common. Because easy perturbations typically make up 50–90% of the test set, the aggregate reflects the regime where baselines already solve the problem and DL models can only add noise. A DL model that learns real perturbation-specific signal on hard perturbations can still appear inferior overall if that same signal introduces variance on easy ones.

To separate these regimes, we stratify model performance by Baseline Saturation. For each test perturbation p we compute the *performance gain* $(\mathrm{DL}_p - \mathrm{baseline}_p) \,/\, (\mathrm{pos}_p - \mathrm{neg}_p + \varepsilon)$, the improvement of the DL model over the scenario-matched baseline normalized by the control gap (sign-flipped for lower-is-better metrics; positive values mean the DL model outperforms the baseline). Perturbations where the positive control does not outperform the negative control by at least 0.05 are excluded, as the metric has insufficient dynamic range to support reliable comparisons.

DL models outperform baselines on hard perturbations. Figure 3 shows a consistent pattern across scenarios and metrics: the performance gain increases with perturbation difficulty. On easy perturbations (left side), all DL models fall below the baseline. On hard perturbations (Baseline Saturation < 0.33, right side), DL models outperform the baseline on the majority of perturbations for five of six metrics.

On $\mathrm{Pearson}_{\mathrm{wt}}$ in UNSEENPERT, PRESAGE, GEARS, and scGPT outperform the baseline on 80%, 68%, and 68% of hard perturbations (median performance gain $+0.30$, $+0.15$, $+0.21$). On fraction of correct direction, which measures whether the model identifies the right sign of gene-level change, win rates reach 71–75% across all three models. $\mathrm{GSEA}_{\mathrm{down}}$ shows the strongest advantage: DL models win on 69–80% of hard perturbations in UNSEENPERT and 61–91% in UNSEENPAIR. This is expected, because repression cascades are perturbation-specific and cannot be approximated by averaging over the training set. The full per-scenario and per-metric breakdown is in Table 12 (Appendix D).

The advantage is in direction, not magnitude. The pattern of win rates across metrics reveals what DL models actually learn on hard perturbations. Metrics that measure the direction and shape of the response show the strongest gains: fraction of correct direction (71–85%), $\mathrm{GSEA}_{\mathrm{down}}$ (61–91%), and Pearson correlation (50–80%). In contrast, MSE, which penalizes magnitude errors quadratically, shows win rates near or below chance for GEARS and scGPT. This means DL models learn *which genes respond and in which direction*, but do not reliably predict the size of the response. A model that gets the sign right but overestimates the magnitude scores worse on MSE than a baseline that predicts zero change, even though the model's prediction is biologically more useful.

The gap between $\mathrm{GSEA}_{\mathrm{up}}$ and $\mathrm{GSEA}_{\mathrm{down}}$ win rates (48–75% vs. 69–91% in UNSEENPERT) supports this further: baselines already capture upregulation signals, which tend to overlap with the shared mean response, but miss downregulation cascades, which are perturbation-specific.

Aggregation masks these gains. The monotonic relationship between difficulty and performance gain (Figure 3) means that aggregate metrics mix two regimes with opposite conclusions. In nearly 30% of testable (dataset, scenario, metric, model) combinations, the DL model loses in aggregate but wins on the hard subset (Appendix E). This is not a failure of DL models but an artifact of how we evaluate them. Baseline Saturation provides the stratification needed to separate these regimes and identify where expressive models add value.

5.3. What characterizes the baseline-resistant frontier?

The previous sections establish that Baseline Saturation varies widely within prediction regimes and that deep learning models gain primarily on resistant perturbations. A natural question follows: what distinguishes resistant from saturated perturbations? Throughout this section, we restrict to DRF-passing perturbations (DRF > 0.1) to ensure that BS values reflect genuine signal rather than noise-floor artifacts.

Difficulty is cell-type-dependent, not gene-intrinsic. If Baseline Saturation were determined solely by which gene

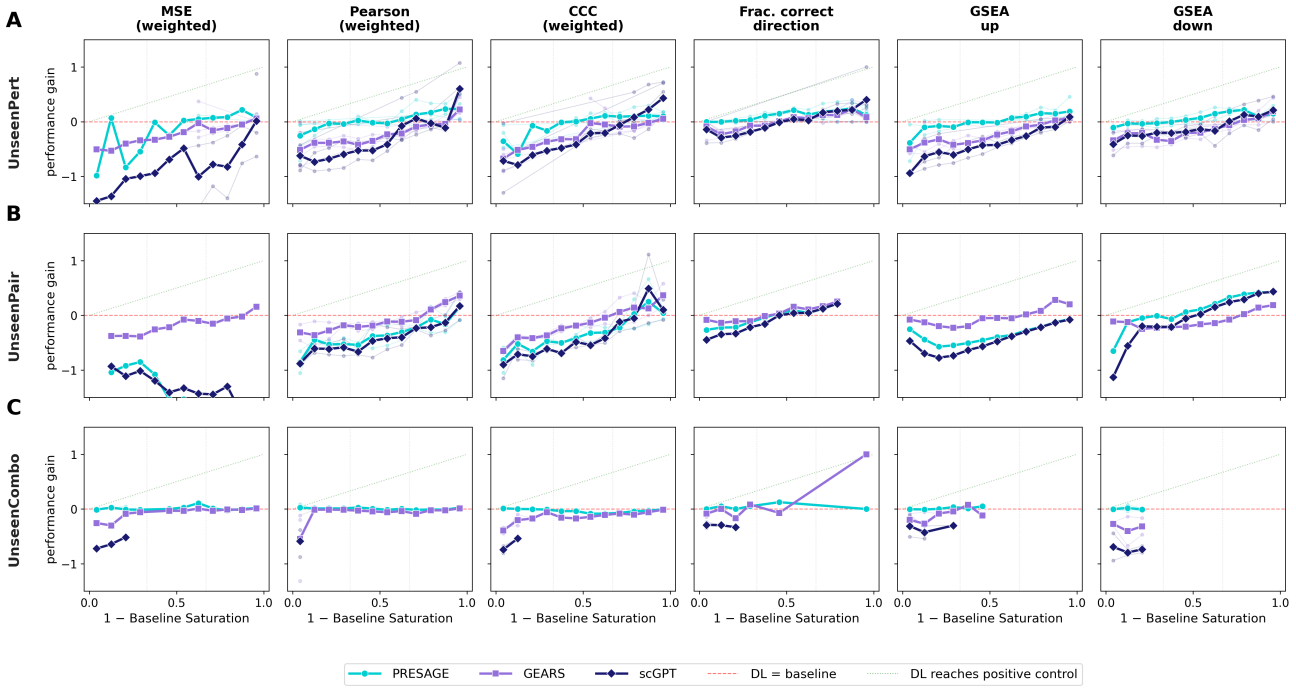

Figure 3. **Performance gain of DL models over the scenario-matched baseline as a function of perturbation difficulty. A.** UNSEENPERT: Adamson16, Frangieh21, Jiang24, McFaline23, Replogle22. **B.** UNSEENPAIR: Jiang24, McFaline23. **C.** UNSEENCOMBO: Norman19, Wessels23, Sunshine23. Each column shows one metric. The x-axis is $1 - \text{Baseline Saturation}$, binned into 12 equal-width bins. The y-axis is the performance gain: $(\text{DL} - \text{baseline})/(\text{pos} - \text{neg})$, sign-flipped for lower-is-better metrics. Per-perturbation values are aggregated to a median per dataset per bin, then averaged across datasets with equal weight. Faint lines show individual datasets; bold lines show the cross-dataset mean. Perturbations where the positive control does not outperform the negative by at least 0.05 are excluded (DRF filter applied).

is knocked out, the same perturbation should have similar saturation across cell types. We test this on two multi-cell-type datasets (Figure 4A–B). In Jiang24 (6 cell types), the concordance of per-gene Baseline Saturation rankings is low (Kendall $W = 0.30$), with pairwise Spearman correlations ranging from -0.44 to 0.84 (Figure 4A). Notably, BXPC3 shows negative correlations with most other cell types: perturbations that are easy in BXPC3 tend to be hard elsewhere, and vice versa. In Replogle22 (K562 vs. RPE1), the correlation is moderate ($\rho = 0.31$, $p < 10^{-38}$, $n = 1{,}379$; Figure 4B): many genes that are easy in K562 are hard in RPE1. These results confirm that whether a perturbation is baseline-resistant depends on the regulatory context in which it acts, not only on the identity of the perturbed gene.

Difficulty is driven by effect uniqueness, not effect strength. Under UNSEENPERT, the scenario-matched baseline is Mean-over-Perturbations (MoP), so perturbations whose response resembles the cell-type-specific mean are well-predicted by construction. Accordingly, the cosine similarity between each perturbation's delta and MoP is moderately correlated with Baseline Saturation ($\rho = 0.55$; Figure 4C): perturbations with high cosine similarity cluster at high BS, while perturbations with unique transcriptomic signatures are predominantly resistant. This relationship is expected given that MoP is the scenario-matched baseline, but the decomposition into direction and magnitude is informative. Effect *strength* does not predict difficulty in the same way: the L2 norm of the perturbation delta is only weakly correlated with Baseline Saturation ($\rho = 0.05$). In contrast, the number of differentially expressed genes is negatively correlated with Baseline Saturation ($\rho = -0.53$): perturbations that affect many genes tend to be baseline-resistant, because their broad transcriptomic effects are less likely to resemble the training-set average. Baseline-resistant perturbations are therefore those with *unique* transcriptomic responses—profiles that point in directions not captured by averaging—regardless of whether the overall effect magnitude is large or small.

Difficulty is a joint function of perturbation, context, and evaluation regime. The cell-type dependence of Baseline Saturation directly explains the large fold-to-fold variation observed for UNSEENCELL (Appendix B). When the held-out cell type has perturbation responses similar to the training cell types, cross-cell-type averaging (MoCT) works well and the regime is saturated. When the held-out cell type responds differently, the same averaging is uninformative

Table 1. **Absolute predictor performance stratified by Baseline Saturation across three generalization scenarios.** Median across DRF-passing perturbations (DRF > 0.1), pooled across datasets within each scenario. *Baseline* denotes the scenario-matched simple-rule predictor. **Bold** marks the best predictor per metric (excluding Tech-dup).

	UNSEENPERT				UNSEENPAIR				UNSEENCOMBO			
	CCC ↑	MSE ↓	Pears ↑	FCD ↑	CCC ↑	MSE ↓	Pears ↑	FCD ↑	CCC ↑	MSE ↓	Pears ↑	FCD ↑
Hard (Baseline Saturation < 0.33; $n = 779$; 186; 31)												
Zero	0.00	0.047	0.00	0.00	0.00	0.046	0.00	0.00	0.00	0.233	0.00	0.00
Baseline	-0.02	0.048	-0.07	0.42	0.17	0.038	0.06	0.59	0.05	0.232	0.09	0.46
BilinearRidge	0.01	**0.047**	0.03	0.55	**0.42**	**0.023**	**0.12**	**0.76**	0.02	0.234	0.07	0.58
Correlation	0.01	0.054	0.08	0.50	-0.01	0.049	0.00	0.40	-0.01	0.233	-0.05	0.39
PRESAGE	**0.06**	0.050	**0.11**	**0.60**	-0.05	0.187	0.02	0.48	0.02	0.232	0.12	0.52
scGPT	-0.01	0.100	-0.02	0.50	-0.08	0.149	-0.00	0.47	**0.09**	**0.018**	**0.15**	0.42
GEARS	-0.01	0.049	-0.02	0.50	0.12	0.037	0.08	0.57	0.04	0.234	0.09	**0.69**
Tech-dup	0.89	0.004	0.91	1.00	0.98	0.001	0.35	1.00	0.97	0.011	0.98	1.00
Moderate ($0.33 \leq$ Baseline Saturation < 0.66; $n = 812$; 153; 83)												
Zero	0.00	0.068	0.00	0.00	0.00	0.047	0.00	0.00	0.00	0.255	0.00	0.00
Baseline	0.13	**0.053**	0.51	0.82	**0.53**	**0.022**	**0.56**	**0.78**	**0.27**	0.210	0.37	0.90
BilinearRidge	0.13	0.057	**0.56**	0.85	-0.07	0.064	-0.01	0.55	0.08	0.238	0.24	0.67
Correlation	0.05	0.079	0.19	0.60	0.00	0.049	0.01	0.46	0.03	0.257	0.14	0.71
PRESAGE	**0.25**	0.059	0.40	**0.85**	0.02	0.134	0.04	0.61	0.20	0.211	0.37	**0.92**
scGPT	0.09	0.101	0.10	0.59	-0.08	0.131	-0.05	0.49	0.22	**0.050**	**0.44**	0.69
GEARS	0.10	0.061	0.26	0.65	0.30	0.034	0.17	0.66	0.13	0.221	0.29	0.87
Tech-dup	0.81	0.012	0.88	1.00	0.98	0.001	0.96	1.00	0.97	0.014	0.97	1.00
Easy (Baseline Saturation ≥ 0.66; $n = 842$; 103; 204)												
Zero	0.00	0.038	0.00	0.00	0.00	0.095	0.00	0.00	0.00	0.114	0.00	0.00
Baseline	0.15	**0.030**	0.27	**1.00**	**0.69**	**0.034**	**0.78**	**0.82**	**0.92**	**0.013**	**0.94**	**0.99**
BilinearRidge	0.17	0.032	**0.28**	**1.00**	-0.18	0.162	-0.21	0.61	0.33	0.083	0.74	0.84
Correlation	0.03	0.040	0.03	0.85	-0.01	0.114	-0.16	0.28	-0.03	0.133	-0.25	0.29
PRESAGE	**0.18**	0.036	0.12	0.99	0.27	0.177	0.30	0.64	0.87	0.015	0.91	0.99
scGPT	0.08	0.060	0.06	0.77	0.09	0.187	0.10	0.54	0.27	0.075	0.54	0.65
GEARS	0.10	0.036	0.09	0.73	0.31	0.062	0.57	0.71	0.58	0.051	0.69	0.85
Tech-dup	0.33	0.019	0.30	1.00	0.99	0.002	0.99	1.00	0.97	0.007	0.97	1.00

and the regime becomes resistant. Mechanistically, each cell type has a different MoP baseline, so a perturbation whose response aligns with the cell-type-specific mean in one context may diverge from it in another. The same logic applies to UNSEENCOMBO: when single-perturbation effects compose additively (as in Norman19, Baseline Saturation $= 0.98$), the Additive baseline saturates the regime. When epistasis is strong (as in parts of Wessels23, where 56% of perturbations are resistant), the combinatorial regime retains genuine headroom. In both cases, the OOD label describes what is absent from training, not how much modeling capacity is required. Difficulty is therefore not a fixed property of a gene or perturbation—it is a joint function of the perturbation identity, the cellular context, and the evaluation regime under which performance is measured.

6. Discussion

Resolving the baselines debate. Our results show that apparently conflicting conclusions about deep learning and simple baselines can hold simultaneously. In baseline-saturated regimes, simple predictors capture most of the signal, leaving little measurable headroom for expressive models. In baseline-resistant regimes, reproducible perturbation-specific structure remains beyond these predictors, and expressive models can add value. Baseline Saturation therefore acts as a conditioning variable for model comparison: the same model can appear competitive or uncompetitive depending on the saturation composition of the test set. The relevant question is not whether deep models beat baselines on average, but where they do so and what characterizes that boundary.

The OOD illusion. The "out-of-distribution" label commonly applied to test splits conflates distributional novelty with modeling demand. UNSEENCOMBO is distributionally novel, since double perturbations are absent from training, yet the Additive baseline saturates most perturbations (median Baseline Saturation $= 0.92$) because single-perturbation effects compose nearly additively. UNSEENPERT, which simply holds out perturbations, retains substantial headroom because predicting the effect of a never-seen perturbation from the training-set average is genuinely hard. UNSEENCELL is particularly instructive: its diffi-

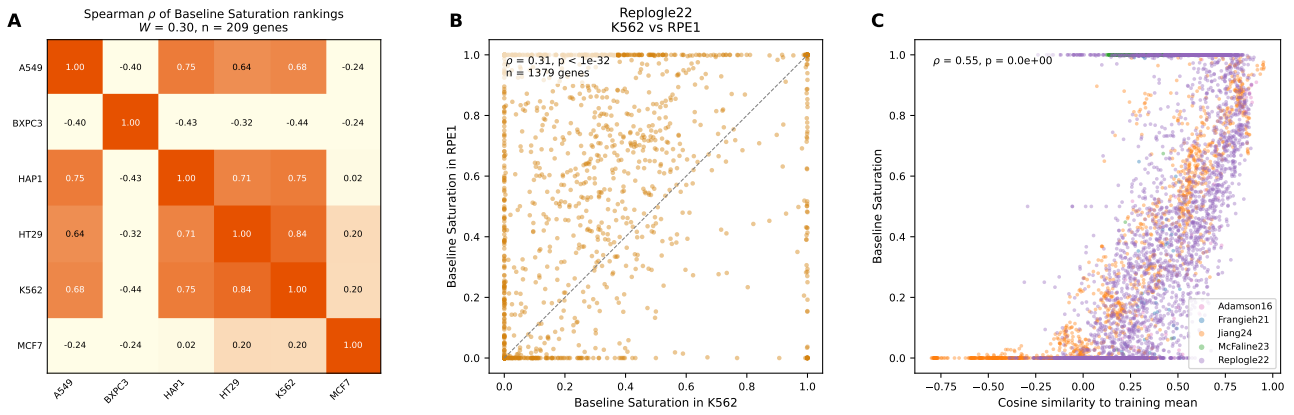

Figure 4. **Biological characterization of the baseline-resistant frontier** (DRF-passing perturbations only, DRF > 0.1). **A.** Pairwise Spearman correlation heatmap of per-gene Baseline Saturation across 6 cell types in Jiang24 (Kendall $W = 0.30$, $n = 209$). BXPC3 shows negative correlations with most other cell types. **B.** Cross-cell-type scatter for Replogle22 (K562 vs. RPE1, $\rho = 0.31$, $n = 1{,}379$). **C.** Baseline Saturation versus cosine similarity to the training-set mean (MoP) across five UNSEENPERT datasets ($\rho = 0.55$). Per-perturbation BS is the median across six calibrated metrics and cross-validation folds, clipped to $[0, 1]$. Perturbations resembling the average cluster at high BS; those with unique responses are predominantly resistant.

culty varies by more than 0.3 across cross-validation folds because the identity of the held-out cell type determines whether cross-cell-type averaging is informative. Under UNSEENPERT, the correlation between Baseline Saturation and cosine similarity to the mean baseline is partly definitional: MoP is the scenario-matched baseline, so perturbations resembling the mean are well-predicted by construction. The non-trivial findings are that (i) effect *strength* does not predict difficulty—only the *uniqueness* of the response direction does—and (ii) difficulty is cell-type-dependent rather than gene-intrinsic: the same perturbation can be saturated in one cellular context and resistant in another (Section 5.3). Benchmarks that report a single aggregate score over these mixtures cannot distinguish genuine model contributions from artifacts of test-set composition.

Limitations and future work. Baseline Saturation depends on the quality of the positive control. When few cells are available per perturbation, the split-half replicate is noisy, which can inflate saturation estimates. We mitigate this through DEG-weighted and masked metrics that concentrate signal above the noise floor and verify that DRF embeds the noise floor on pseudobulk data (Appendix F.6). GSEA metrics require a minimum number of differentially expressed genes per perturbation; for datasets with weak perturbation effects (notably McFaline23 and Frangieh21), many perturbations did not meet this threshold and were evaluated without GSEA, reducing the metric coverage for those datasets. Our analysis evaluates three DL architectures (scGPT, GEARS, PRESAGE); other models may behave differently. The biological characterization identifies what predicts resistance (deviation from the shared response) but not why specific genes produce unique responses; incorporating gene regulatory network structure is an important next step. Finally, we use non-learned baselines as the primary reference; the frontier narrows further with learned baselines (Ridge, Linear Additive), suggesting that much of the remaining headroom is capturable by linear models.

Conclusion. Perturbation prediction benchmarks should move from asking whether deep models beat baselines to asking where they do so and what predicts the boundary. Baseline Saturation provides a principled answer: strong simple baselines define the solved regime, and their failures define the modeling frontier.

Impact Statement

Our results suggest three changes to how perturbation prediction models are evaluated. (1) **Report Baseline-Saturation-stratified results**: at minimum, performance on the resistant subset (Baseline Saturation < 0.33) should accompany the aggregate, since the three-regime breakdown (resistant, moderate, saturated) reveals structure that aggregate reporting hides. (2) **Use qualitative metrics alongside quantitative ones**: the deep learning advantage on hard perturbations is about learning which genes respond and in which direction, not about predicting exact magnitudes. Directional metrics (fraction correct direction) and DEG-recall metrics (GSEA) detect this advantage far more often than MSE (40% vs. 5% reversal rate). (3) **Focus model development on the baseline-resistant frontier**: perturbations where baselines succeed are already solved; the frontier where expressive models can add value consists of perturbations with unique, context-specific transcriptomic signatures.

References

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167:1867–1882, 2016. doi: 10.1016/j.cell.2016.11.048.

Adduri, A. K., Gautam, D., Bevilacqua, B., Imran, A., Shah, R., Naghipourfar, M., Teyssier, N., Ilango, R., Nagaraj, S., Ricci-Tam, C., Carpenter, C., Subramanyam, V., Winters, A., Dong, M., Tirukkovalur, S., Sullivan, J., Plosky, B. S., Eraslan, B., Youngblut, N. D., Leskovec, J., Gilbert, L. A., Konermann, S., Hsu, P. D., Dobin, A., Burke, D. P., Goodarzi, H., and Roohani, Y. H. Predicting cellular responses to perturbation across diverse contexts with STATE. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135.

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

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Frangieh, C. J., Melms, J. C., Thakore, P. I., Geiger-Schuller, K. R., Ho, P., Luoma, A. M., Cleary, B., Jerber, J., Mead, B. E., et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53:332–341, 2021. doi: 10.1038/s41588-021-00779-1.

Huang, Q. et al. X-Atlas: A large-scale dual-guide CRISPR perturbation atlas. *bioRxiv*, 2025.

Jiang, S. et al. Perturb-bench: A comprehensive benchmark for single-cell perturbation prediction. *bioRxiv*, 2024.

Kernfeld, E., Yang, Y., Weinstock, J., Battle, A., and Cahan, P. A systematic comparison of computational methods for expression forecasting. *bioRxiv*, 2024. doi: 10.1101/2023.07.28.551039.

Liang, Y. and Singh, R. Perturbation specificity index: A measure of perturbation complexity for single-cell perturbation prediction. *bioRxiv*, 2026.

McFaline-Figueroa, J. L. and Trapnell, C. A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nature Genetics*, 51:1389–1398, 2019.

Miller, D., Hicks, S., Greene, C. S., Atta, L., and Lotfollahi, M. Simple baselines outperform complex models for in silico perturbation prediction. *bioRxiv*, 2025. doi: 10.1101/2025.01.13.632762.

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., and Weissman, J. S. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

Replogle, J. M., Norman, T. M., Xu, A., Hussmann, J. A., Chen, J., Cogan, J. Z., Meer, E. J., Terry, J. M., Riordan, D. P., Srinivas, N., et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nature Biotechnology*, 38:954–961, 2020. doi: 10.1038/s41587-020-0470-y.

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., et al. Mapping information-rich genotype–phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575, 2022. doi: 10.1016/j.cell.2022.05.013.

Roohani, Y., Huang, K., and Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-02008-y.

Squires, C., Fet, N., Roeder, K., and Uhler, C. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 36, 2024. NeurIPS 2023.

Sunshine, H. and Bhatt, S. Combinatorial perturbation analysis in human primary T cells. *bioRxiv*, 2023.

Vollenweider, V. and Bühlmann, P. Signal, bounds, and baselines: Rigorous evaluation of perturbation prediction methods. *bioRxiv*, 2026. ETH Zürich.

Wessels, H.-H., Methot, S. P., and Satija, R. Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA knockdown. *Nature Biotechnology*, 41:1149–1159, 2023. doi: 10.1038/s41587-022-01643-x.

Wolf, F. A., Angerer, P., and Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: 10.1186/s13059-017-1382-0.

Wu, Z., Huang, L., Craciun, G., et al. PRESAGE: Perturbation response estimation via autoregressive gene expression. *bioRxiv*, 2024.

A. Additional results

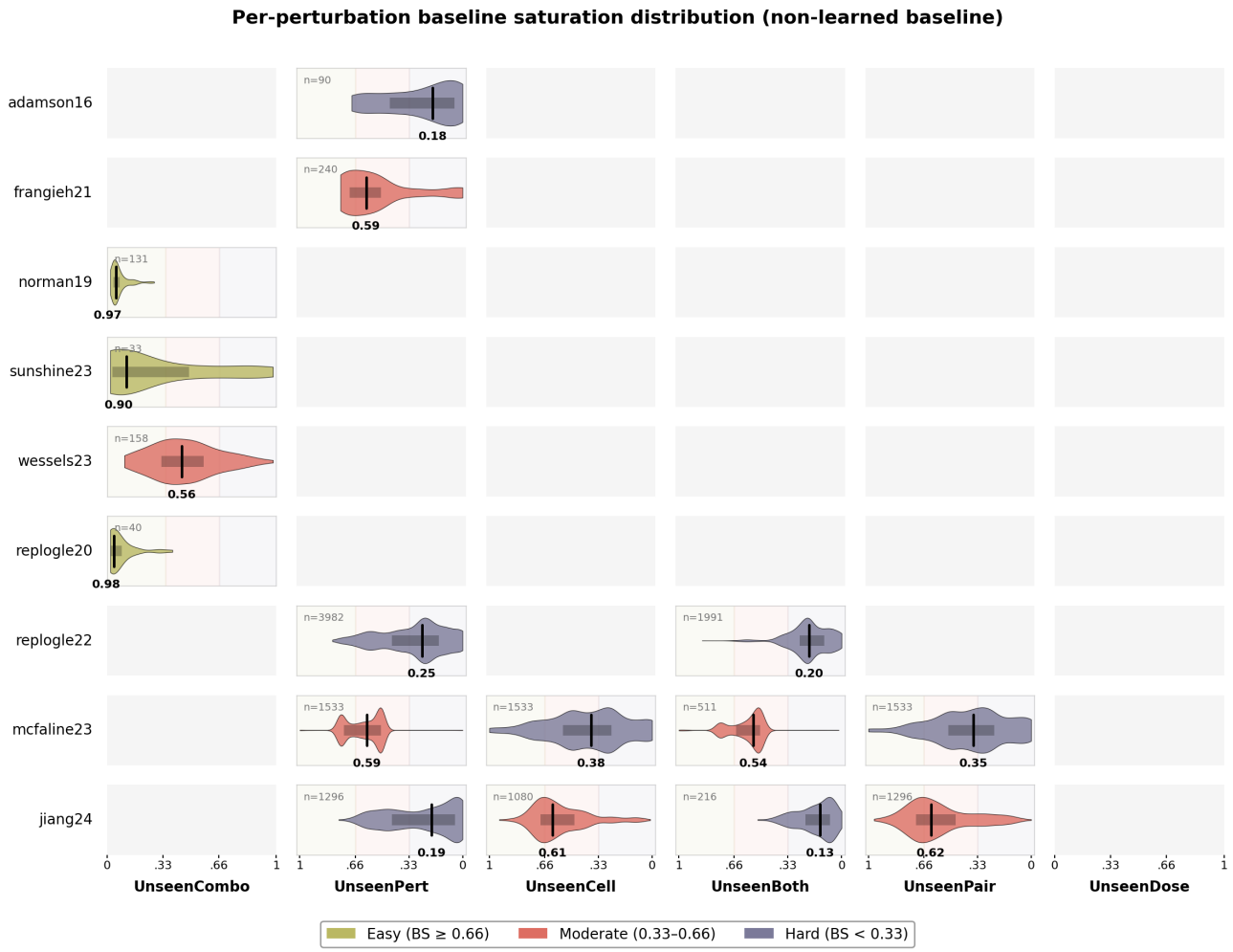

Figure 5. Per-perturbation Baseline Saturation distributions across datasets and prediction regimes. Each violin shows the full distribution for one dataset/regime combination. The wide spread and bimodality show that aggregate scores mask within-split heterogeneity.

B. Cross-fold stability of Baseline Saturation

Table 2. Cross-fold stability of Baseline Saturation.

Dataset	Scenario	Folds	Mean	Median	Std	CV
UNSEENCOMBO						
norman19	UnseenCombo	2	0.966	0.966	0.001	0.1%
replogle20	UnseenCombo	2	0.979	0.979	0.006	0.6%
wessels23	UnseenCombo	2	0.558	0.558	0.015	2.6%
sunshine23	UnseenCombo	2	0.860	0.860	0.078	9.1%
UNSEENPERT						
replogle22	UnseenPert	5	0.250	0.250	0.000	0.0%
mcfaline23	UnseenPert	5	0.586	0.589	0.009	1.5%
frangieh21	UnseenPert	5	0.590	0.613	0.039	6.7%
jiang24	UnseenPert	5	0.183	0.199	0.034	18.5%
adamson16	UnseenPert	5	0.202	0.204	0.046	23.0%
UNSEENCELL						
mcfaline23	UnseenCell	3	0.377	0.364	0.028	7.3%
jiang24	UnseenCell	5	0.611	0.589	0.045	7.4%
UNSEENBOTH						
replogle22	UnseenBoth	5	0.200	0.200	0.003	1.6%
mcfaline23	UnseenBoth	5	0.526	0.536	0.036	6.8%
jiang24	UnseenBoth	5	0.120	0.139	0.040	33.4%
UNSEENPAIR						
mcfaline23	UnseenPair	5	0.349	0.349	0.004	1.3%
jiang24	UnseenPair	5	0.595	0.597	0.005	0.8%
Scenarios with cell-type rotation						*med. 7.3%*
Scenarios without cell-type rotation						*med. 1.5%*

Table 2 reports the mean, median, standard deviation, and coefficient of variation (CV) of per-fold median Baseline Saturation. Baseline Saturation is stable across folds (overall median CV = 5%). Scenarios that rotate cell types across folds (UNSEENCELL, UNSEENBOTH) show $\sim 5\times$ higher variability (median CV = 7%) than those that do not (median CV = 1.5%), because the held-out cell type determines how well cross-cell-type averages approximate the perturbation response.

C. Raw metric scores

Table 3. Adamson16

	$\mathrm{MSE}_{\mathrm{wt}}$	$\mathrm{Pearson}_{\mathrm{wt}}$	$\mathrm{CCC}_{\mathrm{wt}}$	Frac. dir.	Cos. rank	$\mathrm{GSEA}_{\uparrow}$	$\mathrm{GSEA}_{\downarrow}$
UNSEENPERT ($n=80$, 5f)							
Zero	0.0302	0.000	0.000	0.000	0.500	0.090	0.079
Tech-dup	0.0012	0.981	0.978	0.969	0.000	0.882	0.869
Interp-dup	0.0018	0.968	0.965	0.969	0.000	0.916	0.906
MoP	0.0324	0.103	0.042	0.379	0.500	0.419	0.164
Target Scaling	0.0178	0.561	0.315	0.379	0.000	0.419	0.244
GEARS	0.0236	0.010	0.004	0.354	0.438	0.375	0.166
scGPT	0.0148	0.360	0.262	0.392	0.143	0.338	0.183
PRESAGE	0.0295	0.128	0.061	0.411	0.118	0.461	0.269
Ridge	0.0324	0.103	0.042	0.379	0.500	0.419	0.164
Bilinear Ridge	0.0248	0.254	0.112	0.473	0.000	0.534	0.292
Linear Add.	0.0324	0.103	0.042	0.379	0.500	0.419	0.164
Correlation	0.0490	0.199	0.119	0.614	0.559	0.113	0.225
Latent Add.	0.0285	0.101	0.019	0.389	0.529	0.420	0.163

Table 4. Frangieh21

	$\mathrm{MSE}_{\mathrm{wt}}$	$\mathrm{Pearson}_{\mathrm{wt}}$	$\mathrm{CCC}_{\mathrm{wt}}$	Frac. dir.	Cos. rank	$\mathrm{GSEA}_{\uparrow}$	$\mathrm{GSEA}_{\downarrow}$
UNSEENPERT ($n = 39$, 5f)							
Zero	0.0037	0.000	0.000	0.000	0.500	0.211	—
Tech-dup	0.0035	0.215	0.188	1.000	0.213	1.000	—
Interp-dup	0.0030	0.280	0.101	1.000	0.330	1.000	—
MoP	0.0033	0.180	0.067	0.500	0.500	0.099	—
Target Scaling	0.0032	0.273	0.108	1.000	0.032	0.099	—
GEARS	0.0045	0.037	0.021	0.000	0.564	0.214	—
scGPT	0.0041	0.167	0.120	1.000	0.209	—	—
PRESAGE	0.0034	0.161	0.059	0.500	0.444	0.200	—
Ridge	0.0033	0.180	0.067	0.500	0.500	0.099	—
Bilinear Ridge	0.0031	0.288	0.103	1.000	0.000	0.955	—
Linear Add.	0.0033	0.180	0.067	0.500	0.500	0.099	—
Correlation	0.0062	0.043	0.017	1.000	0.415	0.420	—
Latent Add.	0.0035	0.181	0.023	0.250	0.489	0.100	—

Table 5. Jiang24

	MSE_{wt}	$Pearson_{wt}$	CCC_{wt}	Frac. dir.	Cos. rank	$GSEA_{\uparrow}$	$GSEA_{\downarrow}$
UNSEENPERT ($n=216$, 5f)							
Zero	0.0560	0.000	0.000	0.000	0.500	0.066	0.071
Tech-dup	0.0012	0.988	0.988	1.000	0.000	0.942	0.928
Interp-dup	0.0015	0.987	0.986	1.000	0.000	0.971	0.969
MoP	0.0555	0.114	0.025	0.510	0.500	0.420	0.298
Target Scaling	0.0552	0.134	0.034	0.513	0.143	0.420	0.302
GEARS	0.0552	0.026	0.007	0.500	0.405	0.307	0.185
scGPT	0.1505	-0.014	-0.008	0.497	0.571	0.039	0.492
PRESAGE	0.0434	0.534	0.182	0.664	0.190	0.506	0.373
Ridge	0.0555	0.110	0.025	0.511	0.500	0.420	0.298
Bilinear Ridge	0.0534	0.609	0.077	0.704	0.071	0.684	0.413
Linear Add.	0.0555	0.114	0.025	0.510	0.500	0.420	0.298
Correlation	0.0621	-0.160	-0.009	0.372	0.667	0.032	0.057
Latent Add.	0.0574	-0.190	-0.016	0.486	0.524	0.386	0.140
UNSEENCELL ($n=216$, 5f)							
Zero	0.0569	0.000	0.000	0.000	0.500	0.058	0.059
Tech-dup	0.0010	0.991	0.990	1.000	0.000	0.938	0.924
Interp-dup	0.0013	0.991	0.990	1.000	0.000	0.972	0.969
MoCT	0.0332	0.639	0.508	0.746	0.074	0.458	0.426
Target Scaling	0.0557	-0.134	-0.013	0.439	0.186	0.320	0.075
GEARS	0.1412	0.105	0.082	0.584	0.253	0.128	0.569
scGPT	0.1849	-0.116	-0.077	0.437	0.498	0.081	0.347
PRESAGE	0.1749	0.184	0.148	0.624	0.117	0.104	0.698
Ridge	0.0385	0.638	0.434	0.750	0.086	0.451	0.388
Bilinear Ridge	0.0616	-0.323	-0.121	0.403	0.844	0.147	0.127
Linear Add.	0.0332	0.638	0.508	0.752	0.081	0.454	0.394
Correlation	0.0651	-0.147	-0.007	0.432	0.681	0.060	0.100
Latent Add.	0.0559	-0.144	-0.011	0.441	0.519	0.317	0.071
UNSEENBOTH ($n=213$, 5f)							
Zero	0.0529	0.000	0.000	0.000	0.500	0.056	0.064
Tech-dup	0.0010	0.991	0.991	1.000	0.000	0.938	0.927
Interp-dup	0.0012	0.991	0.990	1.000	0.000	0.975	0.970
Grand Mean	0.0552	0.020	0.002	0.441	0.500	0.295	0.082
Target Scaling	0.0552	0.062	0.006	0.441	0.214	0.295	0.083
GEARS	0.1162	0.180	0.126	0.589	0.238	0.133	0.605
scGPT	0.1532	0.013	0.013	0.513	0.595	0.040	0.478
PRESAGE	0.1289	0.259	0.214	0.647	0.119	0.167	0.676
Ridge	0.0516	0.057	0.007	0.474	0.500	0.271	0.101
Bilinear Ridge	0.0579	-0.157	-0.010	0.439	0.667	0.280	0.099
Linear Add.	0.0516	0.058	0.007	0.474	0.500	0.271	0.101
Correlation	0.0558	-0.131	-0.003	0.418	0.667	0.058	0.095
Latent Add.	0.0550	0.016	0.001	0.442	0.548	0.296	0.082
UNSEENPAIR ($n=216$, 5f)							
Zero	0.0535	0.000	0.000	0.000	0.500	0.064	0.071
Tech-dup	0.0012	0.988	0.987	1.000	0.000	0.942	0.928
Interp-dup	0.0014	0.987	0.986	1.000	0.000	0.973	0.971
MoP	0.0540	0.004	0.001	0.500	0.500	0.401	0.285
MoCT	0.0278	0.628	0.509	0.750	0.088	0.447	0.418
Two-way Mean	0.0276	0.671	0.541	0.762	0.091	0.508	0.524
GEARS	0.0370	0.658	0.400	0.755	0.189	0.558	0.481
scGPT	0.1473	-0.017	-0.012	0.495	0.500	0.032	0.495
PRESAGE	0.1654	0.172	0.122	0.600	0.195	0.128	0.602
Ridge	0.0280	0.674	0.479	0.759	0.091	0.528	0.528
Bilinear Ridge	0.0762	-0.043	-0.040	0.623	0.281	0.511	0.470
Linear Add.	0.0266	0.675	0.543	0.758	0.091	0.516	0.525
Correlation	0.0623	-0.133	-0.003	0.393	0.682	0.034	0.079
Latent Add.	0.0567	-0.144	-0.013	0.484	0.511	0.412	0.130

Table 6. McFaline23

	$\mathrm{MSE}_{\mathrm{wt}}$	$\mathrm{Pearson}_{\mathrm{wt}}$	$\mathrm{CCC}_{\mathrm{wt}}$	Frac. dir.	Cos. rank	$\mathrm{GSEA}_{\uparrow}$	$\mathrm{GSEA}_{\downarrow}$
UNSEENPERT ($n=511$, 5f)							
Zero	0.0038	0.000	0.000	0.000	0.500	—	—
Tech-dup	0.0050	0.038	0.034	1.000	0.426	—	—
Interp-dup	0.0037	0.082	0.016	1.000	0.500	—	—
MoP	0.0037	0.082	0.016	0.000	0.500	—	—
Target Scaling	0.0037	0.087	0.017	1.000	0.255	—	—
GEARS	0.0039	0.034	0.015	1.000	0.490	—	—
scGPT	0.0091	0.009	0.008	1.000	0.510	—	—
PRESAGE	0.0085	0.013	0.013	0.000	0.510	—	—
Ridge	0.0037	0.082	0.016	0.000	0.500	—	—
Bilinear Ridge	0.0037	0.113	0.022	0.000	0.049	—	—
Linear Add.	0.0037	0.082	0.016	0.000	0.500	—	—
Correlation	0.0038	0.009	0.000	1.000	0.450	—	—
Latent Add.	0.0038	0.081	0.012	0.000	0.505	—	—
UNSEENCELL ($n=511$, 3f)							
Zero	0.0033	0.000	0.000	0.000	0.500	—	—
Tech-dup	0.0044	0.029	0.026	1.000	0.414	—	—
Interp-dup	0.0033	0.043	0.010	1.000	0.500	—	—
MoCT	0.0051	0.012	0.011	1.000	0.475	—	—
Target Scaling	0.0033	0.049	0.012	1.000	0.271	—	—
GEARS	0.0035	0.022	0.010	0.000	0.514	—	—
scGPT	0.0088	0.005	0.005	1.000	0.505	—	—
PRESAGE	0.0186	0.004	0.003	1.000	0.487	—	—
Ridge	0.0038	0.020	0.010	1.000	0.471	—	—
Bilinear Ridge	0.0033	0.031	0.008	1.000	0.500	—	—
Linear Add.	0.0051	0.012	0.011	1.000	0.475	—	—
Correlation	0.0033	0.007	0.000	1.000	0.459	—	—
Latent Add.	0.0033	0.042	0.006	0.000	0.497	—	—
UNSEENBOTH ($n=511$, 5f)							
Zero	0.0034	0.000	0.000	—	0.500	—	—
Tech-dup	0.0045	0.033	0.030	—	0.401	—	—
Interp-dup	0.0034	0.044	0.011	—	0.500	—	—
Grand Mean	0.0034	0.044	0.011	—	0.500	—	—
Target Scaling	0.0034	0.048	0.013	—	0.272	—	—
GEARS	0.0036	0.023	0.010	—	0.475	—	—
scGPT	0.0094	0.003	0.003	—	0.505	—	—
PRESAGE	0.0161	0.009	0.007	—	0.485	—	—
Ridge	0.0034	0.044	0.011	—	0.500	—	—
Bilinear Ridge	0.0034	0.044	0.011	—	0.500	—	—
Linear Add.	0.0034	0.044	0.011	—	0.500	—	—
Correlation	0.0034	0.008	0.000	—	0.446	—	—
Latent Add.	0.0034	0.044	0.006	—	0.510	—	—
UNSEENPAIR ($n=511$, 5f)							
Zero	0.0038	0.000	0.000	0.000	0.500	—	—
Tech-dup	0.0051	0.039	0.036	1.000	0.402	—	—
Interp-dup	0.0037	0.085	0.015	1.000	0.500	—	—
MoP	0.0037	0.084	0.015	0.000	0.500	—	—
MoCT	0.0047	0.005	0.004	1.000	0.478	—	—
Two-way Mean	0.0046	0.019	0.011	1.000	0.467	—	—
GEARS	0.0039	0.034	0.014	1.000	0.481	—	—
scGPT	0.0086	0.011	0.010	1.000	0.549	—	—
PRESAGE	0.0161	0.008	0.007	0.000	0.505	—	—
Ridge	0.0040	0.030	0.016	1.000	0.463	—	—
Bilinear Ridge	0.0040	0.026	0.007	0.000	0.500	—	—
Linear Add.	0.0047	0.021	0.016	1.000	0.473	—	—
Correlation	0.0039	0.012	0.000	1.000	0.471	—	—
Latent Add.	0.0037	0.078	0.012	0.000	0.480	—	—

Table 7. Norman19

	MSE_{wt}	$\text{Pearson}_{\text{wt}}$	CCC_{wt}	Frac. dir.	Cos. rank	$\text{GSEA}_{\uparrow}$	$\text{GSEA}_{\downarrow}$
UNSEENCOMBO ($n=125$, 2f)							
Zero	0.0993	0.000	0.000	0.000	0.500	0.098	0.090
Tech-dup	0.0025	0.986	0.984	1.000	0.000	0.951	0.943
Interp-dup	0.0036	0.978	0.976	1.000	0.000	0.980	0.966
Additive	0.0090	0.960	0.941	0.997	0.000	0.936	0.923
Matching Mean	0.0251	0.955	0.754	0.990	0.000	0.918	0.910
Target Scaling	0.0708	0.576	0.219	0.703	0.000	0.588	0.394
Global Epistasis	0.0105	0.956	0.909	0.983	0.000	0.914	0.899
GEARS	0.0334	0.751	0.622	0.831	0.141	0.608	0.620
scGPT	0.0735	0.581	0.299	0.635	0.024	0.472	0.472
PRESAGE	0.0118	0.949	0.918	0.993	0.000	0.918	0.913
Ridge	0.0764	0.529	0.170	0.684	0.500	0.560	0.394
Bilinear Ridge	0.0623	0.780	0.339	0.811	0.000	0.737	0.572
Linear Add.	0.0764	0.529	0.170	0.684	0.500	0.560	0.394
Correlation	0.1153	-0.484	-0.072	0.194	0.794	0.000	0.022
Latent Add.	0.0915	0.528	0.059	0.666	0.461	0.557	0.373

Table 8. Replogle20

	MSE_{wt}	$\text{Pearson}_{\text{wt}}$	CCC_{wt}	Frac. dir.	Cos. rank	$\text{GSEA}_{\uparrow}$	$\text{GSEA}_{\downarrow}$
UNSEENCOMBO ($n=39$, 2f)							
Zero	10.4635	0.000	0.000	0.000	0.500	0.306	0.206
Tech-dup	0.0045	1.000	1.000	0.979	0.000	0.951	0.923
Interp-dup	0.0059	1.000	1.000	0.979	0.000	0.966	0.954
Additive	0.0751	0.996	0.993	0.957	0.000	0.899	0.873
Matching Mean	2.7901	0.996	0.727	0.963	0.000	0.893	0.872
Target Scaling	9.7519	0.601	0.045	0.792	0.000	0.778	0.654
Global Epistasis	0.3703	0.996	0.968	0.904	0.000	0.924	0.860
GEARS	4.6582	0.957	0.558	0.953	0.000	0.906	0.689
scGPT	0.0115	0.312	0.169	0.727	0.366	0.468	0.539
PRESAGE	0.4949	0.987	0.958	0.990	0.000	0.877	0.932
Ridge	9.7584	0.620	0.042	0.792	0.500	0.778	0.636
Bilinear Ridge	7.3811	0.980	0.244	0.833	0.000	0.885	0.738
Linear Add.	9.7584	0.620	0.042	0.792	0.500	0.778	0.636
Correlation	9.6266	0.341	0.018	0.750	0.197	0.220	0.487
Latent Add.	10.2452	0.623	0.014	0.792	0.500	0.771	0.629

Table 9. Replogle22

	MSE_{wt}	$\text{Pearson}_{\text{wt}}$	CCC_{wt}	Frac. dir.	Cos. rank	$\text{GSEA}_{\uparrow}$	$\text{GSEA}_{\downarrow}$
UNSEENPERT ($n=1467$, 5f)							
Zero	0.0315	0.000	0.000	0.000	0.500	0.165	0.164
Tech-dup	0.0114	0.719	0.695	1.000	0.010	0.942	0.950
Interp-dup	0.0169	0.632	0.450	1.000	0.005	0.987	0.981
MoP	0.0330	0.149	0.039	0.738	0.500	0.349	0.542
Target Scaling	0.0287	0.358	0.154	0.832	0.008	0.349	0.562
GEARS	0.0337	0.060	0.033	0.621	0.484	0.184	0.479
scGPT	0.0639	0.025	0.023	0.600	0.488	0.248	0.330
PRESAGE	0.0546	0.102	0.095	0.818	0.200	0.377	0.598
Ridge	0.0330	0.149	0.039	0.738	0.500	0.350	0.542
Bilinear Ridge	0.0300	0.097	0.043	0.711	0.078	0.385	0.450
Linear Add.	0.0330	0.149	0.039	0.738	0.500	0.349	0.542
Correlation	0.0364	0.124	0.036	0.721	0.458	0.274	0.175
Latent Add.	0.0330	0.143	0.036	0.734	0.497	0.343	0.535
UNSEENBOTH ($n=1066$, 5f)							
Zero	0.0599	0.000	0.000	0.000	0.500	0.210	0.159
Tech-dup	0.0146	0.781	0.761	1.000	0.010	0.959	0.936
Interp-dup	0.0217	0.548	0.304	1.000	0.071	0.992	0.986
Grand Mean	0.0577	0.242	0.036	0.607	0.500	0.195	0.489
Target Scaling	0.0485	0.321	0.058	0.679	0.144	0.195	0.500
GEARS	0.0544	0.180	0.071	0.667	0.441	0.148	0.584
scGPT	0.0973	0.037	0.030	0.558	0.494	0.270	0.398
PRESAGE	0.1869	0.016	0.014	0.645	0.294	0.281	0.458
Ridge	0.0577	0.242	0.036	0.607	0.500	0.195	0.489
Bilinear Ridge	0.0577	0.242	0.036	0.607	0.500	0.195	0.489
Linear Add.	0.0577	0.242	0.036	0.607	0.500	0.195	0.489
Correlation	0.0599	0.107	0.016	0.636	0.455	0.325	0.148
Latent Add.	0.0579	0.242	0.033	0.602	0.490	0.193	0.485

Table 10. Sunshine23

	MSE_{wt}	$\text{Pearson}_{\text{wt}}$	CCC_{wt}	Frac. dir.	Cos. rank	$\text{GSEA}_{\uparrow}$	$\text{GSEA}_{\downarrow}$
UNSEENCOMBO ($n=13$, 2f)							
Zero	0.0564	0.000	0.000	0.000	0.500	0.204	0.130
Tech-dup	0.0234	0.679	0.660	1.000	0.096	0.996	0.992
Interp-dup	0.0309	0.475	0.265	1.000	0.221	0.999	0.999
Additive	0.0299	0.636	0.551	1.000	0.110	0.995	0.977
Matching Mean	0.0290	0.624	0.384	1.000	0.096	0.995	0.974
Target Scaling	0.0537	0.292	0.057	1.000	0.127	0.748	0.434
Global Epistasis	0.0344	0.517	0.294	1.000	0.160	0.990	0.930
GEARS	0.0585	0.015	0.023	0.166	0.615	0.721	0.152
scGPT	0.0495	0.223	0.074	0.917	0.150	0.725	0.186
PRESAGE	0.0326	0.496	0.291	1.000	0.050	0.995	0.992
Ridge	0.0537	0.272	0.022	0.763	0.500	0.748	0.424
Bilinear Ridge	0.0485	0.653	0.105	0.962	0.000	0.990	0.832
Linear Add.	0.0537	0.272	0.022	0.763	0.500	0.748	0.424
Correlation	0.0639	0.100	0.003	1.000	0.425	0.447	0.744
Latent Add.	0.0555	0.276	0.007	0.769	0.517	0.736	0.412

Table 11. Wessels23

	MSE_{wt}	$\text{Pearson}_{\text{wt}}$	CCC_{wt}	Frac. dir.	Cos. rank	$\text{GSEA}_{\uparrow}$	$\text{GSEA}_{\downarrow}$
UNSEENCOMBO ($n=152$, 2f)							
Zero	0.2153	0.000	0.000	0.000	0.500	0.133	0.138
Tech-dup	0.0082	0.983	0.981	1.000	0.000	0.976	0.899
Interp-dup	0.0144	0.972	0.967	1.000	0.000	0.994	0.948
Additive	0.1463	0.373	0.297	0.886	0.032	0.802	0.815
Matching Mean	0.1719	0.342	0.169	0.875	0.051	0.798	0.808
Target Scaling	0.2003	0.222	0.067	0.736	0.058	0.655	0.254
Global Epistasis	0.1798	0.282	0.118	0.699	0.013	0.716	0.410
GEARS	0.1564	0.313	0.158	0.883	0.135	0.733	0.594
scGPT	0.0515	0.426	0.202	0.650	0.400	0.831	0.167
PRESAGE	0.1511	0.381	0.201	0.902	0.000	0.805	0.795
Ridge	0.2003	0.222	0.067	0.736	0.500	0.655	0.254
Bilinear Ridge	0.1955	0.251	0.079	0.774	0.000	0.689	0.293
Linear Add.	0.2003	0.222	0.067	0.736	0.500	0.655	0.254
Correlation	0.2149	0.141	0.036	0.706	0.391	0.420	0.368
Latent Add.	0.2026	0.216	0.025	0.643	0.538	0.663	0.240

D. Deep Learning models' win rates on hard perturbations

Table 12 reports the fraction of baseline-resistant perturbations (Baseline Saturation < 0.33) where each DL model outperforms the scenario-matched baseline. Perturbations with degenerate control gap ($|\text{pos} - \text{neg}| < 0.05$) are excluded.

Table 12. DL win rates on baseline-resistant perturbations (Baseline Saturation < 0.33), by scenario and metric. Win% = fraction with positive captured fraction; Med. = median captured fraction.

	PRESAGE		GEARS		scGPT	
Metric	Win%	Med.	Win%	Med.	Win%	Med.
UNSEENPERT *(Adamson16, Frangieh21, Jiang24, McFaline23, Replogle22)*						
$\text{Pearson}_{\text{wt}}$	80%	+0.30	68%	+0.15	68%	+0.21
CCC_{wt}	77%	+0.17	53%	+0.01	56%	+0.05
Frac. dir.	72%	+0.25	71%	+0.21	75%	+0.28
Cos. rank	85%	+0.67	55%	+0.14	46%	−0.07
$\text{GSEA}_{\uparrow}$	75%	+0.15	60%	+0.03	50%	+0.00
$\text{GSEA}_{\downarrow}$	80%	+0.16	69%	+0.11	78%	+0.26
MSE_{wt}	69%	+0.11	46%	−0.01	17%	−0.44
UNSEENPAIR *(Jiang24, McFaline23)*						
$\text{Pearson}_{\text{wt}}$	53%	+0.04	71%	+0.21	50%	+0.00
CCC_{wt}	34%	−0.13	62%	+0.09	33%	−0.14
Frac. dir.	85%	+0.19	85%	+0.22	80%	+0.17
Cos. rank	49%	+0.00	61%	+0.29	37%	−0.21
$\text{GSEA}_{\uparrow}$	5%	−0.20	66%	+0.13	5%	−0.21
$\text{GSEA}_{\downarrow}$	91%	+0.37	61%	+0.06	91%	+0.31
MSE_{wt}	0%	−1.92	46%	−0.01	0%	−1.62
UNSEENCOMBO *(Norman19, Wessels23, Sunshine23)*						
$\text{Pearson}_{\text{wt}}$	43%	−0.01	42%	−0.02	100%	+1.51
CCC_{wt}	22%	−0.03	24%	−0.04	80%	+2.00
Frac. dir.	43%	+0.00	76%	+1.00	—	—
Cos. rank	50%	+0.32	64%	+0.30	80%	+2.00
MSE_{wt}	65%	+0.01	49%	−0.00	0%	−0.10

E. Aggregate-vs-hard reversal analysis

Of 165 testable (dataset, scenario, metric, model) combinations, 49 (29.7%) exhibit a reversal: the DL model loses in aggregate but wins on baseline-resistant perturbations (Baseline Saturation < 0.33). The reversal rate is highest for correlation metrics ($\text{Pearson}_{\text{wt}}$: 43%, CCC_{wt}: 34%) and lowest for MSE_{wt} (15%).

F. Methods

F.1. Operation space

All predictors - models and baselines - operate in terms of delta between unperturbed control expression and perturbed expression. The expression values are present as a single pseudobulk vector per cell type per pertubation.

Pseudobulk. The datasets are pre-processed to have groundtruth data present as pseudobulk. By averaging expression across single cell expression present in the dataset.

Delta. Deltas are computed between pseudobulk vectors from the control population of unperturbed cells and population of pertubred cells.

F.2. Baselines

Let $\mathbf{\Delta} \in \mathbb{R}^{C\times P\times G}$ denote the tensor of perturbation effects, where C is the number of cell types, P the number of perturbations, and G the number of genes. Entry $\Delta_{c,p,g}$ is the log-fold-change of gene g in cell type c under perturbation p relative to the unperturbed control:

$$\Delta_{c,p,g} \;=\; \bar{x}^{\text{pert}}_{c,p,g} \;-\; \bar{x}^{\text{ctrl}}_{c,g},$$

where $\bar{x}$ denotes pseudobulk (cell-averaged) log-normalized expression. We denote the set of training (cell type, perturbation) pairs by $\mathcal{T}$, and write $\mathcal{C}_{\text{train}}$ and $\mathcal{P}_{\text{train}}$ for the sets of training cell-type and perturbation indices, respectively.

F.3. Controls

Controls bracket the performance scale: the *negative control* defines the floor (no biological signal), and *positive controls* define the ceiling (best achievable given finite sampling noise).

Negative control: Zero. Predicts no perturbation effect whatsoever:

$$\hat{\Delta}^{\text{Zero}}_{c,p,g} \;=\; 0.$$

Positive control: Technical duplicate (Tech-dup). The N_{cell} cells available for each (cell type, perturbation) condition are randomly split into two disjoint halves A and B of size $\lfloor N_{\text{cell}}/2 \rfloor$. Pseudobulk deltas are computed independently from each half:

$$\Delta^{A}_{c,p,g} = \bar{x}^{A,\text{ko}}_{c,p,g} - \bar{x}^{A,\text{ctrl}}_{c,g}, \qquad \Delta^{B}_{c,p,g} = \bar{x}^{B,\text{ko}}_{c,p,g} - \bar{x}^{B,\text{ctrl}}_{c,g}.$$

Half A serves as the ground-truth target and half B as the prediction. Because both halves measure the same biological condition, the only discrepancy arises from finite cell-sampling noise. Tech-dup therefore estimates the performance ceiling imposed by stochastic cell-to-cell variability.

Positive control: Interpolated duplicate (Interp-dup). A smoother alternative that combines the Tech-dup signal on differentially expressed genes (DEGs) with the Mean-over-Perturbations baseline elsewhere:

$$\hat{\Delta}^{\text{Interp}}_{c,p,g} \;=\; m_{p,g}\,\Delta^{B}_{c,p,g} \;+\; (1-m_{p,g})\,\hat{\Delta}^{\text{MoP}}_{c,p,g},$$

where $m_{p,g} \in \{0,1\}$ is a binary DEG mask (1 if gene g is differentially expressed under perturbation p, determined by t-test with Benjamini–Hochberg correction or magnitude thresholding). This control is less noisy than Tech-dup on the majority of genes with negligible perturbation effects, while preserving the biological signal on DEGs.

F.4. Non-learned baselines

Each evaluation regime has a corresponding MSE-optimal simple-rule baseline, i.e., the conditional expectation $\mathbb{E}[\Delta \mid$ info available at test time$]$. Table 13 summarizes the correspondence.

Mean-over-Perturbations (MoP). Averages over all training perturbations within each cell type:

$$\hat{\Delta}^{\text{MoP}}_{c,p,g} \;=\; \frac{1}{|\mathcal{P}_{\text{train}}|}\sum_{k'\in\mathcal{P}_{\text{train}}} \Delta_{c,p',g}.$$

Table 13. MSE-optimal non-learned baseline for each scenario.

Evaluation regime	Known at test time	Optimal predictor	Baseline
UNSEENPERT	Cell type only	$\mathbb{E}[\Delta \mid c]$	Mean-over-Perturbations
UNSEENCELL	Perturbation only	$\mathbb{E}[\Delta \mid p]$	Mean-over-Cell-Types
UNSEENBOTH	Neither	$\mathbb{E}[\Delta]$	Grand Mean
UNSEENPAIR	Both (unseen combination)	$\mathbb{E}[\Delta \mid c] + \mathbb{E}[\Delta \mid p] - \mathbb{E}[\Delta]$	Two-way Mean
UNSEENDOSE	Full-perturbation effect + dose	$d \cdot \Delta^{\text{full}}$	Scaled Delta
UNSEENCOMBO	Single-perturbation effects	$\Delta_A + \Delta_B$	Additive

For unseen cell types ($c \notin \mathcal{C}_{\text{train}}$), falls back to the global mean across all training cell types and perturbations.

Mean-over-Cell-Types (MoCT). Averages over all training cell types for each perturbation:

$$\hat{\Delta}^{\text{MoCT}}_{c,p,g} \;=\; \frac{1}{|\mathcal{C}_{\text{train}}|} \sum_{b' \in \mathcal{C}_{\text{train}}} \Delta_{c',p,g}.$$

Requires that each test perturbation was observed during training (otherwise undefined).

Grand Mean. A single prediction vector for all (cell type, perturbation) pairs:

$$\hat{\Delta}^{\text{GM}}_{c,p,g} \;=\; \frac{1}{|\mathcal{C}_{\text{train}}|\;|\mathcal{P}_{\text{train}}|} \sum_{b' \in \mathcal{C}_{\text{train}}} \sum_{k' \in \mathcal{P}_{\text{train}}} \Delta_{c',p',g}.$$

Two-way Mean. An additive row–column decomposition of the training tensor:

$$\hat{\Delta}^{\text{2W}}_{c,p,g} \;=\; \mu_g + \alpha_{c,g} + \beta_{p,g},$$

where μ_g is the grand mean, $\alpha_{c,g}$ is the cell-type effect, and $\beta_{p,g}$ is the perturbation effect:

$$\begin{aligned}
\mu_g &= \frac{1}{|\mathcal{T}|} \sum_{(b',k') \in \mathcal{T}} \Delta_{c',p',g}, \\
\alpha_{c,g} &= \frac{1}{|\{p' : (c,p') \in \mathcal{T}\}|} \sum_{\{p' : (c,p') \in \mathcal{T}\}} \Delta_{c,p',g} \;-\; \mu_g, \\
\beta_{p,g} &= \frac{1}{|\{c' : (c',p) \in \mathcal{T}\}|} \sum_{\{c' : (c',p) \in \mathcal{T}\}} \Delta_{c',p,g} \;-\; \mu_g.
\end{aligned} \tag{2}$$

This is equivalent to MoP + MoCT − Grand Mean and is the MSE-optimal predictor when both cell type and perturbation identity are known but their specific combination was not observed during training (S4).

Additive (for combinatorial perturbations). For a combinatorial perturbation (A, B) whose constituent singles are observed:

$$\hat{\Delta}^{\text{Add}}_{c,(A,B),g} \;=\; \Delta_{c,A,g} + \Delta_{c,B,g}.$$

Assumes perturbation effects compose without epistasis, i.e., the joint effect equals the sum of marginal effects (S6).

Matching Mean (for combinatorial perturbations). A conservative alternative to Additive that averages rather than sums:

$$\hat{\Delta}^{\text{MM}}_{c,(A,B),g} \;=\; \frac{\Delta_{c,A,g} + \Delta_{c,B,g}}{2}.$$

Scaled Delta (for partial perturbations). For a partial knockdown at dose $d \in (0, 1)$ of a gene whose full perturbation effect is known:

$$\hat{\Delta}^{\text{SD}}_{c,p,g} \;=\; d \cdot \Delta^{\text{full}}_{c,p,g}.$$

Assumes linear dose–response (S5).

F.5. Learned baselines

Learned baselines use the training split to fit parameters, while remaining far simpler than full neural perturbation-prediction models. All use the same multi-hot feature encoding: $\mathbf{x} = [\mathbf{e}_p; \mathbf{e}_c] \in \{0,1\}^{P+C}$, where $\mathbf{e}_p$ and $\mathbf{e}_c$ are one-hot indicators for the perturbation and cell type, respectively.

Linear Additive. A linear model on the multi-hot encoding, trained by mini-batch SGD on MSE loss:

$$\hat{\Delta}_g \;=\; \mathbf{w}_g^\top \mathbf{x} + c_g,$$

or equivalently $\hat{\mathbf{\Delta}} = \mathbf{W}^\top \mathbf{x} + \mathbf{c}$, with $\mathbf{W} \in \mathbb{R}^{(P+C)\times G}$ and $\mathbf{c} \in \mathbb{R}^G$.

Ridge. The same multi-hot feature encoding but solved in closed form with ℓ_2 regularization:

$$\mathbf{W} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{Y}_c,$$

where $\mathbf{X} \in \{0,1\}^{N\times(P+C)}$ is the design matrix over all $N = |\mathcal{T}|$ training samples, $\mathbf{Y}_c$ is the mean-centered target matrix, and λ controls regularization strength.

Correlation. A parameter-free baseline that predicts perturbation effects from gene–gene co-expression:

$$\hat{\Delta}_{c,p,g} \;=\; -\,\rho(g_p,\, g)\;\cdot\; x^{\text{ctrl}}_{c,g_p},$$

where $\rho(g_p, g)$ is the Pearson correlation between the perturbed gene g_p and target gene g, estimated from the flattened training delta matrix, and x^{ctrl}_{c,g_p} is the control expression level of the perturbed gene in cell type c. For combinatorial perturbations (g_A, g_B), contributions are summed: $\hat{\Delta}_{c,(A,B),g} = -\rho(g_A, g)\, x^{\text{ctrl}}_{c,g_A} - \rho(g_B, g)\, x^{\text{ctrl}}_{c,g_B}$.

Latent Additive. Three two-layer MLPs (ReLU activations, hidden dimension h, latent dimension d) that enforce additivity in a learned latent space:

$$\mathbf{z} = f_{\text{gene}}(\mathbf{x}^{\text{ctrl}}) + f_{\text{pert}}(\mathbf{x}), \qquad \hat{\mathbf{\Delta}} = f_{\text{dec}}(\mathbf{z}),$$

where $f_{\text{gene}} : \mathbb{R}^G \to \mathbb{R}^d$ encodes the control expression profile, $f_{\text{pert}} : \{0,1\}^{P+C} \to \mathbb{R}^d$ encodes perturbation identity, and $f_{\text{dec}} : \mathbb{R}^d \to \mathbb{R}^G$ decodes the predicted delta. By summing in latent space, the model can learn nonlinear gene-expression features while preserving an additive perturbation structure. Trained end-to-end via mini-batch SGD on MSE loss ($h = 256$, $d = 128$).

F.6. Performance metrics - full list

We evaluate every predictor against ground truth using a catalog of metrics drawn from the perturbation-prediction literature. Throughout this appendix we fix notation as follows. Let $k = (c, p)$ index a test item (cell type c, perturbation p), let G denote the number of genes, and let $g \in \{1, \ldots, G\}$ index a gene. We write $\hat{\Delta}_{k,g}$ for the predicted delta and $\Delta_{k,g}$ for the ground-truth delta (log-fold-change with respect to the unperturbed control). Each metric below is defined as a per-row scalar $M(\hat{\mathbf{\Delta}}_k, \mathbf{\Delta}_k)$ acting on a single test item (or, for cross-row metrics, on the full $K \times G$ tensor). Where relevant, $w_g \geq 0$ denotes a per-gene weight and $\mathbf{1}[\cdot]$ an indicator. A gene mask $m_{k,g} \in \{0,1\}$ may restrict the sum to a subset of genes; unless stated otherwise, unmasked genes contribute uniformly ($w_g = 1$, $m_{k,g} = 1$ for all g). All modifiers (masking, weighting, stratification) are deferred to Section G, and every base metric below may be combined with any modifier listed there.

F.6.1. Point-wise distribution metrics

Element-wise comparisons between predicted and ground-truth deltas.

Mean Squared Error (MSE). Lower is better; perfect value 0.

$$\text{MSE}_k \;=\; \frac{\sum_g w_g\, m_{k,g} \left(\hat{\Delta}_{k,g} - \Delta_{k,g}\right)^2}{\sum_g w_g\, m_{k,g}}.$$

Mean Absolute Error (MAE). Lower is better; perfect value 0.

$$\mathrm{MAE}_k \;=\; \frac{\sum_g w_g\, m_{k,g}\,\big|\hat{\Delta}_{k,g} - \Delta_{k,g}\big|}{\sum_g w_g\, m_{k,g}}.$$

Fold-change Gap. Mean absolute error averaged across four equally populated bins of $|\Delta_{k,g}|$ (quartiles of the ground-truth magnitude distribution). Let $B_0, \ldots, B_3$ denote the quartile index sets of $|\Delta_{k,\cdot}|$.

$$\mathrm{FCG}_k \;=\; \frac{1}{4}\sum_{b=0}^{3} \frac{1}{|B_b|} \sum_{g \in B_b} \big|\hat{\Delta}_{k,g} - \Delta_{k,g}\big|.$$

This reweights error so that the high-magnitude tail (a few strongly affected genes) is not drowned by the bulk of unaffected genes. Lower is better; perfect value 0.

F.6.2. Correlation metrics

Shape- or rank-alignment between prediction and target, ignoring absolute scale. Let weighted moments be $\bar{a}_k = \sum_g w_g m_{k,g}\, a_{k,g} \big/ \sum_g w_g m_{k,g}$ for any quantity a.

Pearson correlation. Higher is better; perfect value 1.

$$r_k \;=\; \frac{\sum_g w_g m_{k,g}\,\big(\hat{\Delta}_{k,g} - \bar{\hat{\Delta}}_k\big)\big(\Delta_{k,g} - \bar{\Delta}_k\big)}{\sqrt{\sum_g w_g m_{k,g}\big(\hat{\Delta}_{k,g} - \bar{\hat{\Delta}}_k\big)^2 \;\cdot\; \sum_g w_g m_{k,g}\big(\Delta_{k,g} - \bar{\Delta}_k\big)^2} + \varepsilon}.$$

Spearman correlation. Pearson correlation applied to midrank-transformed values; ties receive the average of the positions they would occupy in the sorted list. Let $\mathrm{rank}(\cdot)$ denote the per-row midrank operator.

$$\rho_k \;=\; r_k\big(\mathrm{rank}(\hat{\boldsymbol{\Delta}}_k),\; \mathrm{rank}(\boldsymbol{\Delta}_k)\big).$$

Rows with constant target or prediction return 0. Higher is better; perfect value 1.

Uncentered R^2. The coefficient of determination computed with the *uncentered* total sum of squares $\sum_g w_g m_{k,g}\, \Delta_{k,g}^2$ as denominator (i.e. taking zero, not $\bar{\Delta}_k$, as the null prediction). This aligns $R^2 = 0$ with the Zero negative control, rather than with the Mean-over-Genes predictor.

$$R^2_k \;=\; 1 \;-\; \frac{\sum_g w_g m_{k,g}\big(\Delta_{k,g} - \hat{\Delta}_{k,g}\big)^2}{\sum_g w_g m_{k,g}\,\Delta_{k,g}^2 + \varepsilon}.$$

Higher is better; perfect value 1; unbounded below.

Centered R^2. The textbook coefficient of determination, with the target mean as null:

$$R^2_{\mathrm{c},\,k} \;=\; 1 \;-\; \frac{\sum_g w_g m_{k,g}\big(\Delta_{k,g} - \hat{\Delta}_{k,g}\big)^2}{\sum_g w_g m_{k,g}\big(\Delta_{k,g} - \bar{\Delta}_k\big)^2 + \varepsilon}.$$

$R^2_{\mathrm{c}} = 0$ corresponds to predicting the per-row mean. Higher is better; perfect value 1; unbounded below.

Concordance Correlation Coefficient (CCC). Lin's concordance coefficient penalises both scale and location mismatch:

$$\mathrm{CCC}_k \;=\; \frac{2\,\mathrm{cov}_w(\hat{\Delta}_k, \Delta_k)}{\mathrm{var}_w(\hat{\Delta}_k) + \mathrm{var}_w(\Delta_k) + \big(\bar{\hat{\Delta}}_k - \bar{\Delta}_k\big)^2 + \varepsilon},$$

where cov_w and var_w denote the weighted covariance and variance under $w_g m_{k,g}$. Higher is better; perfect value 1.

F.6.3. DISCRIMINATION METRICS

Whether a prediction's vector retains enough identity to distinguish its own perturbation from others (PDS) or to flag which genes respond to the perturbation (effect-size AUROC).

Perturbation Discrimination Score (PDS). For each predicted profile $\hat{\boldsymbol{\Delta}}_k$ we find the nearest ground-truth profile $\boldsymbol{\Delta}_j$ under a chosen distance d and ask whether $j = k$:

$$\mathrm{PDS} \;=\; \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\Big[k = \arg\min_{j}\; d\big(\hat{\boldsymbol{\Delta}}_k, \boldsymbol{\Delta}_j\big)\Big].$$

We evaluate three distances (yielding `pds_cosine`, `pds_l2`, `pds_l1`):

$$\begin{aligned} d_{\cos}(u,v) \;&=\; 1 - \frac{u^\top v}{\|u\|_2\,\|v\|_2 + \varepsilon}, \\ d_{L_2}(u,v) \;&=\; \|u - v\|_2, \\ d_{L_1}(u,v) \;&=\; \|u - v\|_1. \end{aligned}$$

If $K > K_{\max}$ ($K_{\max} = 200$ by default), a random subsample of $K_{\max}$ items is drawn with a fixed seed. Higher is better; perfect value 1.

Effect-size AUROC. Measures the separability of responding vs non-responding genes under the prediction. For each row, the binary label is $y_{k,g} = \mathbf{1}[\,|\Delta_{k,g}| > \theta\,]$ and the score is $s_{k,g} = |\hat{\Delta}_{k,g}|$ (threshold $\theta = 0.5$ by default). Using midrank-based Mann–Whitney U with $n_k^+ = \sum_g y_{k,g}$ and $n_k^- = G - n_k^+$:

$$\mathrm{AUROC}_k \;=\; \frac{U_k}{n_k^+\, n_k^- + \varepsilon}, \qquad U_k \;=\; \sum_{g} \mathrm{rank}(s_{k,\cdot})_g\, y_{k,g} \;-\; \tfrac{1}{2} n_k^+\big(n_k^+ - 1\big).$$

Rows with $n_k^+ = 0$ or $n_k^- = 0$ return 0.5. Higher is better; perfect value 1.

F.6.4. EFFECT-SIZE METRICS

Recovery of the *shape* of the per-perturbation response—its direction and magnitude distribution—without requiring per-gene alignment.

Fraction of correct direction. Among genes with a non-zero ground-truth delta, the fraction on which the predicted sign agrees:

$$\mathrm{FCD}_k \;=\; \frac{\big|\{g : \Delta_{k,g} \neq 0,\; \mathrm{sign}(\hat{\Delta}_{k,g}) = \mathrm{sign}(\Delta_{k,g})\}\big|}{\big|\{g : \Delta_{k,g} \neq 0\}\big|}.$$

Rows with fewer than $n_{\min}$ signed genes return NaN (default $n_{\min} = 5$). Higher is better; perfect value 1.

Variance-ratio log error. A global (cross-row) diagnostic that penalises prediction variance collapse or inflation:

$$\mathrm{VRLE} \;=\; \underset{g\,:\,\mathrm{Var}_k(\Delta_{k,g}) \geq \sigma^2_{\min}}{\mathrm{median}} \Big|\log \frac{\mathrm{Var}_k(\hat{\Delta}_{k,g}) + \varepsilon}{\mathrm{Var}_k(\Delta_{k,g}) + \varepsilon}\Big|,$$

where Var_k is the variance across test items for a fixed gene and $\sigma^2_{\min} = 10^{-4}$ is a variance floor excluding genes whose ground-truth variance is numerically negligible. Lower is better; perfect value 0.

F.6.5. ENERGY DISTANCE METRICS

Cross-row distances comparing the empirical distribution of predicted profiles to that of ground-truth profiles, treating the K rows as samples from two distributions on $\mathbb{R}^G$.

E-distance. Szekely's energy distance in the full gene space:

$$\mathrm{Edist}(\hat{\mathbf{\Delta}}, \mathbf{\Delta}) \;=\; \frac{2}{K^2}\sum_{i,j}\big\|\hat{\mathbf{\Delta}}_i - \mathbf{\Delta}_j\big\|_2 \;-\; \frac{1}{K^2}\sum_{i,i'}\big\|\hat{\mathbf{\Delta}}_i - \hat{\mathbf{\Delta}}_{i'}\big\|_2 \;-\; \frac{1}{K^2}\sum_{j,j'}\big\|\mathbf{\Delta}_j - \mathbf{\Delta}_{j'}\big\|_2.$$

Zero iff the two empirical distributions coincide. Lower is better; perfect value 0.

E-distance in PCA space. Identical to EDIST but computed on a PCA projection of rank n_{PC} (default 50). If μ and $V_{n_{\mathrm{PC}}}$ are the mean and principal-component matrix fit on the ground-truth tensor, we project $\tilde{\mathbf{\Delta}}_i = (\mathbf{\Delta}_i - \mu)V_{n_{\mathrm{PC}}}^{\top}$ and similarly for $\hat{\mathbf{\Delta}}$, then

$$\mathrm{Edist}_{\mathrm{PCA}} \;=\; \mathrm{Edist}(\tilde{\hat{\mathbf{\Delta}}}, \tilde{\mathbf{\Delta}}).$$

This reduces gene-space isotropy and emphasises the directions of genuine biological variation.

F.6.6. DE RECALL METRICS

Whether the top of the predicted ranking is enriched for genes that the ground truth flags as differentially expressed (DEGs).

GSEA enrichment score. For each row k, let $S_k \subseteq \{1, \ldots, G\}$ be the signed DEG set and let $\pi_k \in \mathfrak{S}_G$ be the permutation that sorts $\{\hat{\Delta}_{k,g}\}_{g=1}^{G}$ in descending order. Walking the ranked list from rank 1 to rank G, each step contributes

$$r_{k,j} \;=\; \begin{cases} +1/|S_k| & \text{if } \pi_k(j) \in S_k, \\ -1/(G - |S_k|) & \text{otherwise,} \end{cases}$$

and the enrichment score is the maximum of the running cumulative sum:

$$\mathrm{ES}_k \;=\; \max_{1 \le j \le G} \sum_{i=1}^{j} r_{k,i}.$$

Rows with $|S_k| < n_{\text{set,min}}$ (default 5) return NaN. Higher is better; perfect value 1.

GSEA-up. Uses $S_k = \{g : m_{k,g}^{\mathrm{deg}} = 1 \text{ and } \Delta_{k,g} > 0\}$ and ranks predictions in descending order: rewards predictions that place upregulated DEGs at the *top*.

GSEA-down. Uses $S_k = \{g : m_{k,g}^{\mathrm{deg}} = 1 \text{ and } \Delta_{k,g} < 0\}$ and ranks predictions in ascending order ($\arg\mathrm{sort}(-\hat{\Delta})$ descending), rewarding predictions that place downregulated DEGs at the top. In both cases m^{deg} is the per-row DEG mask described in Section G.

F.6.7. BASE METRIC CATALOG

Table 14 summarises the eight composable base metrics and ten standalone metrics implemented in the pipeline. Composable metrics admit all modifiers in Section G; standalone metrics operate on the full $K \times G$ tensor and do not accept masks or weights.

The evaluation pipeline computes 8 composable bases $\times$ modifier choices $+10$ standalone metrics per (scenario, fold, predictor), using bio-relation graphs (GO, PPI, pathway, TF-target) as modifier sources for a total of 110 metrics.

G. Metric modifiers: DEGs, weighting, and masking

Composable base metrics (MSE, Pearson, CCC, etc.) can be combined with modifiers that restrict the gene set or reweight genes. This section describes the modifiers used in the evaluation pipeline.

G.1. Differentially expressed genes (DEGs)

DEGs are computed per perturbation from the second-half single-cell split (held out from pseudobulk ground truth) using scanpy's `rank_genes_groups` with `method='t-test_overestim_var'` and `reference='rest'` (Wolf

Table 14. Base metric catalog. Composable metrics are combined with the modifiers in Section G; standalone metrics are evaluated as single scalars over the full tensor. π denotes the metric's theoretical perfect value.

Category	Metric	Direction	π	Composable
Point-wise	MSE	lower	0	✓
Point-wise	MAE	lower	0	✓
Point-wise	Fold-change gap	lower	0	—
Correlation	Pearson	higher	1	✓
Correlation	Spearman	higher	1	✓
Correlation	R^2 (uncentered)	higher	1	✓
Correlation	R^2 (centered)	higher	1	✓
Correlation	CCC	higher	1	✓
Discrimination	PDS (cosine / L_2 / L_1)	higher	1	—
Discrimination	Effect-size AUROC	higher	1	—
Effect-size	Frac. correct direction	higher	1	✓(weighting only)
Effect-size	Variance-ratio log error	lower	0	—
Energy	E-distance	lower	0	—
Energy	E-distance (PCA)	lower	0	—
DE recall	GSEA-up	higher	1	—
DE recall	GSEA-down	higher	1	—

et al., 2018). This performs a Welch t-test with overestimated variance (conservative) comparing perturbed cells against the rest of the dataset, with Benjamini–Hochberg correction for multiple testing. A gene is classified as a DEG for perturbation k if its adjusted p-value satisfies $p_{\text{adj}} < 0.05$.

From the DEG analysis we derive three arrays stored in the HDF5 file:

DEG mask. A binary matrix $m^{\text{deg}}_{k,g} \in \{0, 1\}$ indicating whether gene g is differentially expressed under perturbation k ($p_{\text{adj}} < 0.05$). Used by GSEA metrics and fraction correct direction to restrict evaluation to genes with detectable perturbation effects.

DEG directions. A signed matrix $d^{\text{deg}}_{k,g} \in \{-1, 0, +1\}$: $+1$ if gene g is significantly upregulated (positive t-score and $p_{\text{adj}} < 0.05$), -1 if significantly downregulated, and 0 otherwise. Used by fraction correct direction to define the ground-truth sign of change.

Per-perturbation weights. For DEG-weighted metrics (denoted by the `_weighted` suffix), each gene receives a weight derived from the t-test score:

$$w_{k,g} = \left(\frac{|t_{k,g}| - \min_g |t_{k,g}|}{\max_g |t_{k,g}| - \min_g |t_{k,g}|} \right)^2,$$

where $t_{k,g}$ is the t-score for gene g under perturbation k. This min–max normalization followed by squaring concentrates weight on the most strongly affected genes while retaining contributions from moderately affected ones. When the range of $|t|$-scores is zero (constant across genes), all weights are set to zero. The weighted metric is then:

$$M_k^{\text{weighted}} = \frac{\sum_g w_{k,g}\, f(\hat{\Delta}_{k,g}, \Delta_{k,g})}{\sum_g w_{k,g}},$$

where f is the base metric's per-gene function (e.g., squared error for MSE, or the contribution to the correlation numerator for Pearson).

G.2. Variance-based weighting

For variance-weighted metrics (denoted `_var_weighted`), genes are weighted by the inverse of their cross-perturbation variance:

$$w_g^{\text{var}} = \frac{1}{1 + \text{Var}_k(\Delta_{k,g})},$$

where Var_k is computed across all test perturbations. This downweights highly variable genes (which tend to be noisy) and upweights genes with stable expression changes.

G.3. Gene masking

Gene masks restrict the metric computation to a subset of genes:

[nosep]

- **Top-k most affected genes per perturbation** (`_top_200`): for each perturbation k, retain only the $k = 200$ genes with the largest $|\Delta_{k,g}|$. This focuses evaluation on genes with the strongest ground-truth response.
- **Top-k expressed genes globally** (`_top_1000_expressed`): retain the 1,000 genes with the highest mean expression across control cells. This restricts evaluation to well-measured genes.

H. Meta-metrics: calibration and evaluation

The two meta-metrics—DRF and Baseline Saturation—are measured per perturbation and operate on per-item performance metric values. Let p index a test perturbation (or, for multi-cell-type datasets, a test item (c, p)). For a given evaluation metric, we write neg_p, pos_p, and baseline_p for the metric values of the negative control, positive control, and scenario-matched baseline on perturbation p.

H.1. DRF: Dynamic Range Fraction

DRF (Miller et al., 2025) measures what fraction of a metric's theoretical dynamic range is realized between the negative and positive controls. Let m^* denote the metric's theoretical perfect score.

For higher-is-better metrics ($m^* > \text{neg}_p$ in expectation):

$$\text{DRF}_p = \frac{\text{pos}_p - \text{neg}_p}{m^* - \text{neg}_p + \varepsilon}, \qquad \varepsilon = 10^{-6}. \tag{3}$$

For lower-is-better metrics ($m^* < \text{neg}_p$ in expectation):

$$\text{DRF}_p = \frac{\text{neg}_p - \text{pos}_p}{\text{neg}_p - m^* + \varepsilon}. \tag{4}$$

Positive-control selection. We evaluate two positive controls: Tech-duplicate (Δ^B, the split-half replicate) and Interp-duplicate (a smoothed variant that uses the Tech-duplicate signal on DEGs and MoP elsewhere; see §F.2). For each metric, we select the positive control whose median value is better (higher for higher-is-better, lower for lower-is-better). This selection is performed once per metric across all test items.

DRF variants. Following Miller et al. (2025), we compute four DRF variants by varying the (negative, positive) pair:

- **DRF** (default): negative = Zero, positive = best of {Tech-dup, Interp-dup}.
- **DRF-mean**: negative = Mean-over-Perturbations, positive = Tech-dup.
- **DRF-ctrl**: negative = Zero, positive = Interp-dup.
- **DRF-interpolated**: negative = Mean-over-Perturbations, positive = Interp-dup.

DRF values are reported unclipped per perturbation. We retain items with $\text{DRF}_p > \tau$ for downstream Baseline Saturation computation (default $\tau = 0$).

H.2. Baseline Saturation

Baseline Saturation quantifies how much of the calibrated dynamic range is captured by simple-rule baselines.

For higher-is-better metrics:

$$\text{Baseline Saturation}_p = \frac{\text{baseline}_p - \text{neg}_p}{\text{pos}_p - \text{neg}_p + \varepsilon}, \qquad \varepsilon = 10^{-8}. \tag{5}$$

Table 15. Existing datasets and model assets used in this work. We do not redistribute raw datasets, pretrained weights, or external model assets; users should obtain them from the original sources.

Asset	Type	Source / identifier	License or usage terms	Redistributed?
Adamson16	Dataset	NCBI GEO: GSE90546	GEO terms[a]	No
Frangieh21	Dataset	Broad Single Cell Portal: SCP1064; raw data via DUOS-000124	SCP / DUOS terms	No
Norman19	Dataset	NCBI GEO: GSE133344	GEO terms[a]	No
Wessels23	Dataset	NCBI GEO: GSE213957	GEO terms[a]	No
Replogle20	Dataset	NCBI GEO: GSE146194	GEO terms[a]	No
Replogle22	Dataset	Figshare+; DOI in original source	CC BY 4.0	No
McFaline23	Dataset	NCBI GEO: GSE114687	GEO terms[a]	No
Jiang24 / PerturBench	Dataset / benchmark source	PerturBench processed datasets via Hugging Face / Lamin	Original data terms; PerturBench code: BSD 3-Clause	No
scGPT	Model / code	`bowang-lab/scGPT`	MIT; checkpoint terms separate	No
GEARS	Model / code	`snap-stanford/GEARS`	MIT	No
PRESAGE	Model / code	`Genentech/PRESAGE`	Genentech Non-Commercial Software License v1.0	No

[a] *GEO terms* denotes the NCBI GEO public-access terms, subject to any submitter-retained rights. For controlled-access or portal-hosted datasets, users should follow the terms provided by the original data source.

For lower-is-better metrics:

$$\text{Baseline Saturation}_p = \frac{\text{neg}_p - \text{baseline}_p}{\text{neg}_p - \text{pos}_p + \varepsilon}. \tag{6}$$

Baseline pool. Baseline Saturation is computed only over non-learned simple-rule baselines. The pool consists of:

1. Mean-over-Perturbations (MoP)
2. Mean-over-Cell-Types (MoCT)
3. Grand Mean
4. Two-way Mean
5. Scaled Delta (partial perturbations only)
6. Additive (combinatorial perturbations only)

Learned models (Linear Additive, Ridge, Correlation, Latent Additive) are excluded.

DRF pre-filter. Only perturbations with $\text{DRF}_p > \tau_{\min}$ (configurable; default $\tau_{\min} = 0$) are included. Perturbations where the control range is degenerate are excluded to prevent unstable Baseline Saturation values.

Best-baseline selection. For each metric, Baseline Saturation is computed independently for every baseline in the pool. The best baseline is the one maximizing $\text{mean}(\text{Baseline Saturation}_p)$ across all DRF-passing perturbations. This selection is global per metric (not per perturbation).

Aggregation. Per-perturbation Baseline Saturation values are clipped to $[0, 1]$ and aggregated as:

$$\text{Baseline Saturation} = \text{median}\big(\text{clip}(\text{Baseline Saturation}_p, 0, 1)\big).$$

I. Existing Assets and Licenses

J. Compute resources

Experiments were run on an internal GPU cluster using up to 8 NVIDIA A100 GPUs with 80GB memory each. CPU preprocessing, pseudobulk construction, baseline evaluation, metric computation, and figure generation were run on standard cluster CPU nodes with 32–64 CPU cores and 256–512GB RAM. Across datasets, raw inputs, intermediate pseudobulk files, model outputs, and generated figures required approximately 1–2TB of storage.