nature biotechnology

Supplementary information

https://doi.org/10.1038/s41587-026-03307-w

Deep learning perturbation models can outperform baselines on calibrated metrics

In the format provided by the authors and unedited

Supplementary Note #1: Defining positive and negative controls

To better understand why the technical duplicate underperforms the mean baseline in some cases (Extended Data Fig. 1), we compared MSE(technical duplicate) and MSE(mean baseline) for each perturbation in Replogle22 K562 GWPS. We found that the technical duplicate underperformed the mean baseline for approximately 95% of perturbations (Extended Data Fig. 1E). Examining representative cases of high (SUPT6H, 435 DEGs), moderate (OGFOD1, 31 DEGs), and low (COX5A, 5 DEGs) perturbation strength showed that genes strongly-affected by a perturbation (adjusted $p_{\text{DEG}} < 0.05$) were predicted more accurately by the technical duplicate, whereas unaffected gene expression ($p_{\text{DEG}} > 0.95$) was predicted more accurately by the mean baseline (fig. SN1). Across all perturbations, technical duplicate performance improved with increasing perturbation strength (Extended Data Fig. 1E).

A downsampling titration experiment further revealed strong dependence on sampling power: increasing the number of cells used to compute the technical duplicate markedly improved its performance relative to the mean baseline (fig. SN2). Together, these findings suggest a straightforward explanation for poor technical duplicate performance: for strongly affected genes, the technical duplicate captures the true signal despite limited sampling, whereas for unaffected genes, the mean baseline provides a more stable estimate because it averages over thousands of cells to calculate unaffected gene expression (fig. SN3).

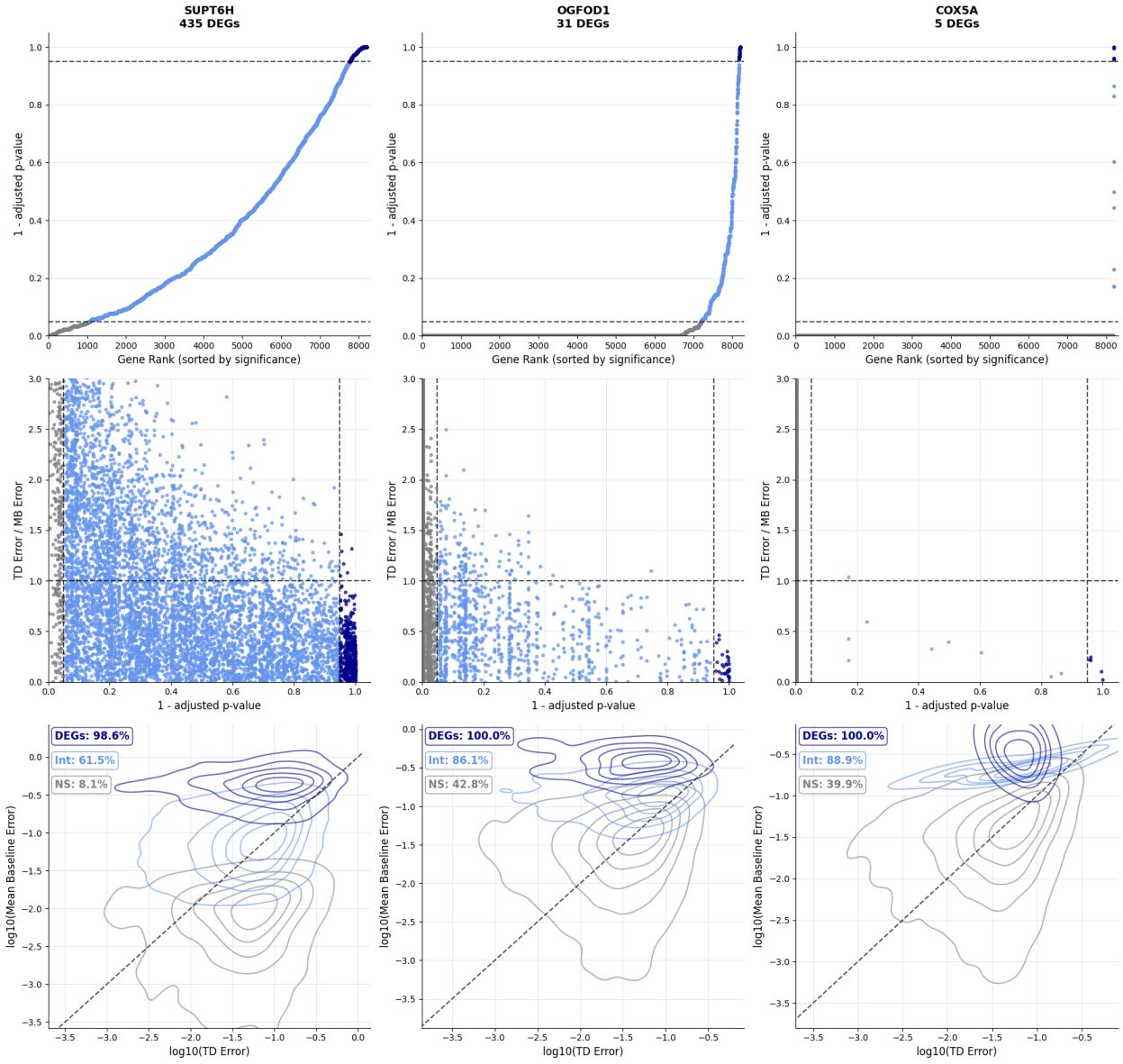

Figure SN1: Relationship between differential gene expression and technical duplicate performance in the Replogle22 K562 GWPS dataset. Top row: rank plots showing the distribution of genome-wide adjusted p-values from differential expression analysis for three representative perturbations (SUPT6H, OGFOD1, and COX5A). Middle row: scatter plots relating gene-wise adjusted p-values to the ratio of technical-duplicate (TD) error (vs. ground truth) to mean-baseline (MB) error. Bottom row: contour plots comparing technical-duplicate and mean-baseline errors, with genes grouped as affected (adjusted $p < 0.05$), unaffected ($p > 0.95$), or intermediate ($0.05 < p < 0.95$). Percentages denote the fraction of genes where the technical duplicate outperforms the mean baseline as a predictor.

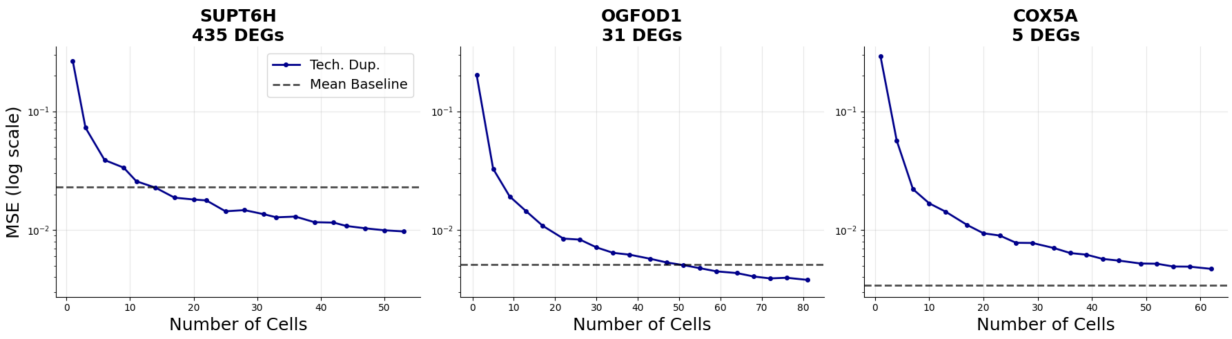

Figure SN2: Effect of sample size on technical-duplicate performance in the Replogle22 K562 GWPS dataset. Line plots show the relationship between the number of cells sampled per perturbation and mean squared error (MSE) against ground truth for three representative perturbations. The mean-baseline performance is shown in each case as a reference.

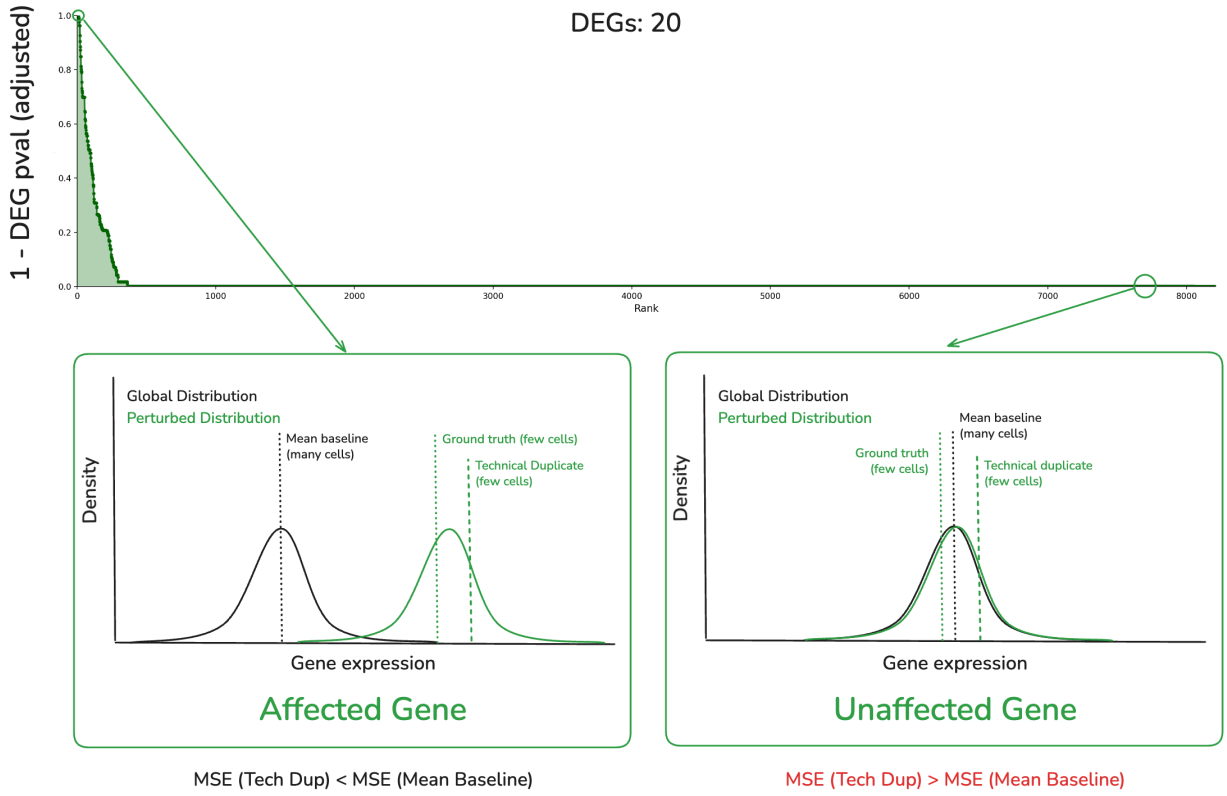

Figure SN3: Conceptual illustration of factors influencing technical-duplicate performance relative to the mean baseline. For genes strongly affected by perturbation, the true expression distribution shifts substantially, allowing the technical duplicate to predict the ground truth more accurately despite limited cell numbers. For weakly or unaffected genes, the mean baseline provides a more stable estimate of unperturbed expression, as the sampling noise in the technical duplicate exceeds the true perturbation effect, leading to poorer predictive performance.

Supplementary Note #2: Reliability of calibration approach

To evaluate the reliability of our calibration approach, we analyzed dataset factors to understand their impact on metric calibration and overall baseline performance: (i) the impact of changing the positive and negative controls for DRF calculation (fig. SN4), (ii) the impact of number of cells per perturbation (fig. SN5), (iii) the impact of the library size per cell (fig. SN6), (iv) number of top highly variable genes (HVGs) selected (fig. SN7), and (v) the relationship between perturbation strength and baseline performance on well-calibrated metrics (fig. SN8). Overall, we found that, compared to typical benchmarking metrics like MSE and Pearson Δ_{ctrl}, well-calibrated metrics amplify perturbation-specific signal and maintain higher calibration, even when data quality is lower (i.e., lower library size or cell count), perturbations are weaker, or selected genes have lower variance on average.

0.1 Choice of controls for DRF calculation

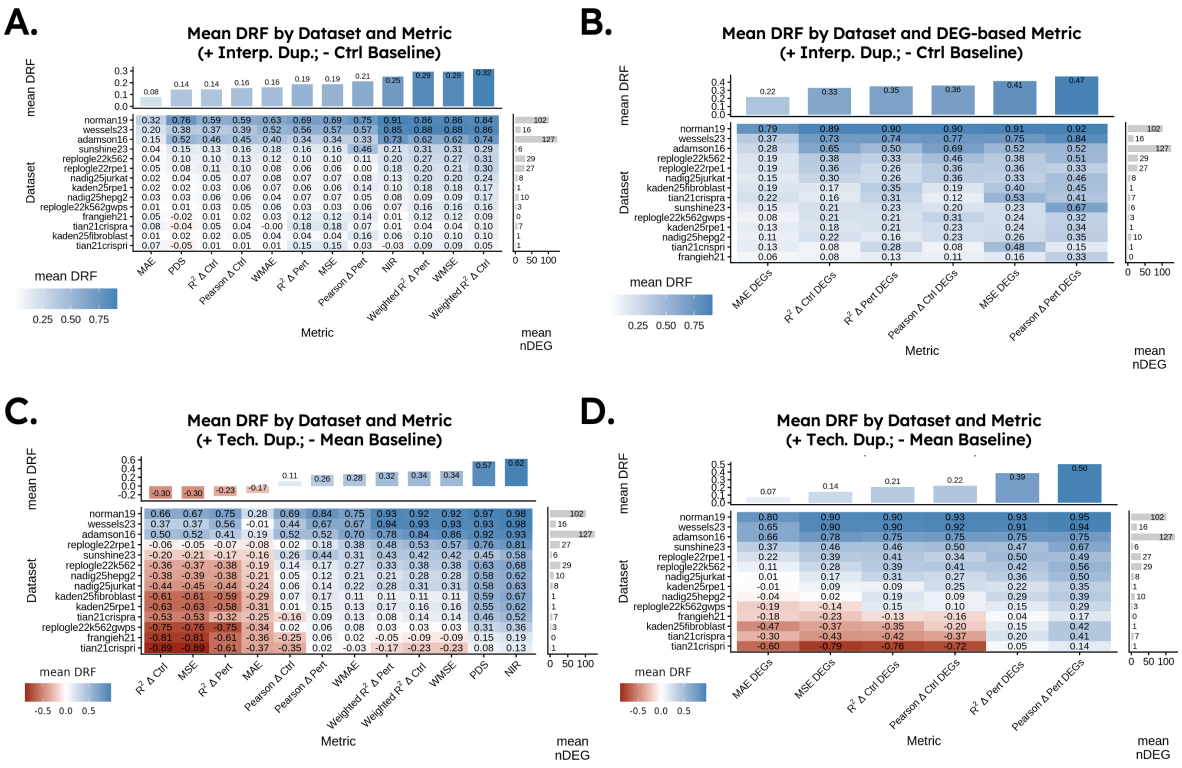

Figure SN4: Calibration DRF with alternative positive and negative controls. (A) Heatmap summarizing the calibration (in terms of mean DRF) of twelve benchmarking metrics across all datasets. For the DRF calculation, the positive control was taken to be the interpolated duplicate, and the negative control was the control cell baseline. Bar plots above and to the right of the heatmap show the mean DRF per dataset and the mean number of differentially expressed genes (DEGs) per perturbation, respectively. (B) Same as (A) except showing DRF for the six DEG-based metrics. (C) Same as (A) except the positive control is the technical duplicate baseline and the negative control is the mean baseline. (D) Same as (C) except for the DEG-based benchmarking metrics.

In this study, we used the interpolated duplicate as the positive control and the mean baseline as the negative control; however, DRF can be computed using any experimentally appropriate

control definitions. To assess DRF robustness to control choice, we recalculated DRF using the control cell baseline as an alternative negative control and the technical duplicate as an alternative positive control (fig. SN4). Although absolute DRF values vary with control choice, the top-ranked metrics by overall calibration remain largely consistent (fig. SN4B–C). In particular, weighted R^2_Δ, NIR, WMSE, and Pearson(Δ_{pert}) on DEGs continue to display strong calibration across datasets. Moreover, MAE and MAE DEGs display consistently poor calibration in all cases.

However, we did observe a few notable differences in calibration due to control choice which reflect the consequences of our previous findings regarding the brittleness of the technical duplicate and the bias of the control cell baseline. Selecting the technical duplicate as a positive control in datasets with weaker perturbations resulted in negative calibration in several cases, especially for metrics less sensitive to perturbation-specific signals (fig. SN4D–E). This is consistent with our prior analysis (Extended Data Fig. 4), and indicates the degree to which the technical duplicate is a suboptimal predictor of unperturbed gene expression compared to the mean baseline, and thus why the interpolated duplicate is a better choice overall. A notable exception is retrieval metrics (PDS/NIR), which showed strong calibration even with the technical duplicate as the positive control; because these metrics assess which perturbation is most similar to the prediction rather than measuring direct errors or correlation, they sidestep the signal dilution problem. Regarding the control cell baseline, it has been previously observed that control cells in perturbation experiments exhibit varying degrees of systematic bias [1, 2], particularly in unweighted error metrics. Thus, when used as a negative control for DRF calculation, it was unsurprising to observe that MSE and MAE appeared marginally better-calibrated as, by comparison, the interpolated duplicate would have been stronger.

Together, these results indicate that our conclusions regarding metric calibration are generally not dependent on a specific control construction while highlighting trade-offs between positive and negative control definitions.

0.2 Number of cells per perturbation

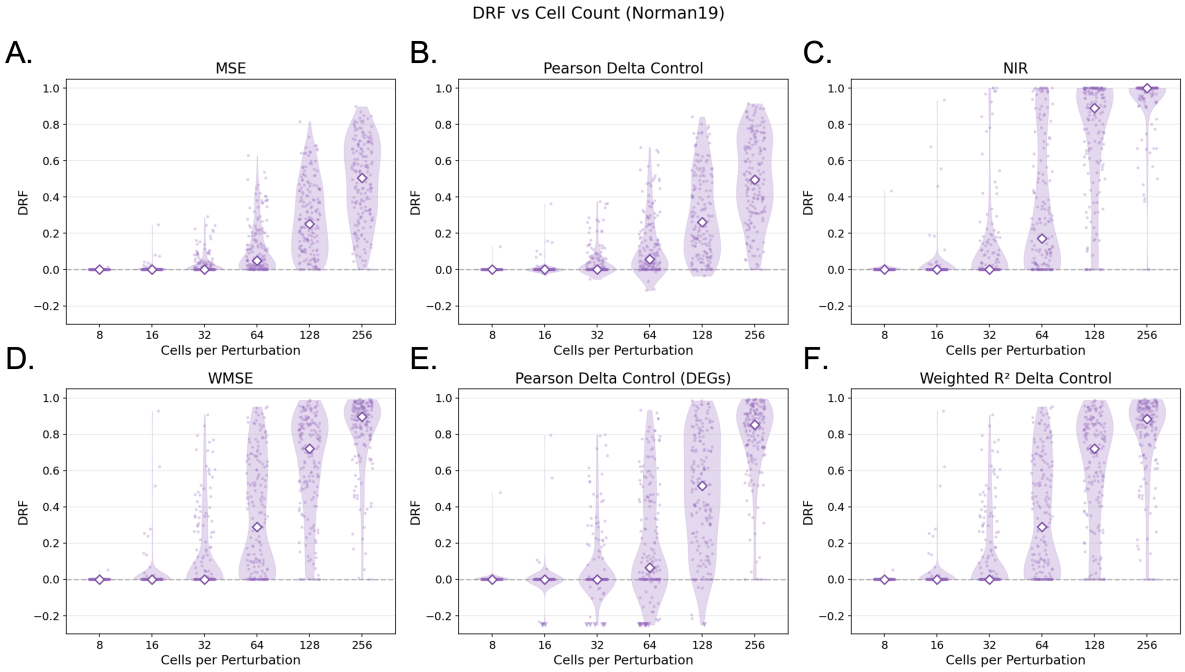

Figure SN5: Relationship between number of cells per perturbation and DRF across metrics in the Norman19 dataset. Various metrics were tested: (A) MSE, (B) Pearson Δ_{ctrl}, (C) NIR, (D) WMSE, (E) Pearson Δ_{ctrl} DEGs, and (F) Weighted $R^2(\Delta_{ctrl})$. Values below -0.25 are clipped for visualization, and are represented by down arrows.

Because the technical and interpolated duplicate baselines estimate perturbation-specific signal from finite samples, we next examined how calibration depends on the number of cells per perturbation using a controlled downsampling analysis of the Norman19 dataset (see Methods). As expected, DRF increased monotonically with sampling depth across all metrics (fig. SN5). However, the rate of improvement differed between metrics. Common metrics such as MSE and Pearson(Δ_{ctrl}) exhibited less calibration at low sampling depths (DRF ≈ 0 at $n = 8$) and only modest separation at intermediate depths (e.g., DRF 0.17 at $n = 64$). In contrast, well-calibrated metrics demonstrated better robustness: at $n = 64$, WMSE achieved DRF = 0.61, NIR = 0.40, and weighted $R^2_{\Delta\text{ctrl}}$ = 0.32. DEG-restricted metrics similarly showed improved robustness relative to their non-DEG counterparts. For example, Pearson(Δ_{ctrl}) on DEGs achieved DRF = 0.27 at $n = 64$, compared to 0.14 for the non-DEG version.

Together, these results indicate that although adequate sampling depth is required for reliable calibration, well-calibrated metrics maintain better sensitivity at lower cell counts.

0.3 Sequencing depth per cell

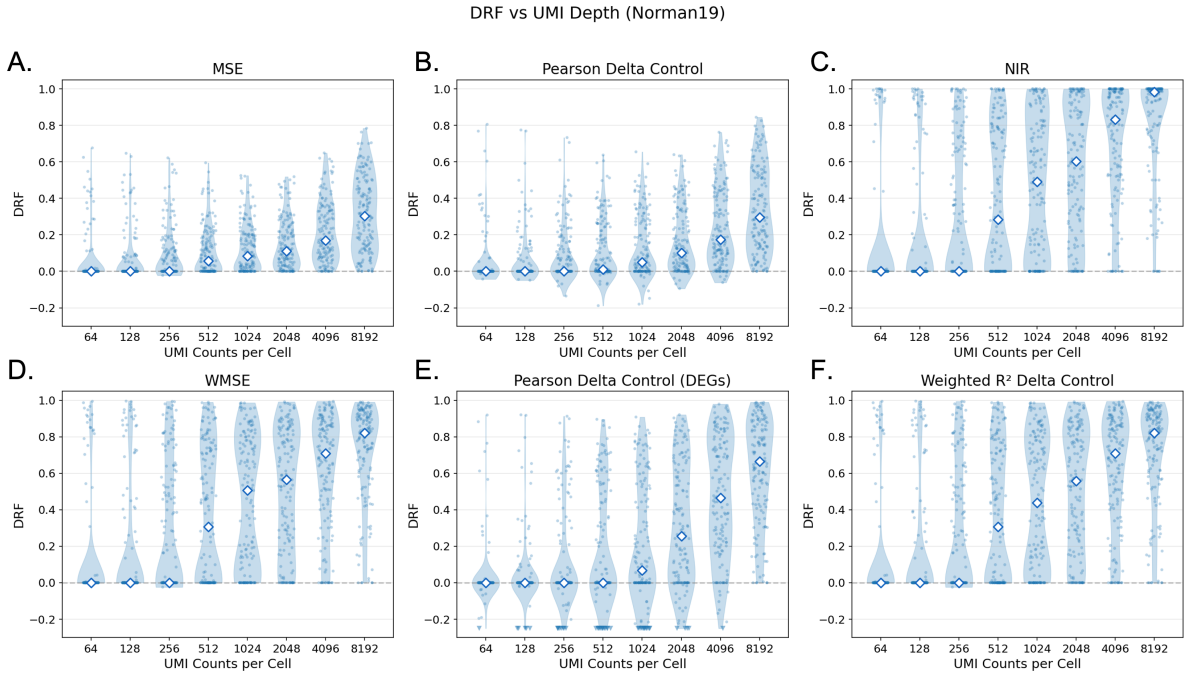

Figure SN6: Relationship between number library size (UMI count per cell) and DRF across metrics in the Norman19 dataset. Various metrics were tested: (A) MSE, (B) Pearson Δ_{ctrl}, (C) NIR, (D) WMSE, (E) Pearson Δ_{ctrl} DEGs, and (F) Weighted $R^2(\Delta_{ctrl})$. Values below -0.25 are clipped for visualization, and are represented by down arrows.

We next examined how calibration depends on sequencing depth per cell with UMI downsampling analysis of the Norman19 dataset (see Methods). As expected, DRF increased monotonically with UMI depth across all metrics (fig. SN6). However, the rate of improvement again differed between metrics. Common metrics such as MSE and Pearson(Δ_{ctrl}) exhibited minimal calibration at low depths (average DRF ≈ 0.05 at 64 UMI counts) and only modest separation at intermediate depths (e.g., average DRF $= 0.13$ at 1024 UMI counts). In contrast, well-calibrated metrics demonstrated greater robustness: at 1024 UMI counts, WMSE achieved DRF $= 0.44$, NIR $= 0.47$, and weighted $R^2_{\Delta_{\text{ctrl}}} = 0.44$.

DEG-restricted metrics similarly showed improved robustness to UMI downsampling relative to their non-DEG counterparts. For example, Pearson(Δ_{ctrl}) on DEGs achieved DRF $= 0.19$ at 1024 UMI counts, compared to 0.13 for the non-DEG version.

Together, these results indicate that, as with cell count per perturbation, adequate sequencing depth per cell is necessary for reliable calibration, and well-calibrated metrics recover meaningful separation at lower UMI depths compared to poorly calibrated alternatives.

0.4 Number of top highly variable genes

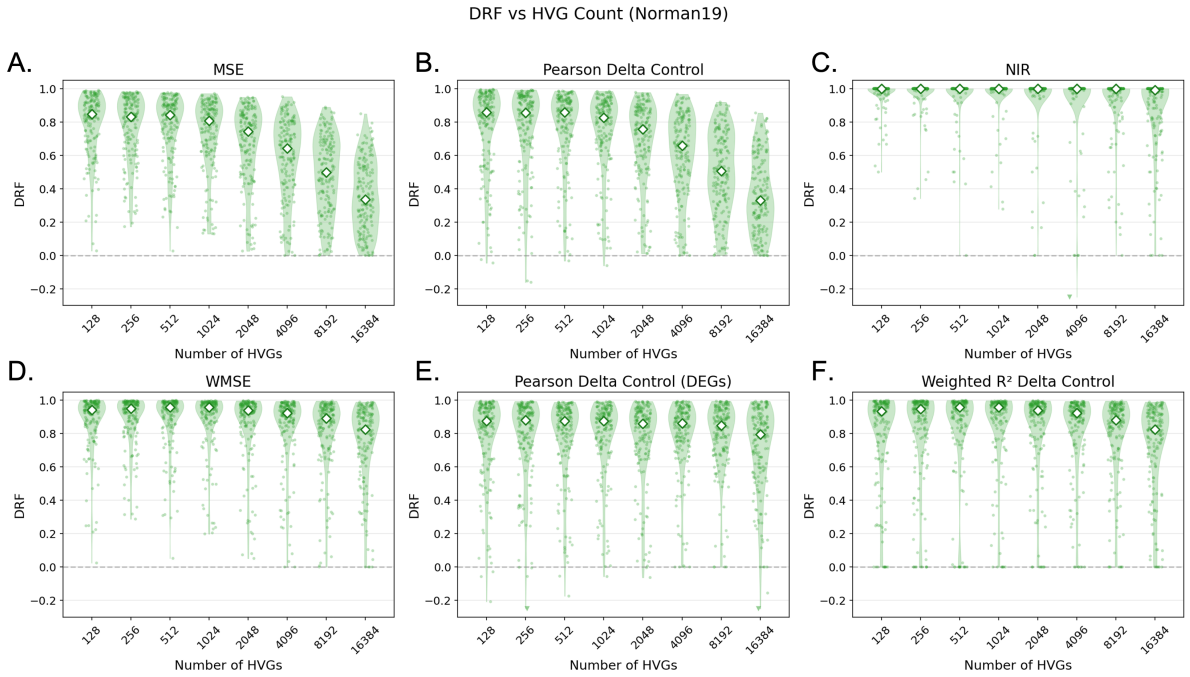

Figure SN7: Relationship between number the number of HVGs selected and DRF across metrics in the Norman19 dataset. Various metrics were tested: (A) MSE, (B) Pearson Δ_{ctrl}, (C) NIR, (D) WMSE, (E) Pearson Δ_{ctrl} DEGs, and (F) Weighted $R^2(\Delta_{ctrl})$. Values below -0.25 are clipped for visualization, and are represented by down arrows.

We next examined how calibration depends on the number of highly variable genes (HVGs) selected during preprocessing using a controlled sweep of $n_{\text{top_genes}} \in \{128, \ldots, 16384\}$ in the Norman19 dataset (see Methods). Unlike cell count or UMI depth, increasing the number of selected genes expands the evaluation space and dilutes perturbation-specific signal with low-variance features, thus exacerbating the signal dilution problem described previously (see Main).

Consistent with this expectation, DRF decreased as larger HVG sets were included (fig. SN7). However, the magnitude of degradation differed across metrics. Common metrics such as MSE and Pearson(Δ_{ctrl}) showed substantial declines in calibration as the number of genes increased (mean DRF decreasing from 0.79 and 0.77 at 128 HVGs to 0.35 and 0.36 at 16384 HVGs, respectively). In contrast, well-calibrated metrics demonstrated greater robustness to signal dilution. For example, WMSE decreased from DRF = 0.88 at 128 HVGs to 0.76 at 16384 HVGs, and weighted $R^2_{\Delta\text{ctrl}}$ from 0.80 to 0.73. DEG-restricted metrics and NIR similarly retained high calibration across the HVG sweep.

These results indicate that while expanding the gene space can dilute perturbation-specific signal, well-calibrated metrics designed to address this issue maintain calibration even under increasingly broad feature selection, whereas common unweighted metrics are more sensitive to signal dilution.

0.5 Perturbation strength impact on calibrated metrics

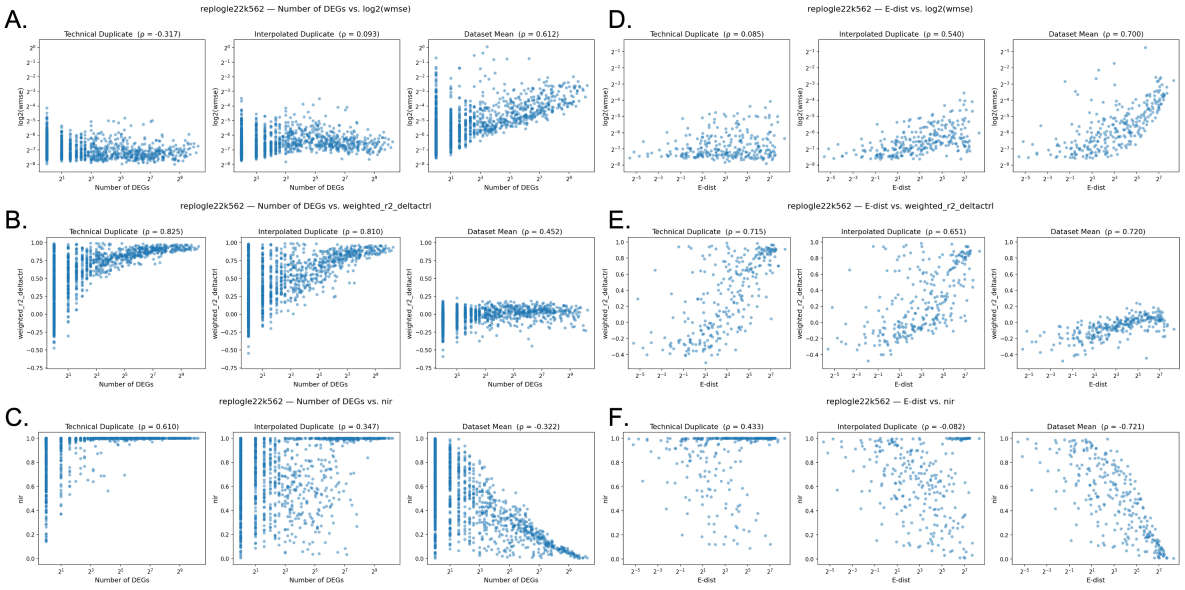

Figure SN8: Relationship between perturbation strength and performance of positive and negative baselines in well-calibrated metrics for the Replogle22 K562 unseen genetic perturbation prediction task. (A) Scatter plots showing the relationship between WMSE (lower is better) and perturbation strength (in terms of number of DEGs) for technical duplicate, interpolated duplicate, and mean baselines. (B) Same as (A) for the weighted $R^2_{\Delta\text{ctrl}}$ metric (higher is better). (C) Same as (A) for the NIR metric (higher is better). (D-F) Same as (A-C) with E-distance [3] as the perturbation strength measure (calculated with a max of 300 randomly-selected perturbations due to quadratic computational complexity).

To further assess validity of well-calibrated metrics, we examined the relationship between perturbation strength and performance of positive controls (fig. SN8), quantifying strength by either the number of DEGs or by E-distance [3]. Consistent with our earlier results, the separation between informative and uninformative baselines increased with effect size across metric families (fig. SN8), thus indicating improved calibration. As an additional check, we evaluated trends in positive control performance with increasing strength. For retrieval- and fit-based metrics (NIR and weighted R^2_Δ), positive controls usually showed improved performance for stronger perturbations (fig. SN8B-C,E-F), consistent with the expectation that larger transcriptional shifts provide clearer perturbation signal. In contrast, direct reconstruction metrics such as WMSE exhibited weaker or non-monotonic trends (fig. SN8A,D), similar to prior reports for MSE [4], likely reflecting the increased difficulty of reconstructing larger distributional shifts. Together, these results indicate that retrieval- and fit-based well-calibrated metrics (NIR, weighted R^2_Δ) respond to changing perturbation strength, while direct reconstruction metrics (WMSE, like MSE) do not show monotonic improvement with stronger perturbations because larger distributional shifts are intrinsically harder to reconstruct point-wise despite increasing the learnable perturbation signal.

References

[1] Mejia, G. M. *et al.* Diversity by design: Addressing mode collapse improves scrna-seq perturbation modeling on well-calibrated metrics. Preprint at https://arxiv.org/abs/2506.22641 (2025).

[2] Viñas Torné, R. *et al.* Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology* **44**, 1050–1059 (2025).

[3] Peidli, S. *et al.* scperturb: harmonized single-cell perturbation data. *Nature Methods* **21**, 531–540 (2024).

[4] Wenteler, A. *et al.* PertEval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. *Proceedings of Machine Learning Research* **267**, 66633–66677 (2025).