nature portfolio

Peer Review File

Deep Learning Perturbation Models *Can* Outperform Baselines on Calibrated Metrics

Corresponding Author: Dr Bo Wang

Parts of this Peer Review File have been redacted as indicated as we could not obtain permission to publish the reports from one of the peer reviewers.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This study introduces a metric (dynamic range fraction) for metrics, specifically for metrics evaluating the quality of models in the perturbation effect prediction space, but potentially with broader applicability. The dynamic range fraction, in short, favors metrics that have a large difference between the performance of a 'trivial baseline' generally a simple statistical predictor such as a mean perturbation effect, and a 'positive control' such as the 'interpolated duplicate baseline'. Specifically this quantifies an achievable score in relation to the maximum (difference from 'trivial' baseline to true target).
The interpolated baseline is a by gene weighting of split half reproducibility and mean baseline, depending on which is likely to be stronger, and is introduced in this study.
Finally the study applies this dynamic range fraction to a suite of metrics designating them as well or un calibrated, and then contextualizes the performance of both older and newer deep learning models across these metrics by the metric calibration. This leads to the opinionated argument from the title, that "Deep Learning-Based Genetic Perturbation Models Do Outperform Uninformative Baselines on Well-Calibrated Metrics". Ultimately, this paper is a direct response to the many benchmarking papers that have come out recently which have highlighted the underperformance of deep learning models relative to simpler methods.

This study makes substantial and welcome contributions in introducing
- a metric for metrics, and
- a strong 'mark to beat'
- extensive benchmarking of models on different datasets with different metrics
- clear explanation of many of the strengths and weaknesses of various metrics

however it oversimplifies the task of benchmarking models in the perturbation prediction space and also substantially misrepresents the chronological reality of the interplay between improvements in modelling and benchmarking that together drive progress in the field.

While active discourse in this field is healthy and welcome, a modelling vs benchmarking attitude is neither going to increase the impact of this work, nor move the field forward faster.
This should, however. be readily addressable.

Major issue number one, over simplification of what is required to benchmark models in the perturbational transcriptomics space.

First, I acknowledge it is surely true that some metrics are more useful than others and agree that using an empirical method like the dynamic range fraction to prioritize them is a fundamentally good idea. However, the paper is written as if this is now a solved problem, repeatedly ignoring complexity, weaknesses and qualifications of the dynamic range fraction, and ultimately the need for complementary metrics.

Example 1 - misrepresenting benchmarking as if it were a solved problem

throughout, wording like the following indicating a solved problem needs to be adjusted to be scientifically accurate

" Calibrated evaluation thus explains prior reports of model underperformance, revealing that deep learning models do outperform uninformative baselines."

"Calibrated evaluation" ->
"The calibration of metrics here" (as this paper does not represent the beginning and end of metric calibration)

"deep learning models do outperform uninformative baselines" ->
"deep learning models can" [because it's hardly every DL model] "outperform uninformative baselines on the most calibrated metrics ..." [as it is not often across the board]

This needs a careful check throughout, also very explicitly including the title of the paper

Example 2- ignoring complexity and weaknesses of the dynamic range fraction

The dynamic range fraction is extremely dependent on the selection of 'uninformative baseline' and 'positive control'. Before recent benchmarking papers introduced the mean perturbation effect baseline, it was (and frankly still is) common to take the controls as a 'no effect' baseline. This would have weighted a metric like pearson delta as far more calibrated. Similarly, a totally different uninformative baseline (say control expression with the CRISPRi target gene downregulated by 100x),
is likely to drastically reorder the 'calibration' levels of metrics. This trivial baseline tends to be performative on retrieval, and would do poorly on delta metrics.

Ideally, I would like to see how the empirical ranking of metrics would change by differences in the choice of baseline.

Example 3 - the need for complementary metrics

Fig 1. C of the paper includes a cartoon depiction of an 'Uncalibrated' metric. Specifically, it is a rather saturated metric, where
there is negligible difference between the 'Trivial baseline and Positive control', and little space to differentiate the models. However, if the authors wanted to accurately represent the state of modelling at the time the
"Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear baselines" was ran; the pink deep learning
bar would be non-negligibly behind the trivial baseline on many metrics. This is not simply a problem of saturation, but that these models do have substantial weaknesses (likely very fixable, once identified and acknowledged).

The major recent example of this, is how the ARC Virtual Cell challenge allowed for clipping of the MAE metric; this allowed for solutions with significant and biologically unrealistic submissions (from exaggerated log fold changes, up to and initially [it was later patched] including submissions with negative expression) to climb high on the leaderboard as neither looking at differential expression nor retrieval alone actually captures any form of overall realism;
especially retrieval does not capture what is close, but what is closest; while differential expression captures what significantly
changes, but not by amount nor direction.

The top 'calibrated' metric (retrieval-like) proposed here, would very much share that qualifier, and should be complemented by absolute error metrics. That is, when developing a model I might not optimize for nor expect major improvements on something like MAE, but
I would still be highly skeptical of a model that performed significantly below baseline on MAE.

Honest presentation of the importance of complementary metrics, and presenting 'calibration' as important for prioritizing metrics to focus
on yes, but not as an excuse to allow 'uncalibrated' (especially simply saturated) metrics to collapse, would greatly improve the positive impact of this paper on the field.

Major issue number 2) misrepresentation of the chronological reality

This manuscript makes strong claims, e.g.

"In the Replogle K562 benchmarking dataset, similar to
others (2, 3, 5, 6, 8), we find that models generally do not outperform the mean and linear baselines when evaluated with poorly calibrated metrics such as MSE and Pearson(Δctrl)
(Fig. S9A,B; table 1)"

while actually looking at Figure S9A,B shows that around half of the DL models, and consistently including the newer models scLAMBDA
and PRESAGE and often also fMLP variants outperform the mean baseline on these 'poorly calibrated' metrics. PRESAGE

and scLAMBDA
Also have a clear edge on the linear baseline. This was not reported in previous benchmarking studies,
for the simple reason that these models are newer than the studies in question. Presumably, the benchmarking studies and the awareness they brought to some
of the challenges of the data and the easy ways to fool ourselves in modelling, had an influence on the development of these models and encouraged the developers there of aim higher

another example is where the authors write
" In the combination-gene task, some models likewise exceeded null
baselines and (especially for PRESAGE) even surpassed the
additive baseline in the Wessels23 dataset."

without providing inline (although the quoted sentence is already in the discussion) mention that PRESAGE is newer than the benchmarking
studies that found additive was the best combinatorical model 'yet'.

Once the authors do mention that the consistently performant models PRESAGE and scLAMBDA are newer, they add that they should be included in future benchmarking studies, as if this was an omission of the benchmarking studies, and not a reality of the chronological order of when work was performed and released.

So yes, more DL models outperform baselines on more calibrated metrics, but the most prominent result is that PRESAGE is consistently a top performer (mostly the top, multiple times above the 'positive control'); and it is above baselines regardless of metric calibration.

A DL model has access to all the same information that a simple baseline does; and should be able to (and clearly can!) learn to outperform it or match it globally. The older models do better on the metrics here, but clearly also
have areas for improvement that the benchmarking studies helped surface. The results of benchmarking studies were imperfect, yes , as this whole
field is learning and shifting rapidly, but they helped spur progress and a new wave of legitimately across-the-board competitive models. Quite frankly, the focus on 'older models would have ranked somewhat better compared to baselines if benchmarked exactly as so' takes away from the progress and excitement of the field and the results of this paper as well.

The wording should be adjusted throughout to bring more acknowledgement to the global progress the field has made; progress that the authors are part of!

As a start, adding 'now' to the title would help.

==== medium importance thoughts

Med 1) these experiments have substantial batch effect, yet the authors chose not to compare to within batch controls for delta metrics (at least I could not find any within-batch matching searching 'batch' and 'gem' (for gem group)).
Given there is confounding between batches and perturbations that allows unmatched metrics to confuse pertubation and batch effect, and that not batch matching inflates batch noise relative to perturbation
effect, I would appreciate it if the authors either a) indicate they perform this and make sure to say it, b) add delta metrics with batch matching,
or c) provide a thorough explanation of why not. I do appreciate not batch matching is extremely common in this field, but in a paper otherwise making some very solid contributions to benchmarking best practice, it seems like a major omission.

Med 2) I notice that the metric WSME has a notably different sort order of models than most other metrics, with
especially the rank of scGPT rising. The idea to weight by comparison to
the whole dataset, and not just controls to circumvent systematic bias is appreciated, and I'd love to have
a better understanding of (and in the future use) this metric.
If the authors know _why_ and what the strength of scGPT (or error of other models)
is, it could improve understanding and trust of this metric. Additionally, I'd like to know the typical number
of genes receiving a meaningful amount of weight.

=== Minor
in introducing the technical duplicate the authors explain it as
“if we repeated the experiment with no batch effects, how predictive would
our new experiment be?” but it is not batch effects coming into play here, it
is variance between individual cells (both technical and biological cell to cell,
not batch to batch, noise)

(Remarks on code availability)
There is a readme. It looks like effort has been made for reproducibility, but I did not try to run it.

Reviewer #2

(Remarks to the Author)
This manuscript tackles a contentious question in single-cell perturbation modeling: whether deep learning perturbation-response models truly add value compared to simpler baselines.
The first half of the manuscript explores the extent to which frequently used benchmarking metrics are able to distinguish between an appropriately chosen positive control and negative control, using a newly introduced meta-metric (Dynamic Range Fraction (DRF)). Across multiple datasets/metrics, they argue that DEG-aware and rank-based metrics (e.g., WMSE, weighted R^2 variants, NIR) are better calibrated than commonly used MSE and Pearson correlation of differentially expressed genes.
In the second half of the manuscript, the authors test 8 different model configurations on unseen gene and unseen combination prediction tasks, showing that for certain well-calibrated metrics certain deep learning methods do outperform linear baselines.

General comments
Given the sparse and noisy nature of biological cell states and single-cell RNA-seq measurements, the introduction of a positive control is an important and welcome contribution. However, because the technical duplicate relies on splitting cells within each perturbation, its performance may be sensitive to the number of cells available per perturbation. While the authors acknowledge this issue (line 468), the manuscript would be strengthened by more explicitly examining how cell number per perturbation influences the interpretation of results and the robustness of the proposed conclusions. Concretely, stratifying perturbations by cell-count bins (or overlaying cell counts) and showing how DRF or control separation changes with cell number would clarify when the positive control is reliable and when it may be noisy or misleading.

One possible extension would be to assess the robustness of the conclusions under a different positive control definition. For example, Jiang et al. (Nature Cell Biology, 2025 [https://www.nature.com/articles/s41556-025-01622-z] (https://www.nature.com/articles/s41556-025-01622-z)) introduced Mixscale scores to quantify the extent to which individual cells respond to a perturbation. Intuitively, well-behaved evaluation metrics should assign higher scores to predictions corresponding to cells with higher Mixscale responses.

The E-distance metric introduced for single-cell data by Peidli et al. (Nature Methods, 2024 https://www.nature.com/articles/s41592-023-02144-y) could also be considered, either as a metric for comparing predicted and ground-truth gene expression or as a way to quantify perturbation effect size. In the latter case, perturbations with larger E-distance would be expected to be easier to predict, and this information could be leveraged to assess metric validity (e.g., models should perform worse for perturbations with lower E-distance, as demonstrated in related work by Wenteler et al). This could serve as an additional sanity check on whether the proposed “well-calibrated” metrics behave monotonically with perturbation strength, and whether DRF conclusions generalize beyond the specific positive control construction.

A recent landmark study in single-cell perturbation prediction is the Arc Institute’s State model ([https://www.biorxiv.org/content/10.1101/2025.06.26.661135v2.full.pdf] (https://www.biorxiv.org/content/10.1101/2025.06.26.661135v2.full.pdf)), which was reported to outperform linear baselines. Including elements of this work could strengthen the present manuscript. In particular, the authors could consider (i) evaluating additional metrics introduced in the State study (e.g., p-value AUPRC, DEG overlap accuracy, and effect size prediction), and/or (ii) including State in the foundation-model benchmarking analysis. This would be especially relevant given that a central claim of the paper is that reports of deep learning models failing to outperform linear baselines largely stem from metric choice (as reflected in the title), whereas the State model reports improved performance even under commonly used metrics such as Pearson correlation.

Relatedly, the manuscript would benefit from clearer positioning of scope and novelty relative to the rapidly developing SOTA landscape in perturbation prediction. Several recent methods report improvements over scGPT and/or linear baselines using distributional and generative modeling approaches (e.g., flow/diffusion paradigms) (e.g., https://arxiv.org/html/2511.02986v1, [https://www.biorxiv.org/content/10.1101/2025.04.11.648220v1] (https://www.biorxiv.org/content/10.1101/2025.04.11.648220v1), [https://openreview.net/forum?id=QSGanMEcUV] (https://openreview.net/forum?id=QSGanMEcUV)). Since the manuscript’s narrative emphasizes metric choice as a key explanation for discrepant findings about whether deep learning adds value, the conclusions would be strengthened by explicitly discussing how these methods fit into the proposed framework (and ideally, by including at least one representative example in the benchmark). Otherwise, it remains somewhat unclear whether the reported “DL vs linear” conclusions generalize beyond the specific set of model classes evaluated here.

While this point is briefly acknowledged in the Discussion (Line 359), statistically significant improvements do not necessarily imply biological usefulness. The manuscript would be strengthened by more explicitly connecting evaluation metrics to concrete biological use cases. For example, which metrics are most appropriate for tasks such as DEG recovery, global transcriptome accuracy, or perturbation retrieval, and do deep learning models provide clear advantages in these settings? Including at least one downstream analysis demonstrating that “well-calibrated” metrics align with biological utility (such as improved pathway or TF target recovery, or biologically meaningful perturbation similarity networks) would help ground the benchmarking results in practical biological relevance.

Specific comments:
Line 118–120: the use of “genes” here is confusing as the authors handle both gene perturbation labels and the genes

whose expression is being measured; consider rephrasing the second clause (and avoiding anthropomorphisms).
Figure S1E: It is interesting that the mean baseline outperforms the technical duplicate for many perturbations, particularly those with few DEGs. Could the authors assess whether this pattern varies with the number of cells per perturbation, for example by overlaying cell counts in a plot like Fig. S1E? More generally, how does the usefulness of the DRF depend on the number of cells per perturbation?
Figure S7A: it could make interpretation easier to more clearly label that higher scores are better with DRF (as shown in the schematic in the main figure).
Figure 2D,E: why does DRF show such a range across datasets? This doesn't seem to be due only to variation in the number of DEGs. It would help to discuss additional drivers (e.g., cell-count distributions per perturbation, sparsity, heterogeneity, batch effects, library size differences), or to provide a decomposition/diagnostic correlating DRF with these factors.
Figure S9B: how did control mean perform on this task?
Figure S9: minor: the color and style of the line for Mean in the legend box does not match the figure.
Lines 220–225: scGPT is the only model evaluated both as a predictive model and as a source of embeddings with an MLP probing head. Could the authors clarify the rationale for including both evaluations only for scGPT (and whether similar probing should be applied consistently to other foundation embeddings/models)?
Line 475: Why are 8192 random cells selected? Similarly, line 518, why are 8192 highly variable genes selected? Please justify or show sensitivity/robustness to these choices.
Line 493: The interpolated duplicate uses adjusted DEG p-values to define the interpolation weight ($\alpha = 1 - p_adj$). Because p-values depend on both effect size and statistical power, and are not directly proportional to biological effect size, this modeling choice may influence the behavior of the positive control in ways that are not immediately transparent. In order to evaluate the impact of this choice, it would be helpful to plot the observed distribution of α (or p-values) across perturbations, or to assess whether alternative definitions based on effect size or ranked significance yield similar conclusions.
Line 521: does the splitting happen on the test set? If yes, could this be made more explicit?
Throughout the manuscript, the authors refer to Pearson Δcontrol as "control-referenced deltas." However, in prior studies this metric is often computed either over all differentially expressed genes or over the top-ranked DEGs. Please clarify how this metric is implemented in the present study, in particular the choice of gene set used for its calculation, as referenced in line 548.
Line 568: the MLP architecture for MLP probing is quite large; what is the impact of this decision? (e.g., sensitivity to probe capacity / risk that probe dominates differences between embeddings).
Line 585: what is the justification for using a one-sided test?
Fig. S9A,B: Given the small absolute effect sizes, the error bars in the forest plots play an important role in interpreting the results. However, the description of the bootstrap procedure (line 596) is underspecified. Please clarify what is being resampled (e.g., perturbations vs. cells), whether the paired structure is preserved, and which statistic is recomputed.

(Remarks on code availability)

Author Rebuttal letter:

Overall comments
We thank reviewers for their careful reading of our manuscript and for their thoughtful and constructive feedback. Below, reviewer comments are shown in red italics, followed by our responses in black text.

We provide our revised manuscript in both PDF and LATEX format. The source .zip includes main.tex (clean revised manuscript) and diff.tex (tracked-changes version generated via latexdiff against the original submission)—and we provide rendered PDFs of both versions. Given the extent of the revisions, we recommend the clean version for readability. All page and line references in this response (e.g., Page 5, Line 345) refer to the clean PDF.

Before addressing our point-by-point response to each reviewer separately, we would like to share a few revisions relevant to both reviewers:

Reformatting to fit Brief Communication limits

To address the limits of the Brief Communication format, we have:

• Condensed the Abstract from 182 to 70 words (in three sentences):

– "Recent benchmarks report that deep learning-based genetic perturbation models fail to outperform uninformative baselines. Herein, we introduce a novel positive control baseline and metric calibration framework, showing—across 14 datasets and 18 metrics—that common benchmarking metrics (e.g., MSE and Pearson Δ) are frequently miscalibrated, with reduced sensitivity to measure positive model performance. Under well-calibrated metrics, we find that deep learning-based genetic perturbation models can outperform

uninformative baselines."

• Removed section headers from Main text. Condensed from ~4200 to ~3200 words.

• Moved details of analyzing the drawbacks of the technical duplicate baseline from "Results: Negative and Positive Controls" along with original fig S2-4 to "Supplementary Note 1". The Main text now summarizes this supplementary note and directs readers to read it for further detail.

Title change

In response to comments from both reviewers, we felt it was appropriate to update our title from "Deep Learning-Based Genetic Perturbation Models Do Outperform Uninformative Baselines on Well-Calibrated Metrics" to "Deep Learning-Based Genetic Perturbation Models Can Outperform

Uninformative Baselines on Well-Calibrated Metrics". We agreed with reviewers our original title was opinionated and could be interpreted as claiming models universally outperform trivial baselines, which is not the case. Our goal with the title is to rebut a prevailing sentiment that models are not yet capable of outperforming baselines, and thus Can is a better choice than do.

Citation of concurrent independent work

During revision, a related preprint by Cole et al. [1] was posted, which systematically benchmarks foundation model embeddings for perturbation response prediction. Their work reaches complementary conclusions to ours—in particular, that perturbation embeddings based on biological prior knowledge can outperform simple baselines. However, it addresses a distinct question (which representations are most informative) rather than the issues surrounding metric calibration that are central to our study. We have added a citation in the Main text to acknowledge this concurrent work (Page 5, Line 266):

Concurrently, an independent preprint by Cole et al. [1] systematically benchmarked foundation model and prior-knowledge-based embeddings for perturbation response prediction. Their work reached complementary conclusions, showing that embeddings encoding prior knowledge can improve prediction, further underscoring the value of prior knowledge for informative representations of genetic perturbations.

Update to DEG-based metrics

During revision, we re-examined the statistical behavior of our DEG-based evaluation metrics. In the original submission, DEG-based metrics were computed over all statistically significant DEGs

per perturbation. While conceptually intuitive, this approach resulted in substantial variability in the number of evaluated genes across perturbations. Because correlation- and error-based metrics are sensitive to the number of observations, perturbations with very few DEGs can yield unstable or extreme values (e.g., Pearson Δ computed on $N = 2$ DEGs is necessarily either 1, –1, or undefined), whereas perturbations with many DEGs produce more stable estimates.

In practice, this variability required excluding perturbations with insufficient DEG counts to encourage stable metric estimation. In datasets with weaker perturbation effects, this sometimes resulted in too few evaluable perturbations to compute stable DEG-based Dynamic Range Fraction (DRF) estimates, leading to blank entries in the original heatmaps for multiple datasets and DEG metrics.

To address this point, we standardized DEG-based evaluation to the top 100 DEGs per perturbation (ranked by adjusted p-value), consistent with recent benchmarking studies [2, 3, 4]. This modification ensures a fixed evaluation space per perturbation, stabilizes metric estimates, enables inclusion of all datasets in DEG-based DRF analyses, and improves comparability across perturbations and datasets.

We decided to select $N = 100$ DEGs to balance the goal of capturing niche, biologically-relevant

perturbation effects with maintaining metric stability and also conducting evaluations in a gene space size more relevant to gene module– and pathway-level modeling goals. In addition, because DEG-based metrics are computed over a different gene feature space than global transcriptome metrics, we now report DEG- and non-DEG metrics separately in DRF heatmaps (Non-DEG heatmap is Fig 2E and DEG heatmap is Fig S5 now). This separation reflects their distinct evaluation objectives and clarifies that calibration comparisons are made within, rather than across, different gene spaces.

Importantly, this refinement strengthens rather than alters our central conclusions: DEG-based metrics (e.g., Pearson Δctrl DEGs) exhibit improved calibration relative to their non-DEG counterparts, particularly in datasets with weaker perturbation effects. A direct comparison of representative figures before and after this standardization is shown above.

Additionally, key analyses from this study (including this one) were rerun in a clean computing environment to help ensure reproducibility. This resulted in minor numerical differences in a small number of values (e.g., DRF of weighted R2 (Δctrl) in Norman19 shifted from approximately 0.85 to 0.84), attributable to minor computational or environment-level differences. These changes were numerically small and did not affect calibration patterns, model rankings, or any conclusions of the study.

Of note, we also added new DEG- and non-DEG-based metrics in response to reviewer feedback, as discussed in the relevant sections below.

Comparison of alternative baseline definitions for calibration metametric

One point raised by both reviewers is that we did not examine alternative definitions of baselines

for the DRF calibration meta-metric, which would be valuable to contextualize the utility and limitations of our calibration approach. In agreeing with this, we have now included a new analysis in which we show the impact of changing the positive baseline (from Interpolated Duplicate to Technical Duplicate) and the negative control baseline (from Mean Baseline to the Control Cell Baseline).

In the above figure (combination of new Fig 2E, Fig S5, and Fig SN4), we show the impact of selecting other positive and negative controls for the DRF calculation across datasets and metrics. While there are baseline-specific findings, the best-calibrated metrics under the standard DRF (NIR, WMSE, Weighted R2 Δctrl , Weighted R2 Δpert , and Pearson Δpert DEGs) are ranked highly regardless of baseline, whereas MAE and MAE DEGs remain poorly calibrated in all cases.

Notable differences in metric calibration between baselines include (i) the marginally improved calibration of MAE/MSE under the DRF with control cell baseline as the negative baseline, (ii) negative DRF values for some metrics where the technical duplicate is used as the DRF positive baseline, and (iii) strong performance of retrieval metrics with the technical duplicate as the positive control. These results are readily explainable based on prior work showing that control cell populations display systematic bias [5, 2]—thus, inflating DRF values when used as the negative baseline in unfiltered error-based metrics—and our results (“Negative and Positive Controls”) showing the

brittleness of the technical duplicate baseline for weaker perturbations in common metrics (Fig S6). Third, retrieval metrics (PDS/NIR) show strong calibration with the technical duplicate as the positive control, because these metrics ask which perturbation is most similar to the prediction rather than measuring direct errors or correlation, thereby sidestepping signal dilution. Overall, this analysis indicates that our main conclusions regarding metric calibration are not dependent on the specific control constructions evaluated here, while revealing key trade-offs in the choice of positive and negative control.

This is now discussed in a new Supplementary Note 2: “Reliability of Calibration Approach” under the subsection “Choice of Controls for DRF Calculation” (truncated for brevity):

...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 $R\Delta$ 2 , 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)... Regarding the control cell baseline, it has been previously observed that control cells in perturbation experiments exhibit varying degrees of systematic bias [5, 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...

Finally, we describe at other points in the work that the choice of positive and negative control is not explicitly set but may be any baselines which are expected to produce robust and uninformative predictions, respectively.

For example, we now include in the main text (Page 5, Line 298):

Beyond perturbation modeling, the DRF framework is not tied to a prescriptive definition of positive or negative controls; instead, it provides a general approach for assessing metric sensitivity relative to any reasonably defined performance baselines.

Evaluation of technical factors influencing the DRF

We thank both reviewers for pointing out that our analysis of factors influencing the performance of the DRF was limited in our original manuscript, and that adding this valuable context would strengthen the description of our calibration approach. In our current revision, we now evaluate several factors which influence the DRF (See new “Supplementary Note 2: Reliability of calibration approach”).

Included here are three data ablation experiments: measuring the impact of (i) the number of cells per perturbation, (ii) the number of UMI counts per cell, and (iii) the number of highly-variable genes (HVGs) selected. Ablation experiments were conducted in the Norman19 dataset due to the higher DRF values, UMI, and cell count overall compared to other datasets, providing greater dynamic range to observe ablation effects.

Cells per perturbation: Starting from the raw data, we subsampled cells per perturbation across a range of target depths (8–256 cells), recomputed preprocessing, DEG analysis, baselines, metrics, and finally recalculated DRF at each level. Calibration improved monotonically with sampling depth across all metrics. However, we observed substantial differences in robustness between metric families. MSE and Pearson(Δctrl) exhibited less calibration at low cell counts and modest calibration at intermediate depths. In contrast, well-calibrated metrics (e.g., WMSE, NIR, R^2 and weighted RΔctrl) achieved greater calibration in general, which was particularly noticeable at lower cell counts. These results are summarized in Fig SN5 (above) and in the “Number of cells per perturbation” subsection of Supplementary Note 2.

Library size: using the scanpy.pp.downsample counts method, we downsampled every cell to a fixed number of UMI counts between 64 and 8192. Similar to our cell count ablation, we observed a strong monotonic relationship between UMI counts per cell and calibration across multiple metrics. Lower sequencing depths resulted in reduced metric sensitivity in all cases as the variance of all estimates increases with dataset noise. However, similar to the cell count analysis,

conventional metrics such as MSE and Pearson Δctrl exhibited worse calibration at the same UMI depth compared to weighted, rank-based, and DEG-based metrics, further exemplifying the utility of well-calibrated metrics to maintain sensitivity under conditions of greater dataset noise. These results are summarized in Figure SN6 (above) and in the “Sequencing depth per cell” subsection of Supplementary Note 2 (“Reliability of calibration approach”).

Highly variable genes: Regarding the number of top highly variable genes selected, we observed an inverse relationship between HVG count and DRF. Increasing the number of genes expands the evaluation space and dilutes perturbation-specific variance with increasingly low-signal features. Under this signal-dilution regime, conventional unweighted metrics (e.g., MSE and Pearson Δctrl) showed marked declines in calibration as the number of selected genes increased. In contrast, weighted, rank-based, and DEG-restricted metrics maintained higher DRF across the HVG sweep, indicating greater robustness to feature expansion and signal dilution. We summarize these results in new Figure SN6 and in the “Number of top highly variable genes” subsection of Supplementary Note #2.

Taken together, we find that variation in DRF across datasets and metrics reflects multiple biological and technical factors that influence signal-to-noise ratio. Importantly, across the regimes we evaluated, the metrics we identified as “well-calibrated” (which, like weighted MSE, were designed to enhance signal in noisy perturbation modeling evaluations) maintained stronger sensitivity to detect positive baseline performance compared to the uninformative mean baseline, especially compared to standard evaluation metrics like MSE and Pearson(Δctrl).

This finding has practical implications for perturbation-response modeling. In realistic settings, datasets often exhibit suboptimal characteristics, such as limited cell counts, reduced sequencing depth, weak perturbation effects, or analysis goals which require predicting gene expression in a broader feature space that includes lower-variance genes. Our analyses show that well-calibrated metrics improve the ability to evaluate models reliably, even under such conditions, whereas conventional metrics are more sensitive to signal dilution and noise.

We summarize these results in new Figure SN7 and in the Supplementary Note 2 “Number of top highly variable genes” subsection.

Reviewer #1

Overall comments

This study introduces a metric (dynamic range fraction) for metrics, specifically for metrics evaluating the quality of models in the perturbation effect prediction space, but potentially with broader applicability. The dynamic range fraction, in short, favors metrics that have a large difference between the performance of a ‘trivial baseline’ generally a simple statistical predictor such as a mean perturbation effect, and a ‘positive control’ such as the ‘interpolated duplicate baseline’. Specifically this quantifies an achievable score in relation to the maximum (difference from ’trivial’ baseline to true target). The interpolated baseline is a by gene weighting of split half reproducibility and mean baseline, depending on which is likely to be stronger, and is introduced in this study. Finally the study applies this dynamic range fraction to a suite of metrics designating them as well or un calibrated, and then contextualizes the performance of both older and newer deep learning models across these metrics by the metric calibration. This leads to the opinionated argument from the title, that “Deep Learning-Based Genetic Perturbation Models Do Outperform Uninformative Baselines on Well-Calibrated Metrics”. Ultimately, this paper is a direct response to the many benchmarking papers that have come out recently which have highlighted the underperformance of deep learning models relative to simpler methods.

This study makes substantial and welcome contributions in introducing

- a metric for metrics, and

- a strong ‘mark to beat’

- extensive benchmarking of models on different datasets with different metrics

- clear explanation of many of the strengths and weaknesses of various metrics

however it oversimplifies the task of benchmarking models in the perturbation prediction space and also substantially misrepresents the chronological reality of the interplay between improvements in modelling and benchmarking that together drive progress in the field.

While active discourse in this field is healthy and welcome, a modelling vs benchmarking attitude is neither going to increase the impact of this work, nor move the field forward faster. This should, however. be readily addressable.

We thank Reviewer #1 for their constructive and thoughtful response. We appreciate both the recognition of this work's contributions and the reviewer's central concern that the original framing risked oversimplifying the challenges of benchmarking perturbation models and underrepresenting the constructive, iterative interplay between benchmarking and model development that has driven progress in the field.

We fully agree that benchmarking for perturbation response modeling is not a solved problem, and that neither metric calibration nor modeling advances alone are sufficient to move the field forward. Rather, progress has emerged from an ongoing feedback loop in which benchmarking studies surface failure modes, motivate improved metrics and diagnostics, and in turn inform the development of stronger, more biologically grounded models. Our intention in this work was not to obfuscate this beneficial feedback loop (or the progressive improvements in modeling driven by it over time), but to provide tools that help interpret benchmarking results more clearly and fairly within this evolving landscape.

In response to the reviewer's comments, we have revised the manuscript to: (i) avoid language

suggesting that metric calibration or benchmarking is complete or definitive; (ii) emphasize the context dependence, assumptions, and limitations of the dynamic range fraction; (iii) explicitly acknowledge the need for complementary metrics, including absolute error–based measures; and (iv) more accurately reflect the chronological development of both benchmarking studies and newer model architectures.

Throughout the revised manuscript, we distinguish more carefully between effects attributable to improved evaluation practices and genuine advances in model design, and we highlight recent models (e.g., PRESAGE, scLambda, and CellFlow) as evidence of modeling progress rather than as counterexamples to earlier benchmarking conclusions. We believe these changes better align the paper with the exciting nature of the progress in this field, while preserving our central finding: that metric calibration plays a critical role in interpreting model performance and in understanding prior reports of underperformance—and that, when metrics are well-calibrated, models can outperform trivial baselines.

We hope the reviewer will find that the revised framing more accurately reflects both the complexity of benchmarking perturbation models and the shared goal of improving model performance along with evaluation rigor.

Major issue number one, over simplification of what is required to benchmark models in the perturbational transcriptomics space.

First, I acknowledge it is surely true that some metrics are more useful than others and agree that using an empirical method like the dynamic range fraction to prioritize them is a fundamentally good idea. However, the paper is written as if this is now a solved problem, repeatedly ignoring complexity, weaknesses and qualifications of the dynamic range fraction, and ultimately the need for complementary metrics.

We thank the reviewer for this important clarification regarding scope and framing. We agree that metric calibration and perturbation model benchmarking are not solved problems, and that the dynamic range fraction (DRF) should be understood as a diagnostic tool for prioritizing and comparing metrics within a given experimental context, rather than as a definitive or universal solution to evaluation.

We have revised the manuscript throughout to avoid language suggesting that benchmarking, calibration, or metric selection are resolved problems (discussed on page 11). In response to your comments and those of Reviewer #2, we now evaluate factors which influence the DRF (beginning on page 6) and the impact of alternative baselines for DRF calculations (beginning on page 4).

We also discuss how robust model assessment requires the use of multiple, complementary metrics,

including absolute error-based measures, to guard against metric-specific failure modes. In particular, we have restructured our “Metrics” methods section to discuss the distinct categories of metrics which we employ, describing their biologically-relevant applications and the complementarity between them. This is discussed further in our response beginning on page 12.

These revisions ensure that the manuscript presents calibration as a framework for organizing and interpreting evaluation results, while explicitly acknowledging its limitations, assumptions, and the ongoing need for complementary metrics as the field continues to evolve.

Example 1 - misrepresenting benchmarking as if it were a solved problem

throughout, wording like the following indicating a solved problem needs to be adjusted to be scientifically accurate:

“Calibrated evaluation” → “The calibration of metrics here” (as this paper does not represent the beginning and end of metric calibration)

“deep learning models do outperform uninformative baselines” → “deep learning models can” [because it’s hardly every DL model] “outperform uninformative baselines on the most calibrated metrics ...” [as it is not often across the board]

This needs a careful check throughout, also very explicitly including the title of the paper

We thank the reviewer for the opportunity to better specify the scope and tone of our manuscript. In response, we have revised the manuscript throughout to avoid language that could be interpreted as presenting benchmarking or metric calibration as a solved problem.

In particular, we:

• Updated the title to replace “Do Outperform” with “Can Outperform”, reflecting that improved performance is demonstrated for multiple deep learning models under the better-calibrated metrics identified here, rather than universally across all models or metrics.

• Adjusted several statements in the Abstract and Main to clarify that our conclusions concern performance under the calibration framework and metrics examined in this work, rather than constituting a definitive resolution of metric calibration more broadly. For example, changing all instances of “calibrated evaluation” to “calibration-aware evaluation” to indicate the evaluation approach incorporates calibration but doesn’t necessarily provide a full solution to it.

• We similarly updated statements to clarify that models often outperformed baselines under better-calibrated metrics, but that this result was not universal. For example, we now clarify that deep learning-based models “usually” or “frequently” outperform baselines, to avoid making universal claims.

These revisions ensure that the manuscript remains scientifically precise while continuing to support the main conclusion: that metric calibration helps explain prior reports of model underperformance, and that under appropriately calibrated metrics, deep learning-based perturbation models can outperform uninformative baselines.

Example 2: Ignoring complexity and weaknesses of the Dynamic Range Fraction.

The dynamic range fraction is extremely dependent on the selection of ‘uninformative baseline’ and ‘positive control’. Before recent benchmarking papers introduced the mean perturbation effect baseline, it was (and frankly still is) common to take the controls as a ‘no effect’ baseline. This would have weighted a metric like pearson delta as far more calibrated. Similarly, a totally different uninformative baseline (say control expression with the CRISPRi target gene downregulated by 100x), is likely to drastically reorder the ‘calibration’ levels of metrics. This trivial baseline tends to be performative on retrieval, and would do poorly on delta metrics.

We thank the reviewer for affording us the opportunity to strengthen these results and our discussion through consideration of alternative baselines for positive and negative calibration controls. In response to this comment and those of reviewer #2, we now include an analysis of alternative

negative and positive controls as discussed on page 4.

Example 3: The need for complementary metrics.

Fig 1. C of the paper includes a cartoon depiction of an ‘Uncalibrated’ metric. Specifically, it is a rather saturated metric, where there is negligible difference between the ‘Trivial baseline and Positive control’, and little space to differentiate the models. However, if the authors wanted to accurately represent the state of modelling at the time the “Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear baselines” was ran; the pink deep learning bar would be non-negligibly behind the trivial baseline on many metrics. This is not simply a problem of saturation, but that these models do have substantial weaknesses (likely very fixable, once identified and acknowledged).

We completely agree with the reviewer and thank them for pointing out the need for greater realism with respect to model performance at the time of Ahlmann-Eltze et al. paper. We have now updated the graphical abstract to more accurately reflect the reality that some perform similarly or worse than trivial baselines.

The major recent example of this, is how the ARC Virtual Cell challenge allowed for clipping of the

MAE metric; this allowed for solutions with significant and biologically unrealistic submissions (from exaggerated log fold changes, up to and initially [it was later patched] including submissions with negative expression) to climb high on the leaderboard as neither looking at differential expression nor retrieval alone actually captures any form of overall realism; especially retrieval does not capture what is close, but what is closest; while differential expression captures what significantly changes, but not by amount nor direction.

The top ‘calibrated’ metric (retrieval-like) proposed here, would very much share that qualifier, and should be complemented by absolute error metrics. That is, when developing a model I might not optimize for nor expect major improvements on something like MAE, but I would still be highly skeptical of a model that performed significantly below baseline on MAE.

Honest presentation of the importance of complementary metrics, and presenting ‘calibration’ as important for prioritizing metrics to focus on yes, but not as an excuse to allow ‘uncalibrated’ (especially simply saturated) metrics to collapse, would greatly improve the positive impact of this paper on the field.

We thank the reviewer for their thorough discussion of the issues the issues inherent in over reliance on single metrics or specific types of metrics, especially in cases where metrics—despite being calibrated—fail to provide a comprehensive depiction of model performance. We agree completely and we address this point by (i) including multiple new metrics, including the MAE, and (ii) discussing the importance of complementary metrics in the manuscript.

Regarding MAE, we agree with the need for error-based metrics as part of comprehensive model evaluation, and agreed that absolute error (L1) based metrics were an important omission from our original manuscript. While we originally included 13 benchmarking metrics, this was expanded to 18 with the inclusion of MAE, weighted MAE (WMAE), MSE/MAE on top DEGs, and perturbation discrimination score (PDS)—an L1-based version of the NIR metric introduced by the ARC institute STATE paper [6] and used in the Virtual Cell Challenge (VCC) [7].

Curiously, we observed that MAE was generally less calibrated compared to MSE overall (page 2). Similarly, while WMAE and MAE DEGs showed better calibration than MAE, they also generally displayed worse calibration overall compared to their L2-based counterparts (WMSE, MSE DEGs). Ultimately, we agree that direct error metrics like MAE and MSE provide a good backstop against certain types of metric hacking, but, based on our calibration results, we suggest that MSE would still be a better metric overall due to its superior calibration.

We now describe this finding along with the broader need for metric complementarity in our Main text (Page 5, Line 302):

Additionally, this framework does not eliminate the need for complementary benchmarking metrics: direct error, reference-based, weighted, and retrieval-based metrics each capture distinct aspects of model behavior and should be combined in holistic model evaluations. For instance, direct error metrics are a useful guardrail against re-

warding implausible model predictions, as recently demonstrated in the Virtual Cell

Challenge [7]. While both MAE and MSE are valuable metrics here, we recommend MSE as it showed better calibration overall (Fig 2E, S5). In practice, a model performing significantly below baseline on either metric would warrant scrutiny regardless as this may indicate a lack of modeling realism.

Major issue number two, misrepresentation of the chronological reality

We thank the reviewer for highlighting the importance of clearly distinguishing the effects of metric calibration from advances in model development, and for emphasizing recent progress in perturbation modeling.

We have revised the manuscript to explicitly separate these two factors throughout. In the Results, we now clarify that earlier models (e.g., scGPT and GEARS) already outperform uninformative baselines when evaluated with well-calibrated metrics, and that their apparent underperformance under conventional metrics reflects metric miscalibration rather than a lack of learnable signal. We further note that more recent models (including PRESAGE, scLambda, and CellFlow) represent advances in model design and are generally capable of outperforming baselines, even under poorly calibrated metrics.

In the main text, we have revised language that previously generalized across models to preserve chronological context. We explicitly acknowledge that strong performance by newer models reflects both improved evaluation practices and progress in model architectures, rather than omissions or errors in earlier benchmarking studies.

These changes retain our central conclusion (that metric calibration is essential for evaluating perturbation-model performance) while more accurately reflecting the rapid evolution and current state of the field.

We will respond to specific examples below:

This manuscript makes strong claims,

e.g.

“In the Replogle K562 benchmarking dataset, similar to others (2, 3, 5, 6, 8), we find that models generally do not outperform the mean and linear baselines when evaluated with poorly calibrated metrics such as MSE and Pearson(Δctrl) (Fig. S9A,B; table 1)”

while actually looking at Figure S9A,B shows that around half of the DL models, and consistently including the newer models scLAMBDA and PRESAGE and often also fMLP variants outperform the mean baseline on these ‘poorly calibrated’ metrics. PRESAGE and scLAMBDA also have a clear edge on the linear baseline. This was not reported in previous benchmarking studies, for the simple reason that these models are newer than the studies in question. Presumably, the benchmarking studies and the awareness they brought to some of the challenges of the data and the easy ways to fool ourselves in modelling, had an influence on the development of these models and encouraged the developers there of aim higher

We agree with the reviewer, and we have revised this section to distinguish between earlier and more recent models (Page 3, Line 183):

In Replogle22 K562, earlier models (scGPT and GEARS) do not outperform baselines under poorly calibrated metrics such as MSE and Pearson(Δctrl) (Fig. S6A,B; Table 1), consistent with prior benchmarks [8, 2, 9, 10, 3]. Under well-calibrated metrics, however, even these earlier models mostly outperform uninformative baselines (Fig. 3B,C; Fig. S6C), suggesting that their perturbation-specific signal was present but obscured by metric choice. More recent models (e.g., PRESAGE and scLambda) go further, often outperforming uninformative baselines even under poorly calibrated metrics; indeed, on every metric at least one deep learning model outperforms all baselines (Table 1, S1, S2). These findings were also replicated in the independent Nadig25 HepG2 dataset [11] (Fig. S6 and Tables S1-2).

Another example is where the authors write “In the combination-gene task, some models likewise exceeded null baselines and (especially for PRESAGE) even surpassed the additive baseline in the Wessels23 dataset.”

without providing inline (although the quoted sentence is already in the discussion) mention that PRESAGE is newer than the benchmarking studies that found additive was the best combinatorical model ‘yet’.

We agree completely with the reviewer and now add the following to the main text (Page 4, Line 242):

and PRESAGE exceeded it on 15 of 18 metrics (Table S1,S2). Notably, PRESAGE is a more recent architecture which was developed after the benchmarking studies that established the additive baseline as a robust predictor compared to other models, exemplifying improving trends in model performance over time in this task.

Once the authors do mention that the consistently performant models PRESAGE and scLAMBDA are newer, they add that they should be included in future benchmarking studies, as if this was an omission of the benchmarking studies, and not a reality of the chronological order of when work was performed and released.

We completely agree and removed our prior assertion about the inclusion of PRESAGE/scLambda in future benchmarking studies. Instead, we include these statements in the main text (Page 5, Line 256):

...as well as newer models spanning diverse architectures—including PRESAGE (attention-based, prior knowledge embeddings), scLambda (VAE), and CellFlow (flow matching)—which stand out for their stronger performance relative to uninformative baselines. Earlier benchmarking studies played a constructive role in surfacing modeling limitations
that these and other architectural advances have begun to address.

Along with the following (Page 6, Line 327):

Finally, as the field continues to evolve rapidly, conclusions drawn from any fixed set of models necessarily reflect a snapshot of perturbation-response modeling and available perturbation datasets at a particular time.

So yes, more DL models outperform baselines on more calibrated metrics, but the most prominent result is that PRESAGE is consistently a top performer (mostly the top, multiple times above the ‘positive control’); and it is above baselines regardless of metric calibration.

We agree that PRESAGE is a standout model in our evaluation and have highlighted this in a new sentence in the main text (Page 5, Line 263):

In our results, PRESAGE is notable both for its strong performance and sophisticated perturbation representation approach based on biological prior knowledge.

A DL model has access to all the same information that a simple baseline does; and should be able to (and clearly can!) learn to outperform it or match it globally. The older models do better on the metrics here, but clearly also have areas for improvement that the benchmarking studies helped surface. The results of benchmarking studies were imperfect, yes , as this whole field is learning and shifting rapidly, but they helped spur progress and a new wave of legitimately across-the-board competitive models. Quite frankly, the focus on ‘older models would have ranked somewhat better compared to baselines if benchmarked exactly as so’ takes away from the progress and excitement of the field and the results of this paper as well.

The wording should be adjusted throughout to bring more acknowledgement to the global progress the field has made; progress that the authors are part of !

As a start, adding ‘now’ to the title would help.

We thank the reviewer for emphasizing the importance of accurately reflecting the substantial progress the field has made in perturbation-response modeling. We agree that recent benchmarking studies have played a constructive role in surfacing limitations and motivating both improved evaluation practices and the development of stronger model architectures, and we have revised the manuscript throughout to better acknowledge this progress.

Specifically, we now distinguish more clearly between effects due to metric calibration and genuine advances in model design, noting that newer models (e.g., PRESAGE and scLambda) represent across-the-board improvements and can outperform uninformative baselines even under conventional, poorly calibrated metrics.

We provide examples of this in our above responses. Below is a recap of relevant quotes from the

new main text.

First (Page 3, Line 183):

In Replogle22 K562, earlier models (scGPT and GEARS) do not outperform baselines under poorly calibrated metrics such as MSE and Pearson(Δctrl) (Fig. S6A,B; Table 1), consistent with prior benchmarks [8, 2, 9, 10, 3]. Under well-calibrated metrics, however, even these earlier models mostly outperform uninformative baselines (Fig. 3B,C; Fig. S6C), suggesting that their perturbation-specific signal was present but obscured by metric choice. More recent models (e.g., PRESAGE and scLambda) go further, often outperforming uninformative baselines even under poorly calibrated metrics; indeed, on every metric at least one deep learning model outperforms all baselines (Table 1, S1, S2).

Second (Page 5, Line 251):

More broadly, across both single and combination prediction tasks, models frequently outperform uninformative baselines under well-calibrated metrics. This includes earlier models such as scGPT and GEARS, whose perturbation-specific signal was present but obscured by metric choice (Fig. S6), as well as newer models spanning diverse architectures—including PRESAGE (attention-based, prior knowledge embeddings), scLambda (VAE), and CellFlow (flow matching)—which stand out for their stronger performance relative to uninformative baselines in these metrics. Earlier benchmarking studies played a constructive role in surfacing modeling limitations that these and other architectural advances have begun to address.

Third (Page 6, Line 327):

Finally, as the field continues to evolve rapidly, conclusions drawn from any fixed set of models necessarily reflect a snapshot of perturbation-response modeling and available perturbation datasets at a particular time. Benchmarks should incorporate both earlier and up-to-date architectures to track whether advances in modeling translate to increased performance in well-calibrated benchmarks.

At the same time, we preserve our central finding that earlier models already outperform uninformative baselines when evaluated with well-calibrated metrics, and that metric choice can substantially obscure this signal. For this reason, while we appreciate the reviewer's suggestion to add "now" to the title, we have opted not to do so, as it could be read as implying that baseline outperformance only emerged with newer models, which would be inconsistent with our findings of earlier models outperforming baselines in better-calibrated metrics. We believe the revised framing more accurately captures both the field's rapid progress and the complementary roles of improved benchmarking and improved modeling.

Medium issue number one, batch effects

These experiments have substantial batch effect, yet the authors chose not to compare to within batch controls for delta metrics (at least I could not find any within-batch matching searching 'batch' and

'gem' (for gem group)). Given there is confounding between batches and perturbations that allows unmatched metrics to confuse pertubation and batch effect, and that not batch matching inflates batch noise relative to perturbation effect, I would appreciate it if the authors either a) indicate they perform this and make sure to say it, b) add delta metrics with batch matching, or c) provide a thorough explanation of why not. I do appreciate not batch matching is extremely common in this field, but in a paper otherwise making some very solid contributions to benchmarking best practice, it seems like a major omission.

We thank the reviewer for raising this important point. We agree that batch effects are an important technical factor in perturb-seq datasets and scRNA-seq more broadly. In this study, we did not apply explicit batch correction or matching for metrics as we did not find batch effects sufficiently strong as to require correction or special handling. Our analysis of batch effect is below:

For the four primary benchmarking datasets emphasized in the Results (Replogle22 K562, Nadig25 HepG2, Norman19, and Wessels23), we calculated PCA, nearest neighbors, louvain clustering, and UMAP. We provide a figure below showing UMAP embeddings with Louvain clustering for each dataset, colored by batch label (left) and cluster label (right). Across all datasets, batch

labels do not align with cluster structure, indicating that batch effects do not dominate the dataset variation. Quantitatively, agreement between batch labels and cluster assignments is minimal, with low adjusted Rand index (ARI) and normalized mutual information (NMI) (ARI ≤ 0.022, NMI ≤ 0.099 across datasets). These results suggest that batch effects are weak and do not dominate the variance in the datasets analyzed here.

While this analysis does not rule out more subtle batch effects within smaller gene expression subsets, attempting to remove these effects could also remove perturbation-specific signals that coincide with batch labels, and, alternatively, batch-matched metrics would complicate comparisons across datasets and metrics—particularly because the number of cells per perturbation within each batch can vary widely, and the subsetting necessary to calculate metrics at the batch level would increase measurement noise and decrease metric calibration as shown in the cell downsampling

analysis on page 6. For this reason, and in keeping with common practice in the field, we elected not to batch correct or do batch matching in our main analysis.

Medium issue number two, WMSE weights and model ordering

I notice that the metric WSME has a notably different sort order of models than most other metrics, with especially the rank of scGPT rising. The idea to weight by comparison to the whole dataset, and not just controls to circumvent systematic bias is appreciated, and I'd love to have a better understanding of (and in the future use) this metric. If the authors know why and what the strength of scGPT (or error of other models) is, it could improve understanding and trust of this metric. Additionally, I'd like to know the typical number of genes receiving a meaningful amount of weight.

We thank the reviewer for this insightful question regarding the behavior and interpretability of WMSE.

In our prior work [5], we showed that the WMSE weighting strategy improved the sensitivity of MSE to small but significant perturbation effects. Moreover, the utility of this weighting approach was supported by our results showing that WMSE as a training objective can reduce mode collapse and improve model performance overall (addressing an issue we noted in GEARS, in particular). In order to help increase confidence in WMSE as a metric, we now discuss these prior findings directly in our revised main text (Page 2, Line 125):

We previously proposed weighted MSE (WMSE) [5] for exactly this purpose: WMSE upweights errors on genes with stronger perturbation effects, even when they comprise a small fraction of the transcriptome—and, when used as a training objective, WMSE can increase perturbation model output variance and improve performance relative to standard MSE.

Moreover, in this work we show that WMSE weights are not just a useful heuristic to increase metric calibration, but they also map onto insights from orthogonal ChIP-seq experiments (Fig S4C-D). In our supplementary analysis using ChEA transcription factor binding sites, we demonstrate that the DEG test statistics (which WMSE weights are a direct transformation of) correctly enrich for TF binding sites corresponding to the correct perturbation label—and that this occurs even in perturbations which are otherwise very poorly calibrated by MSE. This indicates that WMSE preferentially emphasizes genes that encode genuine, perturbation-specific signals.

The reordering of models under WMSE may reflect this property. Models such as scGPT tend to place modest but coherent signal on the correct subset of genes—potentially due to foundation-scale pretraining which could encode gene-gene relationships. However, without a separate ablation study, it may be difficult to pinpoint the precise modeling decisions which lead to any particular model performing well on WMSE (or any other metric). We hope our calibration framework will prove helpful in supporting future efforts to measure the relative impact of various architectures, datasets, and hyperparameters on performance.

In response to the reviewer's request, we also provide a figure above visualizing the gene-wise weight distributions used in WMSE across all four benchmarking datasets for representative perturbations. These plots show that a relatively small subset of genes receives the majority of the weight. While

the exact number of heavily weighted genes varies by perturbation, it is generally on the order of tens to a few hundred genes, consistent with the sparsity of perturbation effects observed in differential expression analyses.

Together, these results show that WMSE ranks models based on their ability to reconstruct gene expression, particularly of meaningfully-perturbed genes—and that our weighting strategy corresponds to perturbation-specific signals that are validated in orthogonal TF binding analyses. We view WMSE as a complementary metric adding substantial value to our larger benchmarking suite—particularly in settings where niche perturbation effects are otherwise obscured by unweighted error metrics.

Minor comments

in introducing the technical duplicate the authors explain it as “if we repeated the experiment with no batch effects, how predictive would our new experiment be?” but it is not batch effects coming into play here, it is variance between individual cells (both technical and biological cell to cell, not batch to batch, noise)

We thank the reviewer for this important clarification and agree that our original wording overstated the role of batch effects in the interpretation of the technical duplicate baseline.

We have revised the manuscript to clarify that the technical duplicate does not represent a hypothetical experiment without batch effects, but rather captures variance arising from cell-to-cell technical noise, biological heterogeneity, and sampling variability within the same biological condition. Accordingly, we now describe the technical duplicate as estimating the reproducibility of perturbation effects across independently sampled cells from the same perturbed population.

Specifically, we now state the following in the “Methods” subsection “Technical duplicate” (Page 8, Line 481):

In other words, it answers the question: if we were to sequence additional cells from the same perturbed biological samples (as in a technical replicate) how predictive would

those new cells be of the perturbation effect observed in the original cells?

We believe this revised framing more accurately reflects the statistical and biological meaning of the technical duplicate and avoids conflating cell-level variability with batch-to-batch effects.

Reviewer #2

This manuscript tackles a contentious question in single-cell perturbation modeling: whether deep learning perturbation-response models truly add value compared to simpler baselines. The first half of the manuscript explores the extent to which frequently used benchmarking metrics are able to distinguish between an appropriately chosen positive control and negative control, using a newly introduced meta-metric (Dynamic Range Fraction (DRF)). Across multiple datasets/metrics, they argue that DEG-aware and rank-based metrics (e.g., WMSE, weighted R^2 variants, NIR) are better calibrated than commonly used MSE and Pearson correlation of differentially expressed genes. In the second half of the manuscript, the authors test 8 different model configurations on unseen genetic perturbation and unseen combination prediction tasks, showing that for certain well-calibrated metrics certain deep learning methods do outperform linear baselines.

We thank the reviewer for their thoughtful summary of the manuscript and for the constructive feedback provided throughout. Below, we address each comment in detail and describe the revisions and additional analyses that have been incorporated to strengthen the manuscript.

Overall comments

Given the sparse and noisy nature of biological cell states and single-cell RNA-seq measurements, the introduction of a positive control is an important and welcome contribution. However, because the technical duplicate relies on splitting cells within each perturbation, its performance may be sensitive to the number of cells available per perturbation. While the authors acknowledge this issue (line 468), the manuscript would be strengthened by more explicitly examining how cell number per perturbation influences the interpretation of results and the robustness of the proposed conclusions. Concretely, stratifying perturbations by cell-count bins (or overlaying cell counts) and showing how DRF or control separation changes with cell number would clarify when the positive control is reliable and when it may be noisy or misleading.

We thank the reviewer for this thoughtful suggestion and agree that the reliability of the positive control may depend on the number of cells available per perturbation.

We note that in our preprocessing pipeline, we imposed both minimum and maximum cell thresholds per perturbation to reduce extreme imbalance (minimum of 12 cells; maximum capped at the dataset average), thereby limiting the possibility that a small number of highly sampled perturbations dominate calibration estimates. In addition, Figure SN2 in Supplementary Note 1 illustrates how technical duplicate performance on MSE improves with increasing cell count, consistent with improved signal estimation at higher sampling depths.

To more directly address the reviewer's suggestion, we have now added a controlled downsampling analysis (see page 6), which shows that cell count impacts calibration but that better-calibrated metrics like WMSE and NIR maintain higher DRF values at low cell counts than common metrics like MSE and Pearson Δctrl .

One possible extension would be to assess the robustness of the conclusions under a different positive control definition. For example, Jiang et al. (Nature Cell Biology, 2025 https://www.nature.com/articles/s41556-025-01622-z) introduced Mixscale scores to quantify the extent to which individual cells respond to a perturbation. Intuitively, well-behaved evaluation metrics should assign higher scores to predictions corresponding to cells with higher Mixscale responses.

The E-distance metric introduced for single-cell data by Peidli et al. (Nature Methods, 2024 https://www.nature.com/articles/s41592-023-02144-y) could also be considered, either as a metric for comparing predicted and ground-truth gene expression or as a way to quantify perturbation effect size. In the latter case, perturbations with larger E-distance would be expected to be easier to predict, and this information could be leveraged to assess metric validity (e.g., models should perform worse for perturbations with lower E-distance, as demonstrated in related work by Wenteler et al). This could serve as an additional sanity check on whether the proposed "well-calibrated" metrics behave monotonically with perturbation strength, and whether DRF conclusions generalize beyond the specific positive control construction.

We thank the reviewer for this thoughtful suggestion. We have now extended our analysis to assess robustness of our calibration approach to alternative control definitions and different dataset factors like cell count (see page 2).

In agreement with the reviewer's perspective, we also examine here the relationship between perturbation strength and metric behavior—which is described in the new "Perturbation strength impact on calibrated metrics" section in Supplementary Note 2 and new Figure SN8 (below). Perturbation strength was quantified both by the number of DEGs and by E-distance (only calculated on 300 random perturbations due to quadratic computational complexity). Consistent with our calibration framework, separation between informative baselines and the uninformative mean baseline increased with effect size in all cases. Moreover, for retrieval- and fit-based metrics (NIR and weighted RΔ2),

positive controls generally exhibited improved performance for stronger perturbations, as expected when larger transcriptional shifts provide clearer signal. In contrast, direct reconstruction metrics such as WMSE showed weaker, possibly inverted (or non-monotonic trends), similar to prior observations for MSE (as in Fig 3a of Wenteler et al. [10]). This may reflect the increased difficulty of accurately reconstructing the larger distributional shifts in expression space that are induced by stronger perturbations. Together, these analyses support the view that well-calibrated metrics respond coherently to perturbation signal, while highlighting differences across metric families.

This analysis is now described in Supplementary Note 2 ("Reliability of calibration approach") subsection "Perturbation strength impact on calibrated metrics":

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 [12]. Consistent with our earlier results, the separation between informative and uninformative baselines increased with effect size across metric families (Fig SN8). As an additional check, we evaluated trends in positive control performance with increasing strength. For retrieval- and fit-based metrics (NIR and weighted RΔ 2), 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 [10], likely reflecting the increased difficulty of reconstructing larger distributional shifts. Together, these results support the view that well-calibrated metrics respond appropriately to perturbation signal while highlighting differences across objective types.

Regarding the suggestion to incorporate Mixscale scores: we appreciate the importance of alternative baseline definitions and now address the broader question of robustness to control choice in the analysis on page 4. However, Mixscale quantifies perturbation response at the individual cell level, whereas our benchmarking framework—along with multiple models and all baselines evaluated here—operates on perturbation-level pseudobulk expression profiles. While one could hypothetically bin cells by Mixscale score and pseudobulk the strongest responders as an alternative positive control, this introduces additional complexity: if ground truth remains the pseudobulk of all non-baseline cells, the positive control would systematically overestimate the perturbation effect rather than providing a better estimate of it. Constructing a matched ground truth within Mixscale bins would address this but also fragment the already limited population of cells per perturbation—and this type of cell-wise downsampling increases noise and reduces calibration, as shown on page

6. Moreover, Mixscale was developed and validated for CRISPRi Perturb-seq data and has not, to our knowledge, been applied across the other perturbation modalities in our benchmark (e.g., CRISPRa, CaRPool-seq), further complicating adoption here.

We believe the perturbation strength analyses presented above along with our prior analysis of alternate control definitions address the spirit of the reviewer's—namely, whether well-calibrated metrics behave coherently with perturbation strength and whether findings are robust to alternate control definitions—while maintaining methodological consistency with the pseudobulk evaluation framework.

A recent landmark study in single-cell perturbation prediction is the Arc Institute's State model ([https://www.biorxiv.org/content/10.1101/2025.06.26.661135v2.full.pdf](https://www.biorxiv.org/content/10.110), which was reported to outperform linear baselines. Including elements of this work could strengthen the present manuscript. In particular, the authors could consider (i) evaluating additional metrics introduced in the State study (e.g., p-value AUPRC, DEG overlap accuracy, and effect size prediction), and/or (ii) including State in the foundation-model benchmarking analysis. This would be especially relevant given that a central claim of the paper is that reports of deep learning models failing to outperform linear baselines largely stem from metric choice (as reflected in the title), whereas the State model reports improved performance even under commonly used metrics such as Pearson correlation.

We thank the reviewer for highlighting the Arc Institute's STATE model, which represents an important recent contribution to single-cell perturbation modeling. We agree that engaging with this work strengthens the manuscript, particularly given its relevance to model–baseline comparisons.

Additional metrics from STATE. We have now incorporated the perturbation discrimination score (PDS) introduced in the STATE study into our evaluation framework (see revised Methods and Results). Included here is the results for the PDS in the Replogle22 K562 unseen genetic perturbation task (below). We found that model rankings are similar to other well-calibrated metrics, particularly NIR which is a related retrieval-based metric.

We also now include MAE, which was employed in the STATE paper and the Virtual Cell Challenge following it, in addition to MSE.

We appreciate the suggestion to include additional metrics such as p-value AUPRC and DEG overlap accuracy. However, these metrics require single-cell predictions, and, as discussed above, the focus of our study requires metrics are calculated at the pseudobulk level—both because of the necessity of comparing with pseudobulk baselines like the mean baseline, but additionally because multiple models in our study (e.g., PRESAGE) only provide pseudobulk-level predictions. We now clarify this explicitly in the Methods section (Page 10, Line 547):

All evaluation metrics were computed on perturbation-level pseudobulk expression pro-

files, ensuring comparability across models and baselines that do not generate single-cell predictions.

Attempted inclusion of STATE in unseen genetic perturbation and combo tasks. We have also explored employing the STATE architecture for the unseen genetic perturbation and unseen-combination prediction tasks evaluated in our study. However, to our understanding and at the time of writing, the publicly available implementation is designed only for context generalization and does not have an official implementation yet for the unseen genetic perturbation and unseen-combo tasks. Therefore, we had previously attempted to implement a version of STATE-TX that would be suitable for both tasks. For example, we tried using ESM2 embeddings to represent perturbations in a learnable space for the unseen genetic perturbation task (similar to CellFlow),

rather than one-hot encoding which is the STATE-TX default. Despite our best optimization efforts (including >1M step training), we were unable to achieve stable or competitive performance in either the unseen genetic perturbation or unseen combo task. For transparency, we provide representative results from one attempt (Q3 2025) on the Replogle22 RPE1 unseen genetic perturbation task (below). In most metrics, our implementation of STATE-TX underperformed other models and most baselines.

Given these challenges, and because the architecture has not yet been described for these specific tasks, we felt that including a partially adapted implementation risked misrepresenting the model's intended use. We therefore chose not to include STATE in the final benchmarking comparison.

Table 1: Representative benchmarking results from an attempted adaptation of the STATE model to the Replogle22 RPE1 unseen genetic perturbation prediction task. Results shown for one train/validation/test split.

Model MSE ↓ WMSE ↓ Pearson(Δ) ↑ Pearson(Δ DEGs) ↑ 2 ↑
RΔ 2 ↑
Weighted RΔ
Control baseline 0.0109 0.0595 -0.0123 -0.2001 -0.3177 -0.2741
Split mean baseline 0.0086 0.0452 nan nan -0.0100 -0.0245
Linear baseline 0.0084 0.0460 0.1279 0.3255 -0.0050 -0.0087
Technical duplicate (2nd half) 0.0082 0.0103 0.5033 0.9671 0.0011 0.7536
scGPT (MSE-only) 0.0093 0.0473 0.0834 0.1864 -0.0508 -0.0335
GEARS (MSE-only) 0.0085 0.0431 0.1411 0.3712 -0.0165 0.0016
scLambda 0.0077 0.0344 0.3424 0.7120 0.0473 0.0899
STATE (tx unseen) 0.0812 0.1039 0.0806 0.3303 -8.2324 -1.1778

Clarification regarding metric choice and model performance. Importantly, the reviewer raises a valuable point regarding our central claim about metric choice. We have refined the manuscript to clarify that while metric miscalibration explains much of the previously reported underperformance of deep learning models, some newer architectures (notably PRESAGE and scLambda) can outperform linear and mean baselines even under poorly-calibrated metrics such as MSE and control-referenced Pearson correlation. However, these improvements are less consistently detectable under poorly calibrated metrics due to limited metric sensitivity to robust predictors.

We now explicitly state in the main text (Page 3, Line 190):

More recent models (PRESAGE and scLambda) go further, often outperforming uninformative baselines even under poorly calibrated metrics...

This clarification ensures that our conclusions are not framed as contradicting the STATE findings. In particular, STATE demonstrates strong performance relative to simple baselines in the context generalization task in their paper, whereas our analysis focuses on unseen genetic perturbation and unseen combo tasks. Our findings therefore complement rather than contradict the STATE results, and we look forward to incorporating STATE in future benchmarks when it is adapted for the unseen single perturbation and unseen combo tasks.

Taken together, we believe these additions and clarifications strengthen the manuscript and better situate our findings within the wider perturbation modeling landscape.

Relatedly, the manuscript would benefit from clearer positioning of scope and novelty relative to

the rapidly developing SOTA landscape in perturbation prediction. Several recent methods report improvements over scGPT and/or linear baselines using distributional and generative modeling approaches (e.g., flow/diffusion paradigms) (e.g., https://arxiv.org/html/2511.02986v1, https://www.biorxiv.org/cont https://openreview.net/forum?id=QSGanMEcUV). Since the manuscript's narrative emphasizes metric choice as a key explanation for discrepant findings about whether deep learning adds value, the conclusions would be strengthened by explicitly discussing how these methods fit into the proposed framework (and ideally, by including at least one representative example in the benchmark). Otherwise, it remains somewhat unclear whether the reported "DL vs linear" conclusions generalize beyond the specific set of model classes evaluated here.

We thank the reviewer for this valuable suggestion. We agree that, given the rapid development of perturbation-response modeling (including recent flow- and diffusion-based generative approaches) it is essential to clarify how our conclusions relate to this evolving landscape.

To address this concern, we made three revisions:

(1) Inclusion of a representative flow-based generative model. We have added CellFlow to the benchmark. In our evaluation, CellFlow generally outperforms uninformative baselines under well-calibrated metrics and, in some settings, ranks among the top-performing models (example from Wessels23 combo prediction shown below—new Fig 3E). This addition ensures that flow-based generative models are explicitly represented in our analysis.

(2) Clearer positioning of architectural diversity. We reorganized the model descriptions in the Methods section ("Benchmarked deep learning models") to emphasize that the benchmark spans multiple architectural families (graph neural networks, transformer-based foundation models, variational autoencoders, flow-matching, and MLP probe models) as well as diverse perturbation representations (including LLM-derived embeddings, ESM2-derived embeddings, etc). This restructuring clarifies that our conclusions are not tied to one set of earlier architectures or specific modeling decisions.

(3) Expanded discussion of scope and generalization. We revised the main text to explicitly state that our central finding holds across multiple architectural paradigms and perturbation representations. Specifically, the passage summarizing results across both tasks now reads (Page 5, Line 251):

More broadly, across both single and combination prediction tasks, models frequently outperform uninformative baselines under well-calibrated metrics. This includes earlier models such as scGPT and GEARS, whose perturbation-specific signal was present but obscured by metric choice (Fig. S6), as well as newer models spanning diverse architectures—including PRESAGE (attention-based, prior knowledge embeddings), scLambda (VAE), and CellFlow (flow matching)—which stand out for their strong performance relative to uninformative baselines in these metrics. Earlier benchmarking studies played a constructive role in surfacing modeling limitations that these and other architectural advances have begun to address.

Overall, by incorporating a representative flow-based model and more clearly describing the diversity of architectures and perturbation representations included in our benchmarking, we believe the manuscript now better establishes that our conclusions are not restricted to specific modeling choices. Rather, we emphasize that advances in modeling (at least partially driven by prior benchmarking efforts) have led to approaches that achieve robust performance, which becomes especially evident when evaluated with well-calibrated metrics.

While this point is briefly acknowledged in the Discussion (Line 359), statistically significant improvements do not necessarily imply biological usefulness. The manuscript would be strengthened by more explicitly connecting evaluation metrics to concrete biological use cases. For example, which metrics are most appropriate for tasks such as DEG recovery, global transcriptome accuracy, or perturbation retrieval, and do deep learning models provide clear advantages in these settings? Including at least one downstream analysis demonstrating that "well-calibrated" metrics align with biological utility (such as improved pathway or TF target recovery, or biologically meaningful perturbation similarity networks) would help ground the benchmarking results in practical biological relevance.

We thank the reviewer for this insightful suggestion. In response, we expanded the manuscript to more explicitly connect evaluation metrics to downstream biological use cases and to evaluate whether well-calibrated metrics align with biological utility.

First, we restructured the Methods section to organize metrics by biological objective (with subsections “Direct error metrics”, “Reference-based delta metrics”, “Retrieval-based metrics”, and “DEG-based metrics”). We now explicitly clarify which evaluation strategies are relevant for distinct downstream use cases, including broader gene expression reconstruction (MSE/MAE), accurate recovery of perturbation-induced transcriptional shifts (delta-based metrics), pathway- and gene module-focused analysis (weighted and DEG-restricted metrics), and perturbation similarity or gene regulatory network (GRN) analyses (retrieval-based metrics such as NIR/PDS).

For example from “Retrieval-based metrics” in Methods (Page 11, Line 585):

Retrieval-based metrics evaluate whether models preserve the relational structure of the perturbation landscape. This is particularly relevant for applications such as network-based analyses, where recovering biologically meaningful relationships among perturbations may be more important than minimizing per-gene reconstruction error.

Second, as you suggested, we introduced new downstream evaluation analyses (see new Methods sections “Pathway recovery analysis” and “Neighborhood recovery analysis”): (i) Pathway Recovery (PR) measures the concordance between predicted and ground-truth gene set enrichment results in perturbation effects using MSigDB Hallmark pathways, and (ii) Neighborhood Recovery (NR) evaluates preservation of the perturbation-perturbation similarity network via k-nearest-neighbor graph overlap between ground truth and predicted perturbation effects. These analyses evaluate whether models which succeed in gene-level evaluation metrics also preserve pathway-level biological responses and perturbation network geometry.

Notably, models that ranked highly under well-calibrated benchmarking metrics (particularly PRESAGE, with strong performance from scLambda, CellFlow, and fMLP-GenePT) also achieved the highest pathway recovery and neighborhood recovery scores. And, in general, we found that model rankings on the most well-calibrated metrics and those in the PR and NR scores were correlated. We report our findings in new Figure S9 and in the main text (Page 5, Line 274):

Importantly, improvements in benchmarking metrics do not necessarily imply usefulness in downstream tasks such as pathway-level interpretation, so we also evaluated whether model performance corresponds to recovery of biological insights through an analysis of two downstream tasks: (i) pathway recovery (PR)—correlation between GSEA calculated on predicted and ground-truth perturbation effects—which is relevant to pathway analysis from model predictions and (ii) neighborhood-structure recovery (NR)—Jaccard similarity between KNN graphs constructed from predicted and observed perturbation shifts—which is relevant to perturbation network analysis (details in Methods). Across datasets, models performing strongly under well-calibrated metrics also showed better pathway recovery (Fig. S9A–D), with PR rankings and well-calibrated metric rankings being correlated (Replogle22 K562 $\rho_{avg} = 0.80$; Wessels23 $\rho_{avg} = 0.87$). Neighborhood recovery showed a similar but more variable pattern (Replogle22 K562 $\rho_{avg} = 0.76$; Wessels23 $\rho_{avg} = 0.51$), with the strongest correspondence between NR and retrieval-based NIR/PDS metrics ($\rho > 0.69$ in all cases), consistent with their related objectives (Fig. S9E–H). These results indicate that well-calibrated metrics capture biologically meaningful differences between models relevant to downstream analysis objectives.

Together, these additions address the downstream biological utility of our work by (i) mapping metric classes to concrete biological use cases and (ii) demonstrating empirically that model performance in well-calibrated metrics corresponds with improved biological recovery in downstream analyses.

Specific comments:

Line 118–120: the use of “genes” here is confusing as the authors handle both gene perturbation labels and the genes whose expression is being measured; consider rephrasing the second clause (and avoiding anthropomorphisms).

We thank the reviewer for pointing out this confusing word choice. We have now corrected the

manuscript text and figures so that “genes” refers to genes which are expressed (and not to perturbations), and “perturbations” (along with “unseen genetic perturbation” and “unseen combinations”) refers to genetic perturbations and related tasks. We have also removed the anthropomorphism implied by “favored”.

Figure S1E: It is interesting that the mean baseline outperforms the technical duplicate for many perturbations, particularly those with few DEGs. Could the authors assess whether this pattern varies with the number of cells per perturbation, for example by overlaying cell counts in a plot like Fig. S1E? More generally, how does the usefulness of the DRF depend on the number of cells per perturbation?

The reviewer brings up an important dataset characteristic which impacts baseline performance and metric calibration. We discuss this issue in depth in response to their prior comment regarding the effect of cell number (page 6). In short, the technical duplicate and interpolated duplicate improve in performance with increasing numbers of cells, along with metric calibration overall, likely owing to the improved power to detect perturbation effects with increased sampling of perturbed populations.

Figure S7A: it could make interpretation easier to more clearly label that higher scores are better with DRF (as shown in the schematic in the main figure).

We have now updated the x-axis label to read “DRF (↑)” and added in the figure caption “higher is better”.

Figure 2D,E: why does DRF show such a range across datasets? This doesn’t seem to be due only to variation in the number of DEGs. It would help to discuss additional drivers (e.g., cell-count distributions per perturbation, sparsity, heterogeneity, batch effects, library size differences), or to provide a decomposition/diagnostic correlating DRF with these factors.

We thank the reviewer for this insightful comment. Indeed, the number of DEGs is not the sole determinant of DRF. Rather, DRF reflects the separation between positive and negative controls, which depends on multiple properties of the dataset that affect signal-to-noise ratio.

We therefore explored several additional drivers of calibration variability—cell counts per perturbation, UMI counts per cells, and number of selected HVGs—which are discussed in detail beginning on page 6. In addition, we discuss the impact of perturbation strength on page 23, and describe our analysis of batch effects on page 18.

In all, we find that our conclusions are robust to dataset factors, and our analysis reveals the benefits of metrics that are well-calibrated overall (e.g., WMSE, NIR) compared to common metrics (e.g., MSE, Pearson Δ_{ctrl}).

Figure S9B: how did control mean perform on this task?

We appreciate the opportunity to clarify our baseline reporting strategy. As the metric in question Pearson Δ_{ctrl} is undefined for the control cell baseline (the prediction vector is all zeros), we chose to omit that baseline here.

We have now clarified our handling of this issue in the Methods section (Page 10, Line 562):

For Pearson Δ, special care is required when the reference baseline coincides with the predictor being evaluated, as the metric is then undefined by construction (e.g., Pearson Δ_{ctrl} for the control baseline, in which case the prediction vector is identically zero). For the purpose of DRF calibration heatmaps, when a Pearson Δ value would otherwise be undefined, we computed DRF using an alternative negative baseline. For example,

DRF for Pearson Δ_{pert} was computed using the control cell baseline as the negative control, rather than the mean baseline, which would yield an undefined reference. This substitution was justified by the observation that identification of well-calibrated metrics from DRF was similar between negative control types (...). In contrast, for model benchmarking analyses, predictors or baselines for which a metric was undefined were omitted from the corresponding plots and summaries to avoid ambiguity in interpretation.

Figure S9: minor: the color and style of the line for Mean in the legend box does not match the figure.

We thank the reviewer for pointing this out. We have now removed the legend, as we felt it didn't add any necessary information, and combined this figure with original figure S10 to create new figure S7.

Lines 220–225: scGPT is the only model evaluated both as a predictive model and as a source of embeddings with an MLP probing head. Could the authors clarify the rationale for including both evaluations only for scGPT (and whether similar probing should be applied consistently to other foundation embeddings/models)?

We agree that this is an important point to clarify, and thank the reviewer for the opportunity to strengthen our manuscript by improving our description of how we handled different models and model-derived perturbation embeddings.

In addition to scGPT, we use GenePT (LLM text embeddings of NCBI gene descriptions) for fMLP probing, which was also used as the feature representation for scLambda. Additionally, ESM2 embeddings are the input for both CellFlow and fMLP-ESM2. ESM2, GenePT, and Geneformer do not provide a mechanism for predicting perturbation responses in the benchmarked unseen genetic perturbation and unseen combo tasks like scGPT does—thus it would not have been possible to measure their performance on a similar footing. Part of our goal with the fMLP model was to evaluate the effects of different embedding strategies when the underlying architecture was not varied—a disentanglement which is not possible by comparing models like CellFlow, scLambda, and scGPT directly.

We have updated the methods to describe this rationale in better detail (under "Probing foundation model-based perturbation embeddings (fMLP)") (Page 12, Line 646):

In addition to open-source models, we evaluated the utility of foundation model-based perturbation representations using an MLP probing approach. This allowed us to test the relative impact of variations in perturbation representation on model performance, independently of complex architectural design, further broadening the generality of our findings.

. . . Description of fMLP approach . . .

By spanning a variety of perturbation representations and model architectures, this benchmark tests whether deep learning-based perturbation models outperform trivial baselines beyond specific modeling choices. Moreover, as the foundation models from which we collected embeddings do not (with the exception of scGPT) provide the capacity to predict unseen genetic perturbations and unseen combo perturbations, our fMLP probing strategy was necessary to evaluate the utility of their perturbation representations.

Line 475: Why are 8192 random cells selected? Similarly, line 518, why are 8192 highly variable genes selected? Please justify or show sensitivity/robustness to these choices.

We are grateful for the opportunity to clarify these choices. Regarding control cell number: First, all datasets have at least 8192 control cells, meaning that this many cells could be consistently selected. Moreover, based on our prior work, we found control signal saturates by 4096 control cells (see Mejia et al. [5] figure 5G-I). Thus, conservatively, we selected double that amount. Regarding highly variable genes, we found that several of the datasets had been pre-filtered to as low as 12k genes. To ensure we would have consistency, even in the case of adding new datasets later, we selected 8192 genes to analyze—thus balancing the desire to evaluate models and metrics in a wider transcriptomic space while making our approach applicable across datasets.

Moreover, we consider the impact of dataset and preprocessing factors like HVG selection and number of cells to be important in contextualizing the calibration methodology we are proposing, and we have now evaluated the robustness of our calibration methodology to the number of HVGs and number of cells selected, as discussed in our prior responses (pages 6 and 8).

Line 493: The interpolated duplicate uses adjusted DEG p-values to define the interpolation weight $\alpha = 1 - \text{padj}$. Because p-values depend on both effect size and statistical power, and are not directly proportional to biological effect size, this modeling choice may influence the behavior of the positive control in ways that are not immediately transparent. In order to evaluate the impact of this choice, it would be helpful to plot the observed distribution of (or p-values) across perturbations, or to assess

whether alternative definitions based on effect size or ranked significance yield similar conclusions.

We thank the reviewer for this thoughtful comment and agree that the interpretation of the interpolation α parameter merits clarification.

First, we have included here representative examples of α across perturbation strengths in Norman19:

To clarify our choice of α:

In our framework, the interpolated duplicate is not intended to make gene-wise predictions that consider perturbation effect size, but rather to serve as positive control with greater consistency and robustness to signal dilution than the technical duplicate.

Using α = 1 – padj provides a natural, bounded interpolation parameter in [0, 1] that can be

interpreted as a confidence-like measure of whether a gene is perturbed, incorporating both effect size and variance. Genes with strong statistical evidence of perturbation are therefore weighted toward the technical duplicate, while genes lacking such evidence are weighted toward the mean baseline. This avoids the need for additional ad hoc scaling or normalization that would be required for unbounded distributions such as test statistics or fold-change estimates.

Importantly, the interpolated duplicate is used to provide a stable positive control in calibration measurement, and our conclusions are largely robust to alternative positive and negative control definitions, as shown in our new control-sensitivity analysis (on page 4)

We have now added a brief clarification to the Methods section (”Interpolated duplicate”) to make the rationale for this interpolation scheme explicit (Page 8, Line 489):

...where α is a vector encoding DEG significance and ⊙ denotes the element-wise product operation between vectors. We define α = 1 – pDEGs per perturbation with the help of the p values of an independent DEG analysis performed only on the held-out technical duplicate set. This choice yields interpolation weights naturally bounded in [0, 1] and provides an interpretable confidence-like measure of whether a gene is perturbed. In other words, α reflects the strength of statistical evidence that a gene is affected by a given perturbation.

We have also updated the caption in Figure 2 to read:

(B) The interpolated duplicate baseline, which blends the technical duplicate and mean baseline using a weight α = 1 – pDEGs from DEG analysis (adjusted p value). The more significantly-perturbed a gene is, the more interpolation is weighted toward the technical duplicate, while less significant or unperturbed genes are weighted toward the mean baseline.

We have also made minor edits to other sections describing our interpolation procedure to increase clarity on this point.

Line 521: does the splitting happen on the test set? If yes, could this be made more explicit?

We thank the reviewer for this question and agree that the ordering of operations was not sufficiently explicit in our original description. The ground-truth (GT) / technical-duplicate (TD) split is performed once globally for all perturbations prior to cross-validation and is fixed across folds. Subsequent train/validation/test splits are carried out at the perturbation-label level.

For perturbations assigned to the training set, both GT and TD cells are available during model fitting. For perturbations assigned to the test set, models never observe either GT or TD cells during training. During evaluation, only the GT half of test perturbations is used as ground truth, while the TD half is used exclusively for baseline construction (e.g., technical and interpolated duplicates).

We have revised the Methods section to clarify this ordering. The section you referenced now reads (Page 9, Line 519):

Next, to construct our technical duplicate ST D and ground truth SGT sets, we perform

a single random cell-wise split for every perturbation in the dataset prior to cross-validation. This split is fixed across folds. Subsequent train/validation/test splits are performed at the perturbation-label level. For perturbations assigned to the training set, both SGT and ST D cells are available during model fitting (i.e., no cell-level withholding occurs within training perturbations). For perturbations assigned to the test set, models never observe either SGT or ST D cells during training. During evaluation, only the ground-truth half SGT of test perturbations is used as the true target, while ST D is used exclusively for baseline construction (e.g., technical and interpolated duplicates). This design ensures that the technical/interpolated duplicate task of predicting one half of the data from the other half is mirrored in model evaluation without introducing data leakage across cross-validation folds.

Throughout the manuscript, the authors refer to Pearson control as “control-referenced deltas.” However, in prior studies this metric is often computed either over all differentially expressed genes or over the top-ranked DEGs. Please clarify how this metric is implemented in the present study, in particular the choice of gene set used for its calculation, as referenced in line 548.

We agree with the reviewer that this could have been clearer. We have now more clearly separated DEG-based metrics from non-DEG metrics. New Figure 2E (heatmap showing calibration across datasets in non-DEG metrics) and S5 (heatmap showing same but in DEG-based metrics) have been formulated to make it clearer when DEG filtering is being applied. These changes and other updates to our DEG-based metric handling are detailed in our introduction within this response (page 2).

We also explicitly describe this in the methods section as well, with a new section describing DEG-based metrics (Page 11, Line 588):

For applications focused on specific genes, gene modules, and pathways, model perfor-

mance can be evaluated on the top N differentially expressed genes (DEGs) per perturbation [4, 3, 2]. Unlike weighted metrics, which apply continuous gene-level weights across all genes in the dataset, DEG-based evaluation performs explicit gene filtering prior to metric calculation.

In this study, we evaluated performance on the top 100 DEGs per perturbation (ranked by adjusted p-value). This provides a consistent evaluation space size across perturbations and datasets while focusing analysis on niche, biologically-informative perturbation effects. Within this subset, we computed unweighted error metrics (MSE, MAE) and delta-based metrics (Pearson Δ and $R\Delta^2$), thereby assessing model accuracy with

respect to perturbation-specific biological signal, an approach relevant to gene module– and pathway-level applications.

Because DEG-restricted metrics are computed over a subset of genes with a different feature space relative to the entire dataset, DEG-based metrics were analyzed and displayed separately when assessing calibration via the Dynamic Range Fraction (DRF).

Line 568: the MLP architecture for MLP probing is quite large; what is the impact of this decision? (e.g., sensitivity to probe capacity / risk that probe dominates differences between embeddings).

We thank the reviewer for this important point. To assess whether probe capacity influences comparative embedding performance, we conducted a size ablation study in which we reduced the fMLP architecture to approximately one-half and one-quarter of its original parameter count (512 hidden; 128 latent), training medium (256 hidden; 64 latent) and small (128 hidden; 32 latent) variants under otherwise identical settings. Across well-calibrated metrics (WMSE, weighted $R\Delta_{ctrl}$, NIR, and Pearson(Δ_{pert}) on DEGs), embedding rankings remained largely stable as probe capacity decreased (New Figure S7; included here). Although absolute performance varied, no embedding exhibited disproportionate degradation, and embedding rankings typically did not change. In particular, fMLP-GenePT and fMLP-ESM2 consistently outperformed fMLP-Geneformer and fMLP-scGPT across size variants. These results suggest that probe capacity is likely not be a strong driver of the observed differences and that performance primarily reflects the information content of the perturbation representations rather than the architecture of the downstream probe.

We now describe these results in the main text (Page 4, Line 197):

Using our fMLP probe, we also observed consistent differences in performance between perturbation embedding features, with GenePT and ESM2-based fMLP models proving more predictive overall. Importantly, these results were found to be consistent across a size ablation study in which we reduced fMLP model capacity to approximately one-half and one-quarter of the original parameter count and trained in the Replogle22 K562 (see Methods). Performance rankings across embeddings remained largely stable across all capacity levels in well-calibrated metrics (Fig. S7), indicating that differences reflect embedding information content rather than probe capacity.

Line 585: what is the justification for using a one-sided test?

In all cases, our statistical hypothesis is that model predictions will outperform the uninformative mean baseline—and thus a one-sided test best-reflects that intention. We now clarify this directly in the “Statistical tests” section of the methods (Page 13, Line 704):

For parametric comparisons, we used paired (dependent) Student’s t-tests, one-sided unless otherwise stated. A one-sided test was chosen because, for each predictor, the statistical hypothesis was that it would outperform the uninformative negative control

(either the mean baseline or control baseline, depending on the specific analysis).

Fig. S9A,B: Given the small absolute effect sizes, the error bars in the forest plots play an important role in interpreting the results. However, the description of the bootstrap procedure (line 596) is underspecified. Please clarify what is being resampled (e.g., perturbations vs. cells), whether the paired structure is preserved, and which statistic is recomputed.

We thank the reviewer for the opportunity to clarify our statistical approach in our forest plots. The mean value we wished to contextualize with 95% CI bars is the $\mu\Delta$, the average Δi (predictor metric score minus mean baseline score), representing the average difference of predictors compared to the mean baseline. Following standard practice for calculating 95% CI around the mean, we sampled the population of per-perturbation (paired) Δi with replacement (and sample size N being the total number of perturbations), recalculating $\mu\Delta$ from the resampled deltas. We repeated this 10,000 times for robustness, and then selected the 2.5% and 97.5% percentile values as our CI bounds.

We have now clarified this section of our methods to better explain our bootstrapping procedure (Page 13, Line 712):

For forest plots, we estimated 95% confidence intervals around $\mu\Delta$ via non-parametric percentile bootstrapping (using scipy.stats.bootstrap): the per-perturbation paired differences $\Delta i = s_{predictor} - s_{negative}$ (metric score difference between predictor and negative control baseline—typically the mean baseline) were resampled with replacement across scored perturbations (10,000 resamples; seed = 42; sample size N is total number of scored perturbations), recomputing the mean on each resample and taking the 2.5th and 97.5th percentiles as confidence bounds.

References

[1] Elijah Cole, Geert-Jan Huizing, Sohan Addagudi, Nicholas Ho, Euxhen Hasanaj, Merel Kuijs, Toby Johnstone, Maria Carilli, Alec Davi, Caleb Ellington, Christoph Feinauer, Pan Li, Romain Menegaux, Shahin Mohammadi, Yanjun Shao, Josiah Zhang, Emma Lundberg, Le Song, Ziv Bar-Joseph, and Eric P. Xing. Foundation models improve perturbation response prediction. bioRxiv, 2026. doi: 10.64898/2026.02.18.706454. URL https://www.biorxiv.org/content/early/2026/02/19/2026.02.18.706454.

[2] Ramon Vi˜ nas Torn´e, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbi´c. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. Nature Biotechnology, Aug 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02777-8. URL https://doi.org/10.1038/s41587-025-02777-8.

[3] Chen Li, Haoxiang Gao, Yuli She, Haiyang Bian, Qing Chen, Kai Liu, Lei Wei, and Xuegong

Zhang. Benchmarking ai models for in silico gene perturbation of cells. bioRxiv, pages 2024–12, 2024.

[4] Lanxiang Li, Yue You, Wenyu Liao, Xueying Fan, Shihong Lu, Ye Cao, Bo Li, Wenle Ren, Yunlin Fu, Jiaming Kong, et al. A systematic comparison of single-cell perturbation response prediction models. bioRxiv, pages 2024–12, 2024.

[5] Gabriel M. Mejia, Henry E. Miller, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Diversity by design: Addressing mode collapse improves scrna-seq perturbation modeling on well-calibrated metrics, 2025. URL https://arxiv.org/abs/2506.22641.

[6] Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state. bioRxiv, 2025. doi: 10.1101/2025.06.26.661135. URL https://www.biorxiv.org/content/early/2025/07/10/2025.06.26.661135.

[7] Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, Reshma Mehta, Benjamin Hsu, Jeremy Sullivan, Chiara Ricci-Tam, Nianzhen Li, Julia Kazaks, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Hani Goodarzi, and Dave P. Burke. Virtual cell challenge: Toward a turing test for the virtual cell. Cell, 188(13):3370–3374, Jun 2025. ISSN 0092-8674. doi: 10.1016/j.cell.2025.06.008. URL https://doi.org/10.1016/j.cell.2025.06.008.

[8] Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods, 22(8):1657–1661, Aug 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02772-6. URL https://doi.org/10.1038/s41592-025-02772-6.

[9] Gerold Csendes, Gema Sanz, Krist´of Z Szalay, and Bence Szalai. Benchmarking foundation cell models for post-perturbation rna-seq prediction. BMC genomics, 26(1):393, 2025.

[10] Aaron Wenteler, Martina Occhetta, Nikhil Branson, Magdalena Huebner, Victor Curean, William Dee, William Connell, Alex Hawkins-Hooker, Pui Chung, Yasha Ektefaie, et al. Perteval-scfm: Benchmarking single-cell foundation models for perturbation effect prediction. bioRxiv, pages 2024–10, 2024.

[11] Ajay Nadig, Joseph M Replogle, Angela N Pogson, Mukundh Murthy, Steven A McCarroll, Jonathan S Weissman, Elise B Robinson, and Luke J O'Connor. Transcriptome-wide analysis of differential expression in perturbation atlases. Nature Genetics, pages 1–10, 2025.

[12] Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. Nature Methods, 21(3):531–540, 2024.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have addressed my concerns.

They have improved the manuscript, both in scientific content and communication style thus increasing the already substantial impact and contributions to the field!

(Remarks on code availability)
I did not check again in any detail in this round of review. I do see and appreciate that the repository has been active; with revisions added.

Reviewer #2

(Remarks to the Author)

The authors have made a considerable effort to address the concerns raised in the first round, and I acknowledge that several additions — the robustness analyses, the downstream biological utility analyses, and the inclusion of CellFlow — represent genuine improvements. However, upon careful examination of the revised manuscript and its supplementary materials, I have identified a set of concerns that are more serious than those I raised in the first round. These concerns go to the integrity of the central empirical claims and, in my view, preclude acceptance in the current form. I detail them below in order of severity.

Major concerns
1. The manuscript's central claim is undermined by selective metric reporting that systematically flatters older, weaker models
This manuscript's stated contribution is to provide a fairer, more rigorous framework for evaluating perturbation response models. It is therefore particularly troubling that the primary benchmarking figure — Figure 3 — relies on a metric selection that, whether intentionally or not, presents a systematically distorted picture of model performance.
The authors justify the use of WMSE and weighted $R^2(\Delta ctrl)$ as the primary reporting metrics in Figure 3 on the grounds that these are the best-calibrated metrics according to their own DRF framework. However, the consequence of this selection is not neutral: it substantially inflates the apparent performance of scGPT relative to a more complete metric view. In Replogle22 K562, scGPT achieves Pearson(Δpert) on DEGs of 0.268 ± 0.009 in the main Table 1 — a figure that receives prominent placement and supports the manuscript's narrative that earlier models outperform baselines under well-calibrated evaluation. Yet in Supplementary Table S1, scGPT's Pearson(Δpert) evaluated over all genes rather than filtered to the top DEGs drops to 0.081 ± 0.003, barely above the mean baseline and far below more recent models. This is not a minor numerical discrepancy — it is a qualitative difference in the conclusion one would draw about whether scGPT carries meaningful perturbation-specific signal.
This discrepancy is not discussed anywhere in the main text. The authors have constructed a framework that identifies well-calibrated metrics, applied that framework selectively to support a narrative of older model rehabilitation, and then declined to report the metrics that would complicate that narrative. A paper that argues for benchmarking transparency while practising selective reporting is internally contradictory in a way that cannot be resolved through revision of framing alone.
The authors must either include Pearson(Δpert) on all genes in Figure 3 and discuss the implications directly, or provide a principled, pre-specified justification for its exclusion that does not amount to post-hoc metric cherry-picking. The current presentation does not meet the standards the paper itself sets.
2. CellFlow is almost certainly misconfigured on the primary benchmarking dataset, rendering the benchmark unreliable for the model class it was specifically included to represent
The inclusion of CellFlow was presented as a direct response to my earlier concern about the absence of flow-based generative models. I therefore examined its results carefully. In Replogle22 K562 — the largest and most prominently featured dataset in the benchmark — CellFlow achieves weighted $R^2(\Delta ctrl) = 0.009 \pm 0.004$, indistinguishable from the uninformative control mean baseline. On Wessels23, the same model achieves weighted $R^2(\Delta ctrl) = 0.154 \pm 0.020$, outperforming most other models. This is not a modest dataset-dependent variation — it is an order-of-magnitude discrepancy that demands explanation.
The explanation is not difficult to find. The benchmark applies 50 input principal components uniformly across all datasets, as confirmed by inspection of the publicly available benchmarking code on GitHub. CellFlow's own documentation and tutorials explicitly state that this hyperparameter should be adapted to dataset scale and complexity, and provide guidance for doing so. Replogle22 K562 spans over 2,000 perturbations across approximately 237,000 cells — by a substantial margin the largest and most complex dataset in the benchmark. Applying a configuration designed for small datasets to this setting, without any dataset-specific tuning, is a methodological error, not a valid test of the model's capabilities.
The consequences for the manuscript are significant. CellFlow was included specifically to demonstrate that the paper's conclusions generalise beyond older transformer- and GNN-based architectures to more recent generative approaches. If CellFlow's near-baseline performance on the primary dataset is an artefact of misconfiguration, this representational goal has not been achieved. The authors cannot claim that their benchmarking conclusions generalise to flow-based generative models on the basis of results that likely reflect a systematic configuration error. This experiment must be repeated with hyperparameters appropriate to the scale of the dataset before these results can be reported.
3. The linear baseline is deliberately weak and the reported model-versus-baseline gap is therefore not meaningful
The linear baseline, as described in Methods §A.5, uses principal components of the pseudobulked training expression matrix as perturbation features. This representation encodes statistical variation in the observed expression space and carries no molecular information about the identity or function of the perturbed gene. It is a deliberately uninformative perturbation representation, and the authors are aware of this — the fMLP probe results reported in the same manuscript

demonstrate that replacing PCA features with ESM2 or GenePT embeddings substantially improves performance even when the underlying model architecture is held fixed.
Recent work by Cole et al. (2026), which the authors themselves cite, has systematically demonstrated that language-model-derived perturbation representations carry substantially richer perturbation-informative signal than PCA-derived features and that this translates directly into improved prediction accuracy across benchmarks. Against this backdrop, the continued use of a PCA-based linear baseline as the primary lower bound is not a neutral methodological choice — it is one that makes the gap between the linear baseline and embedding-equipped deep learning models appear larger than it would under a fair comparison. A linear model equipped with ESM2 or GenePT perturbation embeddings, with all other choices held fixed, would provide a substantially more informative and defensible lower bound. Without this experiment, it remains genuinely unclear how much of the reported deep learning advantage reflects architectural capacity and how much reflects the linear model being given a categorically inferior input representation. This ambiguity is fatal to the manuscript's core comparative claim.

Residual minor concerns from the first round
Several of my earlier minor concerns remain only partially addressed.
On one-sided statistical testing: the authors justify the directional hypothesis by noting that they expect models to outperform the baseline. But systematic underperformance relative to baseline is itself a scientifically meaningful result in this context — indeed it is precisely the finding that prior benchmarking studies reported and that this paper seeks to recontextualise. Using a one-sided test that is structurally incapable of detecting underperformance is therefore not appropriate here. A two-sided test, or at minimum an explicit robustness check, is required.
On WMSE and perturbation strength (Fig. SN8): the authors' conclusion that "well-calibrated metrics respond appropriately to perturbation signal" is directly contradicted by the presented results for WMSE, which shows weak or non-monotonic trends with perturbation strength. Claiming a unified conclusion across metric families when one of the three highlighted metrics behaves differently is an overstatement that should be corrected.

On neighbourhood recovery (Fig. S9E–H): the use of Jaccard similarity as the overlap measure remains unjustified. Jaccard ignores ranking and relative distances within the KNN graph. The authors have not responded to this point.

On pathway recovery scope: restricting the analysis to approximately 50 MSigDB Hallmark gene sets makes it relatively easy to preserve concordance between predicted and ground-truth enrichment profiles. This is not acknowledged as a limitation, and the analysis has not been extended to more granular gene set collections such as GO Biological Process, as suggested in the first round. Given that pathway recovery is presented as a key validation of the manuscript's downstream biological utility claims, this limitation is non-trivial.

(Remarks on code availability)

Author Rebuttal letter:

We thank the reviewers for their continued engagement with this work and we thank the editor for the opportunity to respond to Reviewer #2's new comments. Below we address each point in detail. In summary, we have:

1. Added Pearson(Δpert) on all genes to Figure 3 and Figure S6, and added a candid discussion of inter-metric variability for individual models to the main text, including the specific scGPT example the reviewer highlights.
2. Performed the CellFlow hyperparameter sweep the reviewer requested. Across {50, 100, 200, 400, 800} input principal components on Replogle22 K562, 50 PCs was optimal in 18/18 benchmarking metrics, confirming that our prior CellFlow results reflect optimized performance rather than misconfiguration with respect to the PC parameter.
3. Built fLM-ESM2, a linear model using ESM2 embeddings in place of PCA features, in direct response to the reviewer's request. We also clarify that the existing PCA-based linear baseline is the one introduced by Ahlmann-Eltze et al (the study we directly respond to in our paper) and not a baseline of our own construction.
4. Switched all hypothesis tests to two-sided.
5. Clarified the point about the relationship between WMSE and perturbation strength.
6. Replaced Jaccard with Rank-Biased Overlap for neighborhood recovery and switched pathway recovery from MSigDB Hallmark to GO Biological Process.

None of these changes alter the conclusions of the manuscript. We address each of the reviewer's concerns in detail below.

We provide our revised manuscript in both PDF and LATEX format. The source .zip includes main.tex (clean revised manuscript) and diff.tex (tracked-changes version generated via latexdiff against the original submission)—and we provide rendered PDFs of both versions. All page and

line references in this response (e.g., Line 251) refer to the clean PDF.

Reviewer #2's comments are in red. Our response in black.

Major concerns:

The manuscript's central claim is undermined by selective metric reporting that systematically flatters older, weaker models This manuscript's stated contribution is to provide a fairer, more rigorous framework for evaluating perturbation response models. It is therefore particularly troubling that the primary benchmarking figure — Figure 3 — relies on a metric selection that, whether intentionally or not, presents a systematically distorted picture of model performance. The authors justify the use of WMSE and weighted $R^2(\Delta ctrl)$ as the primary reporting metrics in Figure 3 on the grounds that these are the best-calibrated metrics according to their own DRF framework. However, the consequence of this selection is not neutral: it substantially inflates the apparent performance of scGPT relative to a more complete metric view. In Replogle22 K562, scGPT achieves Pearson(Δpert) on DEGs of 0.268 ± 0.009 in the main Table 1 — a figure that receives prominent placement and supports the manuscript's narrative that earlier models outperform baselines under well-calibrated evaluation. Yet in Supplementary Table S1, scGPT's Pearson(Δpert) evaluated over all genes rather than filtered to the top DEGs drops to 0.081 ± 0.003, barely above the mean baseline and far below more recent models. This is not a minor numerical discrepancy — it is a qualitative difference in the conclusion one would draw about whether scGPT carries meaningful perturbation-specific signal.
This discrepancy is not discussed anywhere in the main text. The authors have constructed a framework that identifies well-calibrated metrics, applied that framework selectively to support a narrative of older model rehabilitation, and then declined to report the metrics that would complicate that narrative. A paper that argues for benchmarking transparency while practising selective reporting is internally contradictory in a way that cannot be resolved through revision of framing alone. The authors must either include Pearson(Δpert) on all genes in Figure 3 and discuss the implications directly, or provide a principled, re-specified justification for its exclusion that does not amount to post-hoc metric cherry-picking. The current presentation does not meet the standards the paper itself sets.

Two threads run through this critique: a structural concern (that our highlighting WMSE and weighted $R^2(\Delta ctrl)$ reflects 'selective metric reporting') and a revision request (that readers see a fuller view of cross-metric model behavior). We address each.

On the structural concern. The DRF framework was applied uniformly to all 18 metrics. DRF scores derive from the performance gap between positive- and negative-control baselines, which are quantities computed independently of any model's predictions. Metric calibration is therefore determined before model evaluation, not after. The metrics reported in Figure 3 (WMSE and weighted $R^2(\Delta ctrl)$) are the transcriptome-wide metrics which displayed the highest calibration across datasets. Had WMAE or Pearson(Δpert) emerged as the best-calibrated metrics in our analysis, they would have been our headline metrics instead, regardless of which models they favored. Thus, our framework has not been applied 'selectively.'

On the revision request. We agree that fuller cross-metric reporting helps readers, and have made three changes:

1. As the reviewer requested, Pearson(Δpert) on all genes is now reported in Figure 3 and Figure S6 alongside WMSE and weighted $R^2(\Delta ctrl)$.
2. A new paragraph in the main text discusses cross-metric variation explicitly (beginning on Line 267: "Our results also show that models exhibit substantial variation across metric families..."). It calls out scGPT directly: best WMSE in Replogle22 K562 (0.027 ± 0.001) yet poor Pearson(Δpert) across all genes (0.081 ± 0.003), barely above the mean baseline. We contrast this with recent architectures such as PRESAGE that are more consistent across metric families.
3. If the editor judges it appropriate for the brief communication format, we are happy to expand Table 1 to include the full metric set currently in Tables S1 and S2, rather than just the well-calibrated metrics.

New Figure 3:
CellFlow is almost certainly misconfigured on the primary benchmarking dataset, rendering the benchmark unreliable for the model class it was specifically included to represent.

The inclusion of CellFlow was presented as a direct response to my earlier concern about the absence of flow-based generative models. I therefore examined its results carefully. In Replogle22 K562 — the largest and most prominently featured dataset in the benchmark — CellFlow achieves weighted $R^2(\Delta ctrl) = 0.009 \pm 0.004$, indistinguishable from the uninformative

control mean baseline. On Wessels23, the same model achieves weighted $R^2(\Delta ctrl) = 0.154 \pm 0.020$, outperforming most other models. This is not a modest dataset-dependent variation — it is an order-of-magnitude discrepancy that demands explanation.

The explanation is not difficult to find. The benchmark applies 50 input principal components uniformly across all datasets, as confirmed by inspection of the publicly available benchmarking code on GitHub. CellFlow's own documentation and tutorials explicitly state that this hyperparameter should be adapted to dataset scale and complexity, and provide guidance for doing so. Replogle22 K562 spans over 2,000 perturbations across approximately 237,000 cells — by a substantial margin the largest and most complex dataset in the benchmark. Applying a configuration designed for small datasets to this setting, without any dataset-specific tuning, is a methodological error, not a valid test of the model's capabilities.
The consequences for the manuscript are significant. CellFlow was included specifically to demonstrate that the paper's conclusions generalise beyond older transformer- and GNN-based architectures to more recent generative approaches. If CellFlow's near-baseline performance on the primary dataset is an artefact of misconfiguration, this representational goal has not been achieved. The authors cannot claim that their benchmarking conclusions generalise to flow-based generative models on the basis of results that likely reflect a systematic configuration error. This experiment must be repeated with hyperparameters appropriate to the scale of the dataset before these results can be reported.

On CellFlow configuration. In response to the reviewer’s suggestion, we have now performed a sweep for optimizing the number of principal components. We varied the number of PCs across {50, 100, 200, 400, 800} on Replogle22 K562 (k-fold #1) and evaluated CellFlow on all 18 metrics. The original choice of 50 PCs was optimal in 18/18 metrics (four representative metrics shown below). Thus, our prior CellFlow results display optimized performance with respect to the number of PCs parameter.

On training steps. To further test whether CellFlow's performance might improve under additional training, we doubled the default training duration from 200k to 400k steps. CellFlow's performance degraded on nearly all metrics under this longer training regime, indicating that the original training duration is not a bottleneck and that additional compute does not recover stronger performance.

On the cross-dataset performance gap. CellFlow's lower performance on Replogle22 K562 relative to Wessels23 reflects task difficulty rather than model configuration. The Replogle22 K562 setting is more challenging: weaker perturbations on average and an unseen single-perturbation prediction task, versus the unseen combination task of Wessels23 (stronger perturbations, with all single perturbations visible during training). The same task/dataset-difficulty pattern is reflected in our positive-control baselines, as the technical and interpolated duplicates perform worse on Replogle22 K562 than on Wessels23 across all 12 comparable (non-error-based) metrics. Moreover, we see a similar pattern for several models in the benchmark, including PRESAGE, scLambda, and GEARS. We have added a brief discussion of this to the main text (section beginning on Line 251: "From our results, we observe that model performance varies substantially across datasets...").

On corroborating evidence from independent studies. Two recent independent studies also evaluate CellFlow and, similar to us, report similarly modest performance in the Replogle22 benchmark while also showing stronger performance in other task settings (e.g., Tahoe100M and PBMC benchmarks). Chen et al. (SCALE; arXiv:2603.17380, 2026) report CellFlow PDCorr of –0.003 and LFCSpear of –0.033 on Replogle, both at or below the mean baseline. Yuan et al. (PerturbDiff; arXiv:2602.19685, 2026) report the same values from their independent evaluation pipeline, with PDS scores at ~0.50 (random performance). Our own results for CellFlow on Replogle22 K562 are slightly more favorable than these independent reports but qualitatively consistent: CellFlow's performance on Replogle22 is near baseline on multiple metrics in independent evaluations, supporting the conclusion that this reflects genuine task difficulty rather than configuration error in our benchmark.

On the framing of "older architectures." We note that even setting CellFlow aside, the benchmark already includes other recent architectures: PRESAGE (2025, graph-attention-based prior-knowledge fusion) and scLambda (2024, LLM-featurized VAE). Thus our conclusion that both older and newer deep learning models can outperform uninformative baselines under well-calibrated metrics does not rest on CellFlow alone.

The linear baseline is deliberately weak and the reported model-versus-baseline gap is therefore not meaningful The linear baseline, as described in Methods §A.5, uses principal

components of the pseudobulked training expression matrix as perturbation features. This representation encodes statistical variation in the observed expression space and carries no molecular information about the identity or function of the perturbed gene. It is a deliberately uninformative perturbation representation, and the authors are aware of this — the fMLP probe results reported in the same manuscript demonstrate that replacing PCA features with ESM2 or GenePT embeddings substantially improves performance even when the underlying model architecture is held fixed.

Recent work by Cole et al. (2026), which the authors themselves cite, has systematically demonstrated that language-model-derived perturbation representations carry substantially richer perturbation-informative signal than PCA-derived features and that this translates directly into improved prediction accuracy across benchmarks. Against this backdrop, the continued use of a PCA-based linear baseline as the primary lower bound is not a neutral methodological choice — it is one that makes the gap between the linear baseline and embedding-equipped deep learning models appear larger than it would under a fair comparison. A linear model equipped with ESM2 or GenePT perturbation embeddings, with all other choices held fixed, would provide a substantially more informative and defensible lower bound. Without this experiment, it remains genuinely unclear how much of the reported deep learning advantage reflects architectural capacity and how much reflects the linear model being given a categorically inferior input representation. This ambiguity is fatal to the manuscript's core comparative claim.

On the choice of linear baseline. The PCA-based linear baseline is not our construction. Rather, it is the baseline introduced by Ahlmann-Eltze et al. in "Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines" (Nature Methods, 2025). In that work, the linear baseline was not ‘deliberately weak.’ Rather, it outperformed trained deep learning models on metrics like MSE, and was a major pillar in that paper's central claim. Because our brief communication directly responds to Ahlmann-Eltze et al., we use the same baseline set (linear, additive, mean, control). Replacing the linear baseline with a foundation-model-equipped baseline would change the paper we are responding to.

On Cole et al. as a parallel. The reviewer cites Cole et al. as evidence that the linear baseline is inappropriately weak. However, Cole et al. performed an evaluation of multiple foundation-model embeddings decoded using either KNN regression (non-linear) or lasso-based linear models. Their findings support rather than undermine our framing: linear models decoding simple representations form a trivial baseline class, while methods leveraging rich embeddings derived from foundation models and prior knowledge form a separate, more capable modeling class. The Ahlmann-Eltze et al. linear baseline belongs to the former, the Cole et al. foundation-based models to the later.

On the requested experiment. To directly address the reviewer's question, we built a linear model using ESM2 embeddings in place of PCA features ("fLM-ESM2"). fLM-ESM2 performs similarly to the PCA-based linear baseline on most metrics, with improvement on NIR. We are happy to include this result if desired, though we note that it does not change any conclusion of the manuscript.
On the fMLP comparison. To clarify one point: the reviewer notes that "the fMLP probe results... demonstrate that replacing PCA features with ESM2 or GenePT embeddings substantially improves performance even when the underlying model architecture is held fixed." The fMLP is a non-linear deep learning model, and not a simple linear baseline. The fMLP results compare embedding choices (not including PCA) within a non-linear architecture; they do not test whether PCA vs. ESM2/GenePT inputs would change the linear baseline's behavior. The fLM-ESM2 experiment above is the relevant test of that question, and the results (similar performance, with metric-specific exceptions) do not support the conclusion that PCA is substantially weaker than ESM2 in providing linear predictor features.

Minor points:

On one-sided statistical testing: the authors justify the directional hypothesis by noting that they expect models to outperform the baseline. But systematic underperformance relative to baseline is itself a scientifically meaningful result in this context — indeed it is precisely the finding that prior benchmarking studies reported and that this paper seeks to recontextualise. Using a one-sided test that is structurally incapable of detecting underperformance is therefore not appropriate here. A two-sided test, or at minimum an explicit robustness check, is required.

In response to the reviewer’s request, we have regenerated all hypothesis tests using two-sided t-tests. In the process we noticed that the forest plots had already been computed with two-sided tests, so those results are unchanged. The updated figures are Fig. S1A–D and Fig. S3B. None of the conclusions of the manuscript are affected.
On WMSE and perturbation strength (Fig. SN8): the authors' conclusion that "well-calibrated

metrics respond appropriately to perturbation signal" is directly contradicted by the presented results for WMSE, which shows weak or non-monotonic trends with perturbation strength. Claiming a unified conclusion across metric families when one of the three highlighted metrics behaves differently is an overstatement that should be corrected.

The reviewer is right that we overgeneralized across metric families. We meant only that each metric responded in a manner consistent with the mechanism it is built on, but the original wording was ambiguous. We have rewritten the relevant sentence to SN8:

“Together, these results indicate that retrieval- and fit-based well-calibrated metrics (NIR, weighted R2 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.”

On neighbourhood recovery (Fig. S9E–H): the use of Jaccard similarity as the overlap measure remains unjustified. Jaccard ignores ranking and relative distances within the KNN graph. The authors have not
responded to this point.

We agree that Jaccard ignores rank order and that this matters for k-nearest-neighbor graph comparisons. We have replaced Jaccard similarity with Rank-Biased Overlap (RBO), which weights agreement in higher-ranked (nearest) neighbors more heavily and respects the ordering induced by relative distances in delta space. Figure S9 has been updated along with the correlations reported in the section beginning on Line 303: “Importantly, improvements in benchmarking metrics do not necessarily imply...”. None of the conclusions of the manuscript are affected.

On pathway recovery scope: restricting the analysis to approximately 50 MSigDB Hallmark gene sets makes it relatively easy to preserve concordance between predicted and ground-truth enrichment profiles. This is not acknowledged as a limitation, and the analysis has not been extended to more granular gene set collections such as GO Biological Process, as suggested in the first round. Given that pathway recovery is presented as a key validation of the manuscript's downstream biological utility claims, this limitation is non-trivial.

We have switched the pathway recovery analysis from MSigDB Hallmark to Gene Ontology Biological Process. Figure S9 has been updated along with the correlations reported in the section beginning on Line 303: “Importantly, improvements in benchmarking metrics do not necessarily imply...” None of the conclusions of the manuscript are affected.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I appreciate the continued work of the authors in responding to review requests and strengthening their manuscript. I think the paper is again improved after the latest round of revision and is a strong contribution to the field, especially it is a key and needed perspective in the iterative give and take progress between modelling and benchmarking.

Please find thoughts on the latest modifications and if requested points were covered below:

> Regarding the concern about selective metric reporting and showing a fuller picture of metrics.

As I read it, the focus of this paper is on introducing DRF as an empirical method for testing metric calibration, and the importance of calibrated metrics. Given that, I both appreciate and find sufficient, the addition of Pearson(delta pert) to figure 3 and provided wording changes. This more clearly emphasizes the differences between metrics and models, while retaining focus.

> Regarding the ask to tune CellFlow

I do appreciate seeing the results of a recent distributional model in the independent evaluation, here. Similarly the ask to tune CellFlow is reasonable, as hyperparameter tuning indeed leads to far more conclusive comparisons when evaluating models. This tuning means the impact of model architecture/loss/task design–and not just optimization or familiarity level–is benchmarked. That said, doing so could hardly be called industry standard. GPU time costs money and energy and there is thus undoubtedly an onus on modellers to provide already optimized (and if needed dynamic) hyperparameter defaults

respecting how they would like their model to be benchmarked and used. I cannot confirm from cross referencing the CellFlow repository and documentation that I would find it clear which key parameters to focus on in tuning.

I appreciate and acknowledge that the authors have precisely fulfilled the request as stated and included larger PCA dimensions. They have addressed the concern that CellFlow showed lower performance on the Replogle dataset only due to too small of a PCA dimension and the authors have given a thorough explanation and contextualization for the difference in performance. I enjoyed the report and learned something while reading.

To nevertheless pre-emptively raise the obvious follow up. I do doubt CellFlow is optimally tuned in this configuration. When the smallest dimension performs best; it would be typical to also try decreasing the dimension (I can see from the CellFlow repository that 30 is also used). However, I would only ask the authors to do this if, on their own conscience, all or most of the other models received substantially more optimization and tuning than CellFlow did. As far as I can see from the text and reviews, this is not the case and default parameters were used throughout. Moreover, differences are small and would be highly unlikely to change conclusions.

Minor additional comment here (disregard if this relates only to the review and not manuscript test). PRESAGE uses rich context on genes, including context derived from knowledge graphs processed with node2vec, but it does not use graph attention.

> Regarding the concern that linear baseline is weak
In my opinion, a weak linear baseline is a feature not a bug; it is first and foremost a lower bound to indicate what should be unambiguously possible to achieve with a more complex model, and to keep ourselves honest. Since the authors have already implemented and calculated fLM-ESM2, I would include it as it disentangles prior knowledge from linear vs MLP and contributes to the comprehensiveness of the benchmark. (this opinion is lightly held)

> Regarding the minor concerns
These appear to be fairly addressed such that further optimization would not change key conclusions.

I continue to see this paper as an important contribution to the field and I trust the authors to address any outstanding minor concerns above without further review.

(Remarks on code availability)
I performed no additional review of the code in this review round.

Reviewer #2

(Remarks to the Author)
I thank the authors for the thorough and responsive revision. The new experiments fully address my concerns, and I appreciate the care taken in running them rather than reframing.

One framing point to consider. Given the cross-metric variability now documented in Figure 3, I would encourage tuning down language that singles out any one method as a "top performer in the text and abstract as well.
As the revised analysis shows, model ranking depends substantially on the metric family, and naming a specific method as best risks misrepresenting that nuance and being unfair to methods that lead on other axes. I would suggest framing the contribution as a general evaluation framework, with the empirical finding stated at the class level: that deep learning methods, taken as a group, can outperform existing baselines under well-calibrated metrics. This is both more defensible given the data and more useful to the community than a single-method ranking.

This is a strong paper, and with that adjustment, I expect it to be a useful and well-cited resource for the perturbation-modelling community. Congratulations to the authors.

(Remarks on code availability)

Author Rebuttal letter:

We thank the reviewers for their approval of this work and for their valuable feedback.

Reviewer comments are in red. Our response in black.

Reviewer #1:

To nevertheless pre-emptively raise the obvious follow up. I do doubt CellFlow is optimally tuned in this configuration. When the smallest dimension performs best; it would be typical to also try decreasing the dimension (I can see from the CellFlow repository that 30 is also used). However, I would only ask the authors to do this if, on their own conscience, all or most of the other models received substantially more optimization and tuning than CellFlow did. As far as I can see from the text and reviews, this is not the case and default parameters were used throughout. Moreover, differences are small and would be highly unlikely to change conclusions.

As the reviewer notes, we did not perform substantially more optimization or tuning with other models and so we agree additional attempts to tune CellFlow would not be warranted here.

Minor additional comment here (disregard if this relates only to the review and not manuscript text). PRESAGE uses rich context on genes, including context derived from knowledge graphs processed with node2vec, but it does not use graph attention.

Our comment about graph attention was related just to the review. In the text, we simply mention that PRESAGE uses the attention mechanism (as in global pooling) during prediction. To increase clarity in the text, we changed our description of PRESAGE as “a delta-predicting model with attention-based prior knowledge featurization” to “a delta-predicting model with prior knowledge-based featurization.”

> Regarding the concern that linear baseline is weak
In my opinion, a weak linear baseline is a feature not a bug; it is first and foremost a lower bound to indicate what should be unambiguously possible to achieve with a more complex model, and to keep ourselves honest. Since the authors have already implemented and calculated fLM-ESM2, I would include it as it disentangles prior knowledge from linear vs MLP and contributes to the comprehensiveness of the benchmark. (this opinion is lightly held)

We thank the reviewer for their support and flexibility on this request. Due to space limitations and our perception that the fLM-ESM2 results would not substantially improve this work, we have elected not to include fLM-ESM2 results.

Reviewer #2:

One framing point to consider. Given the cross-metric variability now documented in Figure 3, I would encourage tuning down language that singles out any one method as a "top performer in the text and abstract as well.
As the revised analysis shows, model ranking depends substantially on the metric family, and naming a specific method as best risks misrepresenting that nuance and being unfair to methods that lead on other axes. I would suggest framing the contribution as a general evaluation framework, with the empirical finding stated at the class level: that deep learning methods, taken as a group, can outperform existing baselines under well-calibrated metrics. This is both more defensible given the data and more useful to the community than a single-method ranking.

We thank the reviewer for this insightful suggestion. We were careful in the text not to single out a “top performer” as we agree with the reviewer that the primary purpose of this work is to highlight the importance of metric calibration and argue that, with calibrated metrics, deep learning perturbation models can outperform uninformative baselines. We feel that it would be challenging to make this point without showing the results of benchmarking on calibrated metrics. To decenter these results, we have removed Table 1 and the references to it in the text. We have also lightly edited the text where we felt we might be giving the impression of crowning a top-performer to clarify our intents.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/