**Reviewer #2:**

**Comment:** The manuscript lacks context regarding recent benchmarking studies (e.g., Wenteler et al. 2024; Wu et al. 2024) that report foundation models show no consistent improvement over simple baselines. The authors should clarify how their work differs from and extends these prior efforts.

**ANS:** We thank the reviewer for this suggestion. In the revised Introduction, we have added discussion and citations of recent benchmarks by Wenteler et al., Wu et al., Kedzierska et al., and Ahlmann et al.[[1]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Recent%20studies%20have%20provided%20initial,generalize%20across%20broader%20biological%20contexts). We explain that those studies provided valuable initial comparisons but were limited in dataset diversity and task coverage[[2]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=prediction%2C%20and%20reported%20that%20complex,generalize%20across%20broader%20biological%20contexts). Our work significantly expands on these by evaluating **25 datasets** across multiple tasks (including combinatorial perturbations and cell-type transfer), using a comprehensive suite of **24 evaluation metrics**. This broader scope allows us to observe patterns (e.g., the impact of perturbation effect size and cell-type differences) that were not addressed in earlier studies. In summary, we now clearly state how our benchmark builds on and goes beyond prior efforts, while noting that our results generally show foundation models performing better than simple baselines **overall**, with important caveats discussed in the manuscript.

**Comment:** The evaluation scope in the original submission seemed narrow (17 datasets, 9 metrics). Please expand the dataset diversity and include additional performance metrics to strengthen the benchmarking conclusions.

**ANS:** We agree and have substantially broadened the evaluation in the revised manuscript. Specifically, we increased our dataset collection from **17 to 25 datasets**, incorporating newly released resources (such as multi-cell-type data from the Virtual Cell Challenge) and two new in-house datasets[[3]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=immune%20populations%20%28e,type%20transfer). These additions cover a wider range of cell types and perturbation conditions, including primary immune cells and cross-species data. We have also expanded the set of evaluation metrics from **9 to 24**, adding measures of differential expression recovery and distributional similarity (Wasserstein distance, MMD) alongside the original error, goodness-of-fit, and correlation metrics[[4]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=We%20further%20classified%20each%20combination,Notably%2C%20buffering%20accuracy%20in%20seen0)[[5]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Foundation%20models%20exhibited%20variance%20compression,objectives%20in%20the%20foundation%20model). By adopting this **multi-metric** evaluation framework, we can better capture different aspects of model performance (e.g., ability to recover DE genes and capture variance in responses). These changes make our benchmarking more comprehensive, and the conclusions drawn are now supported by a much richer analysis of diverse scenarios.

**Comment:** Only 8–9 models were initially evaluated, and some important methods (e.g., scVIDR, PerturbNet or others like CellOT) were missing. The authors should justify their model selection and, if possible, include additional models or clarify exclusions.

**ANS:** We appreciate this concern and have updated the model selection accordingly. In the revision, we evaluate **12 single-cell perturbation prediction models** spanning four broad categories[[6]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=To%20evaluate%20the%20models%20in,given%20in%20Supplementary%20Tables%202). This includes models that were previously omitted: for example, we added **PerturbNet** (a conditional normalizing flow method) and **scVIDR** (a variational domain adaptation model), in addition to the originally tested models. We now provide a clear description of all model categories and their representative methods in the text and Supplementary Table 2[[6]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=To%20evaluate%20the%20models%20in,given%20in%20Supplementary%20Tables%202). We also clarify any exclusions: for instance, **CellOT** was not included due to implementation challenges and its narrower scope[[7]](file://file-NfGQeCRofQinhQWcsFawnw#:~:text=model60). By expanding and clarifying our model set, we ensure a fair and thorough comparison that addresses the reviewer’s concern.

**Comment:** The three prediction tasks (unseen perturbations, combinatorial perturbations, and cell-type transfer) were not clearly described in the original manuscript. Please clearly define each task and how the training/testing design differs. A schematic or figure would help.

**ANS:** We have revised the text to clearly delineate each task and added a schematic **Figure 1c** to illustrate the experimental design. In the **Methods/Results** section, we now explicitly define: (1) the *Unseen Perturbation* task, where models predict the effects of single-gene perturbations not seen in training; (2) the *Combinatorial Perturbation* task, focusing on predicting outcomes of gene combinations (with definitions of “seen0/seen1/seen2” depending on training exposure); and (3) the *Cell State Transfer* task, which tests generalization of perturbation responses to **unseen cell types**[[8]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=%5Citem%20%5Ctextbf,changes%20under%20previously%20unseen%20perturbations)[[9]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=line). Each task’s setup (training vs. testing data splits and goals) is now described in detail, and **Fig. 1c** visually outlines these scenarios for clarity[[10]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=reflect%20the%20varying%20complexity%20of,Using%20training%20data)[[11]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=%5Citem%20%5Ctextbf,cell%20type%20or%20cell%20line). These additions ensure that readers understand the differences among the tasks and how they reflect distinct aspects of model generalization.

**Comment:** The claim that foundation models “converge on population averages, struggling to capture heterogeneous responses” is interesting. However, the manuscript should provide evidence for this phenomenon and quantify it.

**ANS:** We have added a specific analysis to support this point. In the revised Results, we examine the **variance compression** effect in fine-tuned foundation models. We show that after fine-tuning, foundation models’ predictions have reduced variance, tending toward the dataset-wide mean response[[5]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Foundation%20models%20exhibited%20variance%20compression,objectives%20in%20the%20foundation%20model). This is evidenced by significantly higher Wasserstein distance and MMD values for foundation models compared to other model classes (indicating a poorer alignment with the true distribution of single-cell responses)[[5]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Foundation%20models%20exhibited%20variance%20compression,objectives%20in%20the%20foundation%20model). In practical terms, foundation models under-predict the heterogeneity of perturbation effects. We discuss this result in the context of training objectives: because preserving distribution variance was not explicitly enforced during fine-tuning, these models bias toward predicting population-average changes. This new analysis (see **Figure 2e**) quantitatively confirms our statement about foundation models and heterogeneous responses, as requested.

**Comment:** The manuscript should further analyze how perturbation effect size influences model performance. For example, are models performing differently on large-effect vs small-effect perturbations?

**ANS:** We agree that this is an important aspect, and we have included additional analysis on perturbation effect sizes. In the revised text, we report that **perturbation effect magnitude is a key driver of performance trends**[[12]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=modalities%2C%20and%20prediction%20tasks,is%20associated%20with%20reduced%20accuracy). Specifically, in tasks assessing **direct gene expression predictions**, we observe that smaller-effect perturbations yield higher correlation scores (since perturbed profiles remain closer to controls), whereas larger-effect perturbations are more challenging and lead to lower raw-score performance. Conversely, when evaluating relative changes (delta-based metrics), the trend reverses: **larger-effect perturbations yield better delta correlation and direction accuracy** because their signals rise above technical noise[[13]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=When%20moving%20to%20an%20analysis,for%20the%20models%20to%20capture). We have added text to highlight these patterns and included a panel (**Figure 2d**) illustrating that delta-Pearson correlation and direction accuracy increase with perturbation effect size. This analysis underscores how effect size impacts predictability, addressing the reviewer’s point.

**Comment:** For the combinatorial perturbation task, it would be informative to know whether models capture interaction effects (synergy or buffering) rather than just additive responses.

**ANS:** We have performed a new analysis of interaction effects in combinatorial perturbations. In the revised Results, we classify each two-gene perturbation as **synergistic** or **buffering** by comparing its observed effect to the sum of single-gene effects[[4]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=We%20further%20classified%20each%20combination,Notably%2C%20buffering%20accuracy%20in%20seen0). We then evaluated model performance on these categories. We found that models generally predict **buffering interactions** more accurately than synergistic ones[[14]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=%28Fig,also%20modeling%20their%20joint%20effects). For example, across unseen combinations (**“seen0”** cases), predictions were conservative, tending to underestimate extreme synergistic effects (leading to lower accuracy on synergies)[[15]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=%28Fig,also%20modeling%20their%20joint%20effects). This tendency is consistent with previous observations in the literature and is now described in the text with new results (see **Fig. 3c,d**). By analyzing interaction types, we provide deeper insight into how well models capture true perturbation interactions, as requested.

**Comment:** The performance in the cell-type transfer task (Task 3) seems relatively poor, even for foundation models. Could the authors comment on why foundation models, which are trained on many cell types, did not excel in cross-cell-type prediction?

**ANS:** We agree that this is an important point to discuss. In the revised Discussion, we note that **no model achieved strong performance in Task 3 (cell state transfer)**, including foundation models[[16]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=PerturbNet%2C%20which%20consistently%20achieved%20substantially,MMD%20added%20another%20dimension%20of)[[17]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=cross,training%20scFM%20on%20snapshot%20scRNAseq). We propose a possible explanation: although single-cell foundation models are pre-trained on millions of cells, their architectures may not capture the **cell-type-specific perturbation dynamics** required for transfer to unseen cell types[[18]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Single,help%20them%20understand%20perturbation%20dynamics). Moreover, pretraining on static transcriptomic profiles might not convey how perturbation responses translate across divergent cell types. We have added this reasoning to the discussion, suggesting that current foundation model training paradigms might not optimally address cross-cell-type perturbation generalization[[19]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=cross,heterogeneity%20of%20true%20perturbation%20responses). In support of this, we also highlight that datasets with larger transcriptional differences between source and target cell types showed notably worse transfer accuracy[[12]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=modalities%2C%20and%20prediction%20tasks,is%20associated%20with%20reduced%20accuracy), emphasizing the challenge of this task. We now explicitly acknowledge these limitations and the need for future improvements in modeling perturbation effects across cell types.

**Comment:** Lastly, the authors should highlight their main findings more clearly. What are the most significant insights from this benchmarking study?

**ANS:** We have revised the Abstract and Introduction to more clearly convey our key findings. We now emphasize two overarching insights up front[[20]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Motivated%20by%20these%20developments%2C%20we,is%20associated%20with%20reduced%20accuracy): (1) **Perturbation effect size is a dominant factor** influencing predictive performance – smaller perturbations are easier for models to predict in absolute terms, while larger perturbations are easier to capture in relative terms (but still hard to fully recover in terms of DE genes). (2) **Cell-type differences limit generalization** – in particular, when models are applied to a new cell type, a greater transcriptomic distance between training and target cell types is associated with reduced prediction accuracy[[21]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=biological%20and%20technical%20factors%20influencing,is%20associated%20with%20reduced%20accuracy). Additionally, throughout the Results we point out that while foundation models generally outperform simpler baselines on many metrics, **PerturbNet** stood out by best recovering differentially expressed genes in single and double perturbation tasks[[22]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Evaluation%20of%20DE%20gene%20overlap,4)[[23]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=At%20the%20delta%20and%20DE,3e), and no single model dominated the challenging cell transfer task. These points are now clearly stated in the manuscript so that readers can easily grasp the most significant conclusions of our study.

[[1]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Recent%20studies%20have%20provided%20initial,generalize%20across%20broader%20biological%20contexts) [[2]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=prediction%2C%20and%20reported%20that%20complex,generalize%20across%20broader%20biological%20contexts) [[3]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=immune%20populations%20%28e,type%20transfer) [[4]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=We%20further%20classified%20each%20combination,Notably%2C%20buffering%20accuracy%20in%20seen0) [[5]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Foundation%20models%20exhibited%20variance%20compression,objectives%20in%20the%20foundation%20model) [[6]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=To%20evaluate%20the%20models%20in,given%20in%20Supplementary%20Tables%202) [[8]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=%5Citem%20%5Ctextbf,changes%20under%20previously%20unseen%20perturbations) [[9]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=line) [[10]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=reflect%20the%20varying%20complexity%20of,Using%20training%20data) [[11]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=%5Citem%20%5Ctextbf,cell%20type%20or%20cell%20line) [[12]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=modalities%2C%20and%20prediction%20tasks,is%20associated%20with%20reduced%20accuracy) [[13]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=When%20moving%20to%20an%20analysis,for%20the%20models%20to%20capture) [[14]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=%28Fig,also%20modeling%20their%20joint%20effects) [[15]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=%28Fig,also%20modeling%20their%20joint%20effects) [[16]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=PerturbNet%2C%20which%20consistently%20achieved%20substantially,MMD%20added%20another%20dimension%20of) [[17]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=cross,training%20scFM%20on%20snapshot%20scRNAseq) [[18]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Single,help%20them%20understand%20perturbation%20dynamics) [[19]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=cross,heterogeneity%20of%20true%20perturbation%20responses) [[20]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Motivated%20by%20these%20developments%2C%20we,is%20associated%20with%20reduced%20accuracy) [[21]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=biological%20and%20technical%20factors%20influencing,is%20associated%20with%20reduced%20accuracy) [[22]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=Evaluation%20of%20DE%20gene%20overlap,4) [[23]](file://file-1A2NSD5s15HzHb83zLiXNo#:~:text=At%20the%20delta%20and%20DE,3e) revised\_manuscript.md

<file://file-1A2NSD5s15HzHb83zLiXNo>

[[7]](file://file-NfGQeCRofQinhQWcsFawnw#:~:text=model60) manuscript\_with\_figure\_and\_supfig.pdf

<file://file-NfGQeCRofQinhQWcsFawnw>