DETECTING CELL-LEVEL TRANSCRIPTOMIC CHANGES OF PERTURB-SEQ USING CONTRASTIVE FINE-TUNING OF SINGLE-CELL FOUNDATION MODELS

Wenmin Zhao, Ana Solaguren-Beascoa, Grant Neilson, Regina Reynolds, Louwai Muhammed, Liisi Laaniste, and Sera Aylin Cakiroglu
Cosyne Therapeutics
London, UK
{millie, ana, grant, louwai, liisi, aylin}@cosyne.com

ABSTRACT

Genome-scale perturbation cell atlases are an exciting new resource to understand the transcriptomic and phenotypic impact of single-gene activation or knockdown. However, in terms of differentially expressed genes identified, the signal detected in these data atlases is low, leading to the exclusion of most data from downstream analyses. Recent advances in single-cell foundation models have shown promise in capturing complex biological insights. However, their application to perturbation analysis, especially in predicting perturbed single-cell transcriptomes, remains limited. In this paper, we focus on learning representations of single-cell transcriptomes that capture subtle, yet important, transcriptome-wide changes, and we propose a novel fine-tuning strategy using contrastive learning to leverage single-cell foundation models for this task. We pre-train a single-cell foundation model and fine-tune on a genome-scale perturbation dataset using a contrastive loss, which minimises the distance between cell embeddings from unperturbed cells while maximising the distance between perturbed and unperturbed cells. We validate and test the model on unseen perturbations, demonstrating its ability to identify global biologically meaningful transcriptional changes not captured by traditional differential expression methods. Our approach provides a novel framework for analysing single-cell perturbation data and offers a more effective means of identifying perturbations that drive systemic gene expression changes.

1 INTRODUCTION

Understanding the transcriptomic and phenotypic outcomes of perturbed gene expression (e.g. knockdown or activation) at the single-cell level has the potential to improve our understanding of development, disease mechanisms, and cell-state engineering, with applications in target identification for drug discovery. High-throughput methods based on CRISPRi perturbations have recently been applied to this end at genome scale: Replogle et al. (2022) perturbed just under 10k expressed genes, resulting in a dataset of close to 2M single-cell transcriptomes and new preprints signpost genome-scale perturbation cell atlases as a new frontier of single-cell data (Nourreddine et al., 2024). Simultaneously, advanced deep-learning techniques have been applied in the prediction of perturbed single-cell transcriptome data (Roohani et al., 2023; Hao et al., 2024). Recent developments in training so-called 'single-cell foundation models' on large-scale single-cell transcriptomic data (Theodoris et al., 2023; Yang et al., 2022; Cui et al., 2023; Hao et al., 2024), have renewed excitement about learning fundamental representations of biology and their application to understanding co-regulatory effects. However, on the task of predicting gene expression under perturbation, these models have been outperformed by simple baselines (Bendidi et al., 2024; Ahlmann-Eltze et al., 2024a; Gaudelet et al., 2024).

The main objective of perturbation studies is to identify genes that drive meaningful transcriptional changes. Usually, a transcriptomic perturbation signature is defined by identifying genes exhibiting the most significant changes in expression upon perturbation of the target gene. However, differential expression analyses often treat each gene in isolation, failing to account for the intricate interde-

pendencies between genes that are crucial to understanding perturbation effects at the cellular level. This approach also has unique challenges when applied to single-cell perturbation data, due to high sparsity (e.g. zero-inflated gene expression distributions), as well as varying perturbation efficiency of single-guide RNAs (sgRNAs), which is difficult to distinguish from biological signal or technical noise. Together with a limited number of cells per perturbation, this results in difficulties in identifying differentially expressed genes in 70- 89% of the genome-wide perturbations performed (Replogle et al., 2022; Nourreddine et al., 2024). It is unclear how many of these perturbations give rise to a meaningful transcriptomic signal, and they are commonly excluded from any downstream analyses. To fully utilise genome-scale perturbation data atlases, approaches need to identify perturbations that elicit a change in the overall transcriptomic state of the cell instead of focusing only on differentially expressed genes.

In this paper, we introduce a method for learning representations of single-cell transcriptomes that encode information about the perturbation state (perturbed/unperturbed) based on a cell's whole transcriptome. To this end, we present a novel fine-tuning strategy for single-cell foundation models using contrastive learning. By fine-tuning these models on single-cell perturbation data, we aim to improve their ability to capture and predict perturbation-driven transcriptomic changes even when no differentially expressed genes can be detected. We show that our approach captures more signals in the data than previous methods, allowing for a more comprehensive use of perturbation analysis in single-cell biology.

2 Related Work

2.1 Perturbation Analysis

There is no standard analysis for identifying perturbations that induce transcriptomic changes in single-cell RNA sequencing (scRNA-seq) data. Most approaches rely on differential gene expression analysis (Replogle et al., 2022; Nourreddine et al., 2024), using statistical tests like the Mann-Whitney U test or regression models fitted to expression values with a negative binomial distribution (Alessandrì et al., 2019; Chen et al., 2025). A drawback of these methods is the focus on large individual gene-level changes to determine the impact of perturbations, which may miss larger global shifts created by small combinatorial changes. Approaches that aim to determine distribution shifts throughout the transcriptome use dimensionality reduction with Principal Component Analysis (PCA), followed by computing distance metrics such as the energy distance (Peidli et al., 2024). Other approaches that are commonly used to project high-dimensional scRNA-seq data into lower-dimensional embeddings include Variational Autoencoders (VAEs), exemplified by models such as scVI (Lopez et al., 2018), which have recently been compared in benchmarks of perturbation analyses (Bendidi et al., 2024). Recently published single-cell foundation models, such as Geneformer (Theodoris et al., 2023) and scGPT (Cui et al., 2023), are transformer-based models that are pre-trained on large-scale single-cell atlases and then fine-tuned on a range of downstream tasks, where they often outperform existing methods. For perturbation analyses, however, these models are still outperformed by simple methods like PCA (Bendidi et al., 2024; Ahlmann-Eltze et al., 2024b).

2.2 Contrastive Learning

Contrastive learning is an approach that helps to extract meaningful representations by contrasting similar and dissimilar pairs of data points; it leverages the assumption that similar examples should be closer in a learned embedding space, while dissimilar examples should be farther apart. This type of training is commonly used in image-caption pre-training and fine-tuning (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022; 2023), where aligning image and text representations in a shared embedding space enables good performance on zero-shot transfer tasks, such as classification and retrieval. Recently, there has also been a surge in the use of contrastive learning to fine-tune encoder-only language models for improved sentence representation learning (Gao et al., 2021; Zhang et al., 2022; Chuang et al., 2022). In particular, contrastive learning can alleviate the problem of anisotropy, where sentence embeddings occupy only a narrow cone in the embedding space, leading to poor sentence representation and low performance in downstream tasks (Xu et al., 2023).

3 METHODS

3.1 PRE-TRAINING ON LARGE SCALE SINGLE-CELL RNA-SEQ DATA

We downloaded scRNA-seq data from $\sim 33M$ unique cells across 265 datasets in the census dataset (version 2023-07-25) from the CellXGene data portal (CZI Single-Cell Biology Program et al., 2023). Data processing followed Theodoris et al. (2023), representing each single-cell transcriptome as a sequence of gene names of maximum length $2,048$ ordered by their median-normalised expression. We excluded cancer cells and cells with < 500 expressed genes. Our single-cell foundation model is based on a bidirectional transformer encoder-only architecture (BERT) similar to Geneformer (Theodoris et al., 2023), receiving a single-cell transcriptome as an ordered sequence of gene names of maximum length $2,048$. The model was pre-trained with a masked language modelling task (masking 15% of input tokens) for three epochs. For more implementation details see Appendix A.1, and a schematic of the model is shown in Figure 1A. We compared our model and Geneformer on a dataset that was recently published and, therefore, did not form part of the training data for either model (Heimlich et al., 2024). The two models performed comparably, with our model slightly outperforming Geneformer in reproducing the overall ranking of highly expressed genes (Appendix A.2 and Supplementary Figure A.1).

3.2 CONTRASTIVE FINE-TUNING ON PERTURB-SEQ

We leveraged the largest genome-scale perturbation dataset to date of 9,866 perturbations (knockdowns) in 1.98 million K562 lymphoblast cells (Replogle et al., 2022). For each cell, we recorded the perturbed gene or noted it as "unperturbed" if a non-targeting sgRNA was used. There are $75,000$ unperturbed cells, and the median number of cells per perturbation is ~ 200 (Figure A.2). For each perturbation, we identified differentially expressed genes (DEGs) between perturbed and unperturbed cells using a Mann-Whitney U test (Supplementary Figure A.3). Using the E-distance between perturbed and unperturbed PCA embeddings as described in Section 3.4, we observed a Pearson correlation of 0.54 between E-distance and DEG counts. We selected $1,541$ perturbations with ≥ 20 DEGs for fine-tuning, as these are more likely to exhibit larger overall transcriptional changes, resulting in a stronger training signal. In our evaluation, however, we assessed the model's generalisation to putative perturbations for which no DEGs were detected; that is, perturbations that induce subtle yet widespread transcriptional changes, even when the expression of individual genes do not show large shifts. Perturb-seq data was normalised with gene medians from the pre-training dataset before ranking genes by their expression into an ordered sequence of gene names, as described above. We grouped perturbed cells based on their perturbed target gene, and split cells by target gene into train, validation and test sets by a 80/10/10 rule. We split the unperturbed cells using the same 80/10/10 rule, and randomly sampled size-matched sets from the training set for each training perturbation to provide examples of control/dissimilar cells during training.

We used the final checkpoint of our pre-trained model for fine-tuning. A single training sample consisted of a pair of single-cell transcriptomes that are either "similar" or "dissimilar": a perturbed and unperturbed cell constituted a dissimilar pair, whereas two unperturbed cells formed a similar pair. The cell embeddings for both transcriptomes in the pair were obtained from the model as described in Section 3.3. The following contrastive loss was calculated for embedding pairs in each batch and back-propagated through all layers of the model:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left[y_i \cdot d_i^2 + (1 - y_i) \cdot \max(0, m - d_i)^2 \right], \tag{1}$$

where N is the batch size, m is the margin hyperparameter, and $d_i = \|\mathbf{o}_1^i - \mathbf{o}_2^i\|_2$ is the Euclidean distance between the two embedding vectors $\mathbf{o}_1^i$ and $\mathbf{o}_2^i$ of the single-cell transcriptomes belonging to the ith pair in the batch, and $y_i \in \{0, 1\}$ is the corresponding binary label, where 1 indicates that the pair is similar, and 0 dissimilar. The term $\max(0, m - d_i)$ ensures that the dissimilar loss is non-negative and only contributes when d_i is less than the margin. A schematic of the contrastive fine-tuned model is shown in Figure 1.

We ensured that cells with the same perturbed gene were in the same batch when training on dissimilar pairs, alternating between training on batches of similar and dissimilar pairs. Note that we did not explicitly train the model to learn separations between perturbations. We fine-tuned the models

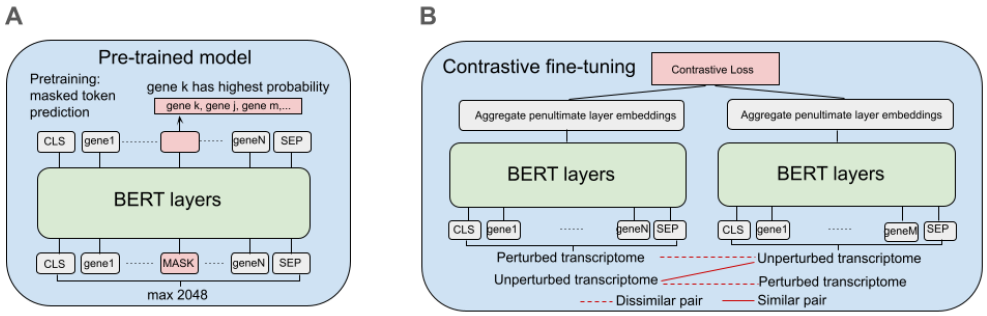

Figure 1: A: Schematic of the pre-trained foundation model. The model is trained with a masked language modelling task where input tokens are randomly masked and the model is trained to predict the right gene name. B: Schematic of the model fine-tuning task. The input to the model is a pair of transcriptomes: a perturbed and unperturbed cell form a dissimilar pair, or two unperturbed cells are a similar pair. The gene embeddings of the penultimate layer of the pre-trained model are aggregated into a cell embedding and the contrastive loss shown in Equation 1 is calculated for embedding pairs in each batch and back-propagated through all layers of the model.

using a margin of 20, a learning rate of $1e-5$, using the AdamW optimiser with a weight decay of $1e-2$. Fine-tuning of 10 epochs took 7 days on a single NVIDIA T4 15G GPU.

3.3 Cell embeddings

To obtain cell embeddings from the pre-trained or fine-tuned model, we performed a forward pass on an input transcriptome and extracted the contextualised embedding for each input gene from the penultimate layer. To compute the cell embedding, we took the mean of all of the cell's gene embeddings to form a single embedding vector of dimension 256, similar to Theodoris et al. (2023).

3.4 Distance Metrics

Our objective is to learn embeddings of single-cell transcriptomes that capture transcriptomic differences between perturbed and unperturbed cells. To evaluate our model, we used a set of metrics designed to measure how well the embeddings separated perturbed from unperturbed cells, as well as different pairs of perturbations. Although the model was not directly trained to do the latter, this evaluation tests whether the model captures more general differences in transcriptomic states. To measure the quality of the embeddings in this regard, we used the following metrics:

- **Energy distance (E-distance)**: This metric measures the statistical dispersion between two groups, making it particularly useful for quantifying the separation between distributions. Here, it captures the difference in expression profiles between perturbed and unperturbed cells, taking into account the variation within the two groups (Peidli et al., 2024). Higher values indicate greater separation, suggesting that a model is better at distinguishing transcriptional changes caused by perturbations. The distance between the distributions P and Q is defined as

$$D(P,Q) = 2\mathbb{E}[d(X-Y)] - \mathbb{E}[d(X-X')] - \mathbb{E}[d(Y-Y')], \tag{2}$$

  where, $X, X' \sim P$ and $Y, Y' \sim Q$ are samples drawn from P and Q, respectively, and d is the Euclidean distance. When calculating D for pairs of perturbations, X and Y are the whole sets of the cell embeddings for each perturbation target, whereas when computing D for perturbed and unperturbed cells, Y is replaced by embeddings of random samples from the unperturbed cell population (size-matching X).

- **Cosine E-distance:** Similar to E-distance, but d in Equation 2 is taken to be the cosine distance instead of the Euclidean distance. This change makes the distance invariant to the magnitudes of the embedding vectors that are compared.

- **High-dimensional Wasserstein distance:** This metric quantifies the minimal effort required to transform one distribution into another. While the E-distance emphasizes group separation, this metric provides a more nuanced measure of difference in distributions.

To compare the separation quality of different model embeddings, we L2-normalised each embedding to account for variations in embedding dimensions. We then applied a z-score normalisation to each metric using a control distribution, which was derived from pairs of embedding groups randomly sampled from the unperturbed cell population, to account for different variability and scales in the embeddings of the unperturbed distribution. The normalisation of metric D is given by

$$D_{norm} = \frac{D(P, Q) - \mu_{\text{control}}}{\sigma_{\text{control}}},$$

where μ_{control} and σ_{control} denote the mean and standard deviation of the control distribution, and D_{norm} is the normalised distance.

4 Results

4.1 Contrastive fine-tuning allows the encoding of distinct transcriptomic states in cell embeddings

For a visual assessment of the impact of contrastive fine-tuning, Figure 2 shows UMAP projections of embeddings of perturbed and unperturbed cells in the test set from the pre-trained and the fine-tuned model. Before fine-tuning, there is significant mixing between the embeddings of the perturbed and unperturbed cells, indicating that the pre-trained model is unable to distinguish between them. There is also a separate cluster that contains a mixture of perturbed and unperturbed cells, likely capturing batch effects. After contrastive fine-tuning, however, the model's embeddings separate better between perturbed and unperturbed cells: embeddings of the unperturbed cells now cluster together, and no strong batch effects are apparent. While there is some overlap of unperturbed and perturbed cell embeddings, there is a more pronounced distinction between them, suggesting that the fine-tuning process has improved the model's ability to detect transcriptomic changes.

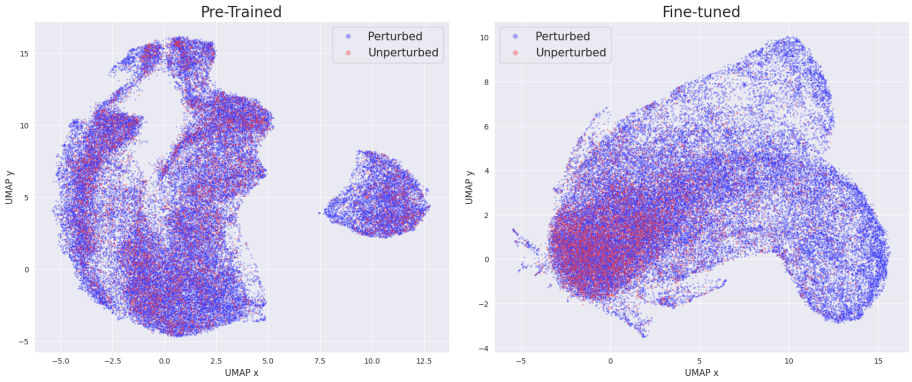

Figure 2: UMAP visualisation of cell embeddings in the test set from the pre-trained model before (left) and after contrastive fine-tuning (right). Perturbed cells ($N = 47,349$) of 232 perturbations are shown in blue, and unperturbed cells ($N = 7,470$) in red.

Figure 3 shows examples of cell embeddings for two perturbations, *TAF10* and *DHDDS*, and their separation from a size-matched random sample of unperturbed cells from the test set. These perturbations rank among the top five in the test set with the largest E-distance in the pre-trained model. In both cases, cell embeddings of perturbed cells separate from embeddings of unperturbed cells after contrastive fine-tuning, indicating the model's increased ability to distinguish cells with different transcriptomic states.

While the model was not explicitly trained to distinguish transcriptomes from different perturbation targets, we hypothesised that if the model uses transcriptomic states to distinguish unperturbed and

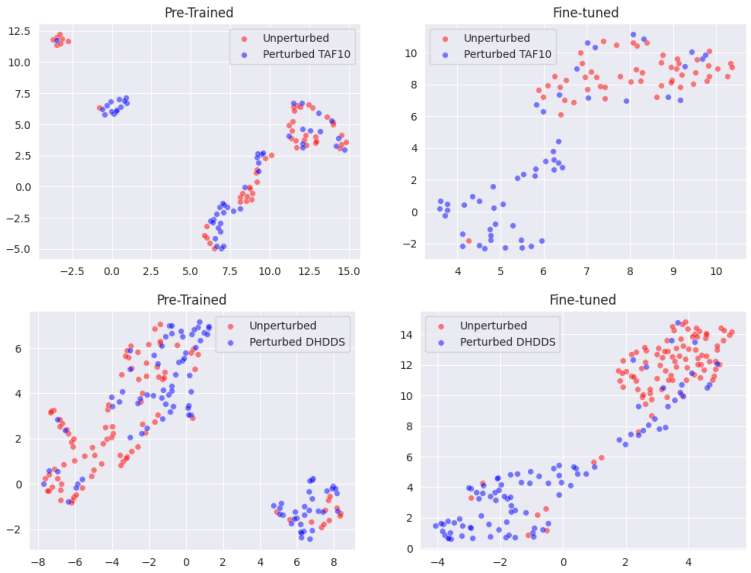

Figure 3: UMAP visualisation of embeddings of perturbed (blue) and unperturbed (red) cells from the test set for the perturbation targets *TAF10* (top) and *DHDDS* (bottom), before (left) and after contrastive fine-tuning (right).

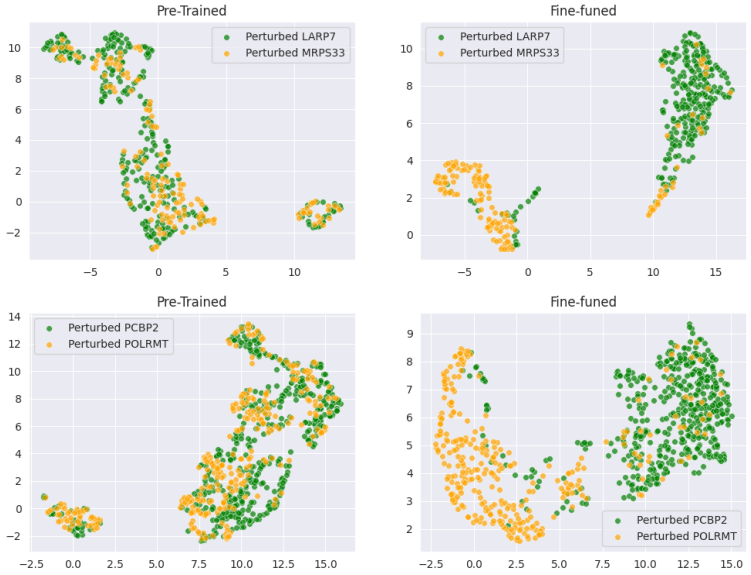

Figure 4: UMAP visualisation of embeddings of perturbed cells from the test set with perturbation targets *LARP7* vs *MRPS33* (top) and *PCBP2* vs *POLRMT* (bottom), before (left) and after contrastive fine-tuning (right).

perturbed cells, then it should capture these changes also for different perturbation targets. Figure 4 shows UMAP visualisations of cell embeddings before and after contrastive fine-tuning for the perturbation targets *LARP7* vs *MRPS33* and *PCBP2* vs *POLRMT*. In both cases, the fine-tuned model is able to distinguish between transcriptomes arising from the different unseen perturbations, further showcasing the model's ability to capture more nuanced perturbation effects after contrastive fine-tuning.

4.2 CONTRASTIVE FINE-TUNING OUTPERFORMS STRONG BASELINE MODELS

To assess the ability of our fine-tuned model to capture transcriptomic states upon perturbation, we quantified the model's performance with three distance metrics described in Section 3.4. We compared our model against five other approaches, some of which have been shown to outperform deep learning models in perturbation analysis (Bendidi et al., 2024):

1. Pre-trained: the single-cell foundation model before fine-tuning.
2. Highly variable genes (HV): using the expression of the top 300 highly variable genes across the whole dataset (including both perturbed and unperturbed cells).
3. PCA: applying PCA on the whole dataset (log normalised) and selecting the top 200 principal components.
4. scVI Linear: a variational autoencoder with a single encoder and decoder layer trained on raw expression counts of all cells in the training set of the contrastive model (Lopez et al., 2018). See A.3 for training and implementation details.
5. scVI 5L: scVI where both the encoder and decoder contain 5 layers trained on raw expression counts of all cells in the training set of the contrastive model (Lopez et al., 2018). See A.3 for training and implementation details.

As an initial visual assessment of the methods, we compared the UMAPs of PCA and scVI Linear on the same perturbations used in Figures 3 and 4. The UMAPs of both models are shown in Figure A.4 and A.5; both models displayed mixing of the perturbed and unperturbed cells, but better separation than the pre-trained model.

Next, we compared the models on the metrics described in Section 3.4. To ensure fair comparison across methods, we normalised embeddings and distance metrics as described in Section 3.4. The control distributions for various distance measures across different models (Supplementary Figure A.6) highlight the necessity of normalisation for a fair comparison. The contrastive model has the highest median across all 3 distance metrics in the case of separating perturbed from unperturbed transcriptomes (Figure 5, Table 1). PCA proved a competitive baseline on the E-distances, although performing worse than the contrastive model (E-distance medians: PCA$= 49.4$ vs fine-tuned$= 84.14$). In contrast, the pre-trained model without fine-tuning and the highly variable genes showed relatively low performance across all metrics, suggesting limited ability to separate cells based on perturbation-induced transcriptomic changes. The scVI Linear model was the only model to outperform the contrastive model on any of the metrics: while the contrastive model was superior on the Wasserstein distances, scVI Linear had a higher E-distance and Cosine E-distance between perturbation pairs (E-distance medians: scVI Linear $= 137.9$ vs fine-tuned$= 102.0$). For all metrics the contrastive model exhibits the largest positive skew, suggesting that it can separate certain perturbations extremely well.

These results demonstrate that the contrastive model effectively captures the differences in distributions between perturbed and unperturbed cell profiles, as well as variations within different perturbations. However, scVI Linear exhibited greater separation between perturbation pairs, as indicated by the larger median E-distances, likely due to its ability to leverage latent representations. The high variation of E-distances and Cosine E-distances in the contrastive model suggests that while it effectively distinguishes certain perturbations, others are not optimally separated.

4.3 QUALITATIVE ASSESSMENT OF PREDICTED LARGE TRANSCRIPTOMIC CHANGES FOR CELLS WITHOUT DIFFERENTIALLY EXPRESSED GENES

The model was fine-tuned only on perturbations for which at least 20 DEGs were identified, as we could assume enough signal in this data for the model to learn distinguishing patterns between perturbed and unperturbed cells. Of real interest, however, is whether the model can also be used to identify perturbations that elicited a global transcriptomic change even though no DEGs could be called. To this end, we set out to assess the quality of the contrastive model's embeddings for perturbed transcriptomes that (i) were excluded from fine-tuning due to having 0 DEGs and (ii) exhibited a high distance metric compared to unperturbed cells in the embedding space of the contrastive model.

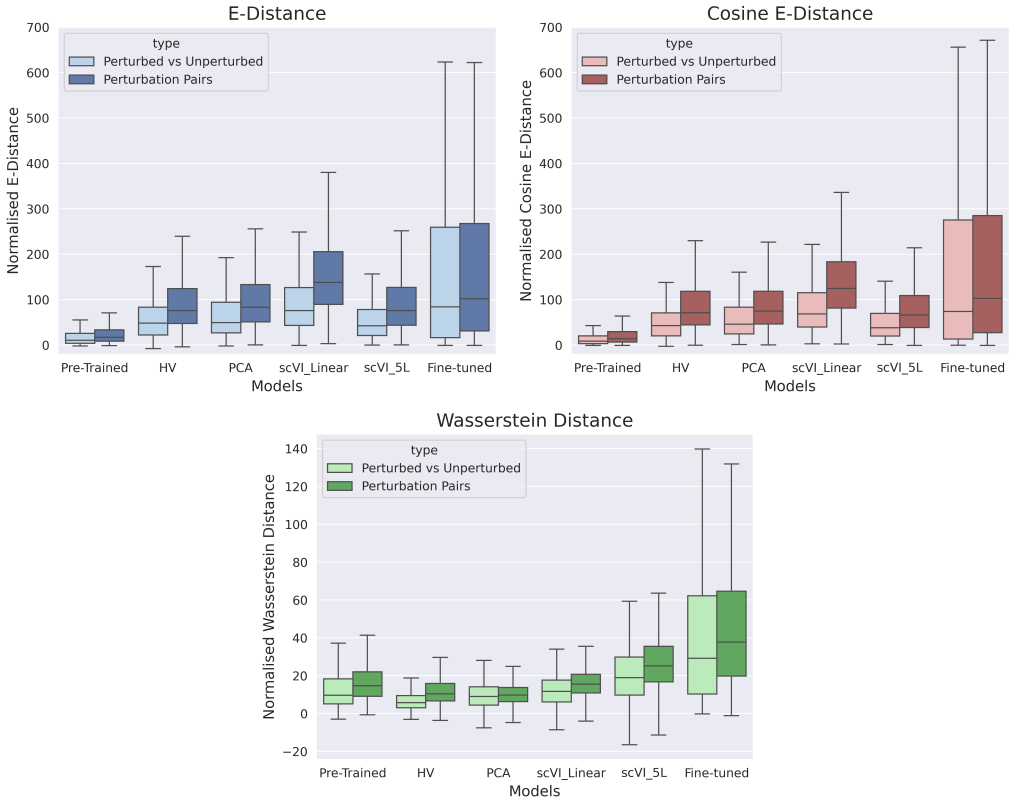

Figure 5: Distance distributions for perturbations in the test set across models. Shown are distances between embeddings of perturbed and unperturbed cells (light colour, $N = 232$ perturbations) and between perturbed cells (dark colour, $N = 232^2$ perturbation pairs).

In the absence of ground truth, we cannot be certain whether the model detected a global transcriptomic shift for these perturbations or whether the large distances were an artefact of the model. To validate perturbed transcriptomes where the model detected substantial transcriptomic shifts, we examined whether perturbation targets were enriched for gene sets that are particularly prone to inducing significant transcriptional changes when targeted for down-regulation. We focused on genes with critical roles in cellular function. For example, cell cycle genes regulate important processes such as cell growth, DNA replication, and division, ensuring proper cell proliferation and genomic integrity. Similarly, essential genes are indispensable for cell survival and fundamental biological functions, with their disruption typically leading to cell death or failure of vital processes.

To this end, we obtained a list of 663 cell cycle genes (The Gene Ontology Consortium et al., 2023; Ashburner et al., 2000) and $2,058$ essential genes (Replogle et al., 2022). We extracted the cell embeddings of $6,248$ perturbations with 0 DEGs (which were excluded during fine-tuning), obtained embeddings of the same transcriptomes from PCA and scVI Linear for comparison, and ranked the perturbations by their normalised distances from unperturbed cells (from the test set). Regardless of the distance metric used, the contrastive model placed perturbations targeting cell cycle and essential genes more often into the top n most distant embeddings than PCA or scVI Linear for different values of n (Supplementary Figure A.7). To evaluate the significance of the enrichment, we conducted a permutation test (with $10,000$ permutations) for the top n perturbations (according to their E-distances), comparing them to randomly sampled perturbations for different values of n (Figure 6). There was a significant enrichment (p-value < 0.001, Bonferroni correction for multiple tests) for cell cycle genes for all $n \leq 2,500$ and for essential genes for all $n \leq 2,000$. Results were similar when using the two other distance metrics (Supplementary Figure A.8).

In summary, our contrastive model was able to identify biologically relevant perturbations, even when the number of DEGs failed to capture their effects.

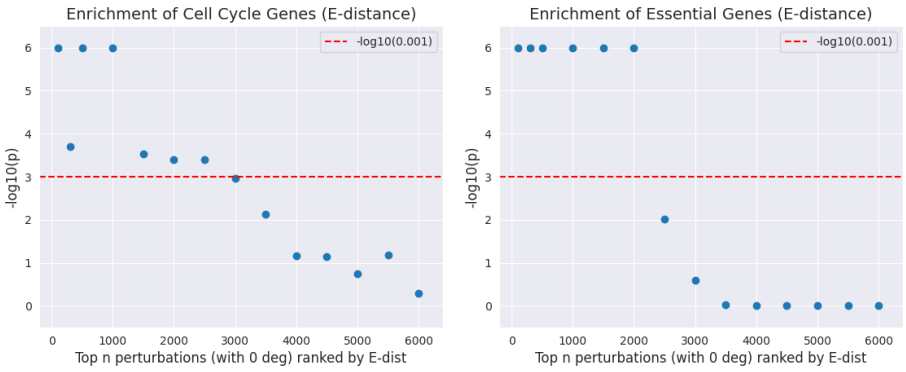

Figure 6: Enrichment analysis of cell cycle (left) and essential genes (right) for perturbations with 0 DEG ranked by normalised E-distance based on the embeddings from the contrastive model. Shown is the log-transformed enrichment p-value for different numbers of top n perturbations.

5 Discussion

We introduced a novel fine-tuning strategy for single-cell foundation models on Perturb-seq data that leverages contrastive learning to capture transcriptome-level changes in the cell embeddings. We demonstrated the performance of our approach for perturbation analysis by benchmarking our fine-tuned model against existing approaches using three distance metrics. We focused our comparisons on distance metrics instead of additional clustering as it allows a direct comparison of the separation of the embeddings and ranking of perturbations, independent of the clustering techniques used. We showed that our model identified perturbations that would have been overlooked by traditional differential expression analysis, showing significant enrichment in biologically relevant pathways and functions, including cell cycle regulation and gene essentiality.

In future work, we will explore other contrastive loss functions (Oord et al., 2019) and hybrid methods that integrate contrastive and cross-entropy losses on perturbation targets (Gunel et al., 2021). While our current model was trained on high-DEG perturbations, we will experiment with expanding training to cells with low-DEG perturbations, and evaluate the impact on predictive performance. To mitigate signal dilution, we will explore self-supervised learning approaches that do not rely on predefined similarity labels to learn transcriptome representations (Kim et al., 2021). Further, we will explore explainability techniques developed for BERT-based models to identify genes driving the most significant transcriptomic shifts (Aken et al., 2020; Talebi et al., 2024).

In summary, our method allows for a more comprehensive analysis of perturbation data, thus aiding the identification of perturbations that induce significant transcriptomic changes, and, ultimately, the understanding of disease mechanism.

References

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear methods, September 2024a. URL http://biorxiv.org/lookup/doi/10.1101/2024.09.16.613342.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear methods, September 2024b. URL http://biorxiv.org/lookup/doi/10.1101/2024.09.16.613342.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. VisBERT: Hidden-State Visualizations for Transformers, November 2020. URL http://arxiv.org/abs/2011.04507. arXiv:2011.04507 [cs].

Luca Alessandrì, Maddalena Arigoni, and Raffaele Calogero. Differential Expression Analysis in Single-Cell Transcriptomics. In Valentina Proserpio (ed.), *Single Cell Methods*, volume 1979, pp.

425–432. Springer New York, New York, NY, 2019. ISBN 978-1-4939-9239-3 978-1-4939-9240-9. doi: 10.1007/978-1-4939-9240-9_25. URL http://link.springer.com/10.1007/978-1-4939-9240-9_25. Series Title: Methods in Molecular Biology.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1061-4036, 1546-1718. doi: 10.1038/75556. URL https://www.nature.com/articles/ng0500_25.

Ihab Bendidi, Shawn Whitfield, Kian Kenyon-Dean, Hanene Ben Yedder, Yassir El Mesbahi, Emmanuel Noutahi, and Alisandra K. Denton. Benchmarking Transcriptomics Foundation Models for Perturbation Analysis : one PCA still rules them all, November 2024. URL http://arxiv.org/abs/2410.13956. arXiv:2410.13956 [cs].

Yunshun Chen, Lizhong Chen, Aaron T L Lun, Pedro L Baldoni, and Gordon K Smyth. edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*, 53(2):gkaf018, January 2025. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkaf018. URL https://academic.oup.com/nar/article/doi/10.1093/nar/gkaf018/7973897.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings, 2022. URL https://arxiv.org/abs/2204.10298. Version Number: 1.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI. preprint, Bioinformatics, May 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.04.30.538439.

CZI Single-Cell Biology Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. preprint, Cell Biology, November 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.10.30.563174.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings, 2021. URL https://arxiv.org/abs/2104.08821. Version Number: 4.

Thomas Gaudelet, Alice Del Vecchio, Eli M. Carrami, Juliana Cudini, Chantriolnt-Andreas Kapourani, Caroline Uhler, and Lindsay Edwards. Season combinatorial intervention predictions with Salt & Peper, April 2024. URL http://arxiv.org/abs/2404.16907. arXiv:2404.16907 [cs, q-bio].

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo Da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, February 2022. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-021-01206-w. URL https://www.nature.com/articles/s41587-021-01206-w.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning, April 2021. URL http://arxiv.org/abs/2011.01403. arXiv:2011.01403 [cs].

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, June 2024. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-024-02305-7. URL https://www.nature.com/articles/s41592-024-02305-7.

J. Brett Heimlich, Pawan Bhat, Alyssa C. Parker, Matthew T. Jenkins, Caitlyn Vlasschaert, Jessica Ulloa, Joseph C. Van Amburg, Chad R. Potts, Sydney Olson, Alexander J. Silver, Ayesha Ahmad, Brian Sharber, Donovan Brown, Ningning Hu, Peter Van Galen, Michael R. Savona, Alexander G. Bick, and P. Brent Ferrell. Multiomic profiling of human clonal hematopoiesis reveals genotype and cell-specific inflammatory pathway activation. *Blood Advances*, 8(14): 3665–3678, July 2024. ISSN 2473-9529, 2473-9537. doi: 10.1182/bloodadvances.2023011445. URL https://ashpublications.org/bloodadvances/article/8/14/3665/515374/Multiomic-profiling-of-human-clonal-hematopoiesis.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, June 2021. URL http://arxiv.org/abs/2102.05918. arXiv:2102.05918 [cs].

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Assessing the limits of zero-shot foundation models in single-cell biology. preprint, Bioinformatics, October 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.10.16.561085.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-Guided Contrastive Learning for BERT Sentence Representations, June 2021. URL http://arxiv.org/abs/2106.07345. arXiv:2106.07345 [cs].

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0229-2. URL https://www.nature.com/articles/s41592-018-0229-2.

Sami Nourreddine, Yesh Doctor, Amir Dailamy, Antoine Forget, Yi-Hung Lee, Becky Chinn, Hamza Khaliq, Benjamin Polacco, Monita Muralidharan, Emily Pan, Yifan Zhang, Alina Sigaeva, Jan Niklas Hansen, Jiahao Gao, Jillian A Parker, Kirsten Obernier, Timothy Clark, Jake Y Chen, Christian Metallo, Emma Lundberg, Trey Ideker, Nevan Krogan, and Prashant Mali. A PERTURBATION CELL ATLAS OF HUMAN INDUCED PLURIPOTENT STEM CELLSd, November 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.11.03.621734.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs].

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3): 531–540, March 2024. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-023-02144-y. URL https://www.nature.com/articles/s41592-023-02144-y.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data*

Mining, pp. 3505–3506, Virtual Event CA USA, August 2020. ACM. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3406703. URL https://dl.acm.org/doi/10.1145/3394486.3406703.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, July 2022. ISSN 00928674. doi: 10.1016/j.cell.2022.05.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867422005979.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, August 2023. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-023-01905-6. URL https://www.nature.com/articles/s41587-023-01905-6.

Salmonn Talebi, Elizabeth Tong, Anna Li, Ghiam Yamin, Greg Zaharchuk, and Mohammad R. K. Mofrad. Exploring the performance and explainability of fine-tuned BERT models for neuroradiology protocol assignment. *BMC Medical Informatics and Decision Making*, 24(1):40, February 2024. ISSN 1472-6947. doi: 10.1186/s12911-024-02444-z. URL https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02444-z.

The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil Dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, Monte Westerfield, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil Dos Santos, Steven Marygold, Victor Strelets,

Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *GENETICS*, 224(1):iyad031, May 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031. URL https://academic.oup.com/genetics/article/doi/10.1093/genetics/iyad031/7068118.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://www.nature.com/articles/s41586-023-06139-9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

Lingling Xu, Haoran Xie, Zongxi Li, Fu Lee Wang, Weiming Wang, and Qing Li. Contrastive Learning Models for Sentence Representations. *ACM Transactions on Intelligent Systems and Technology*, 14(4):1–34, August 2023. ISSN 2157-6904, 2157-6912. doi: 10.1145/3593590. URL https://dl.acm.org/doi/10.1145/3593590.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10):852–866, September 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00534-z. URL https://www.nature.com/articles/s42256-022-00534-z.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-Shot Transfer with Locked-image text Tuning, June 2022. URL http://arxiv.org/abs/2111.07991. arXiv:2111.07991 [cs].

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training, September 2023. URL http://arxiv.org/abs/2303.15343. arXiv:2303.15343 [cs].

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. Unsupervised Sentence Representation via Contrastive Learning with Mixing Negatives. *Proceedings of the*

AAAI Conference on Artificial Intelligence, 36(10):11730–11738, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i10.21428. URL https://ojs.aaai.org/index.php/AAAI/article/view/21428.

A APPENDIX

A.1 IMPLEMENTATION

Hyperparameters were chosen to allow for distributed learning: max learning rate, 1×10^{-3} scaled by the number of GPUs; a learning scheduler, linear with warm-up (10k steps) and linear decay; Adam optimizer with weight decay parameter 0.001. The training was distributed over 4 GPUs in one node with a minibatch size 11 and 2 gradient accumulation steps.

To speed up pre-training we used dynamic padding combined with a length-grouped sampler to minimise computation on padding. This sampler takes a randomly sampled megabatch and then orders minibatches by their length in descending order. Mini-batches are then dynamically padded, minimising the computation wasted on padding as sequences of similar lengths are grouped. The authors of Geneformer extended an existing version of this sampler from Huggingface transformers for the distributed case (Theodoris et al., 2023; Wolf et al., 2020). However, neither of these samplers shuffle the mini-batches within the megabatch before passing them to the model, which resulted in a 60x-performance-drop of the trained model in our tests (in terms of training and test perplexity on smaller sample datasets) compared to model runs not employing the grouped-length batching. We implemented a shuffling of the mini batches which slightly diminishes the speed up during training.

For efficient data parallelisation across the GPUS, we used Deepspeed (Rasley et al., 2020). Overall, pre-training was achieved in just over 7 days distributed across one node with four Nvidia A10G 24GB GPUs.

A.2 PRE-TRAINING EVALUATION

To compare our pre-trained model to Geneformer (Theodoris et al., 2023), we evaluated both models on a dataset of $\sim$ 66k peripheral blood mononuclear cells (PBMCs) that was published after both models were trained (Heimlich et al., 2024).

We computed macro-averaged hits@k metrics on masked tokens at different thresholds in the 2000 highest expressed genes in 10k randomly sampled PBMCs (Figure A.1 A). Macro averaging gives equal weight to each gene when computing the accuracy, giving a sense of the model's performance overall. Here, an instance is one prediction instance, e.g. one of the masked genes in the input we ask the model to fill in. If one gene occurs much more often among the first 2000 genes of the input sequence and is, therefore, more often masked, a model could "cheat" overall metrics by always predicting that one gene. Macro-averaging per gene combats this bias. Both models performed similarly on this task.

To assess further the performance of the models, we followed Kedzierska et al. (2023) and investigated how well the models can reproduce the correct gene rankings per cell given the masked input. To do this, we compare the ground truth order with the predicted order of genes and compute the Spearman correlation coefficients for different cut-offs in all $\sim$ 66k PBMCs. Similar to Kedzierska et al. (2023), we find that both models struggle to correctly predict the positioning for lower-expressed genes, with our model performing better on reproducing the input rankings of the higher-expressed genes (Figure A.1 B).

A.3 IMPLEMENTATION AND TRAINING OF SCVI

We used the implementation of scVI from *scvi-tools* (Gayoso et al., 2022) and trained a linear VAE with a single encoder and decoder layer (denoted in the text by "scVI Linear") and a model where both encoder and decoder consisted of 5 hidden layers (denoted in the text by "scVI 5L"). The training data contained all unperturbed and perturbed cells from the training dataset of the contrastive model. The raw scRNA-seq counts of the cells without any further filtering or preprocessing formed the input to the models. We used a random 90/10 split for training and test sets to monitor model convergence. We used a Zero-Inflated Negative Binomial as the likelihood function, and hyperparameters for both models were set at $n\_latent = 300$, $dropout\_rate = 0.1$, and $max\_epochs = 500$ with early stopping.

A.4 SUPPLEMENTARY FIGURES

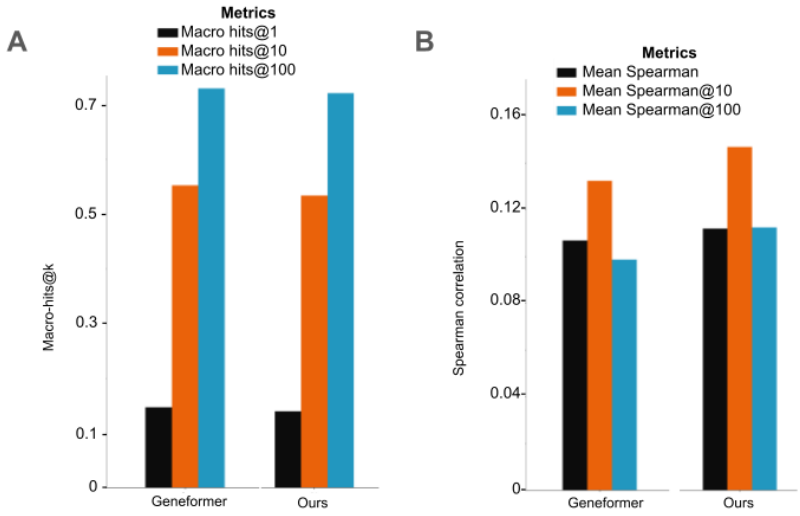

Figure A.1: A: Macro-averaged hits@k metrics for $k = 1, 10$ and 100 measuring the performance of correct masked token prediction of the models in 15% masked genes in 10k randomly sampled PBMCs. B: Mean Spearman correlations at different thresholds between ground-truth and predicted gene rankings for the top 100, 500 and 2000 expressed genes per cell in 66k PBMCs. 15% of the input genes were masked at random and the model was asked to generate full ranking outputs for all positions.

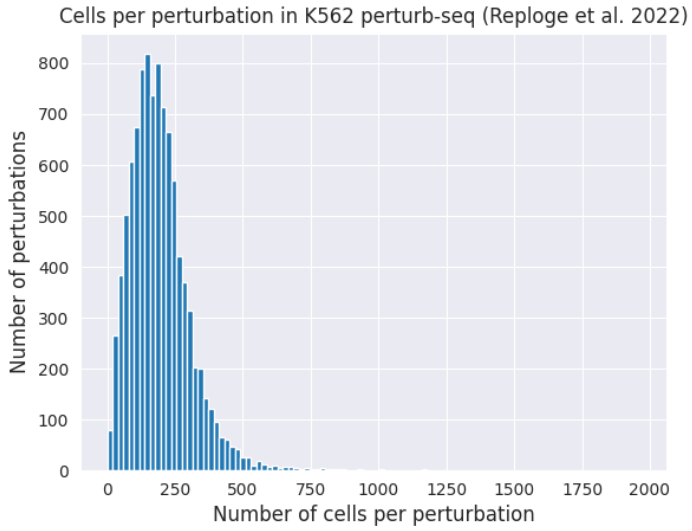

Figure A.2: Histogram of the number of cells per perturbation.

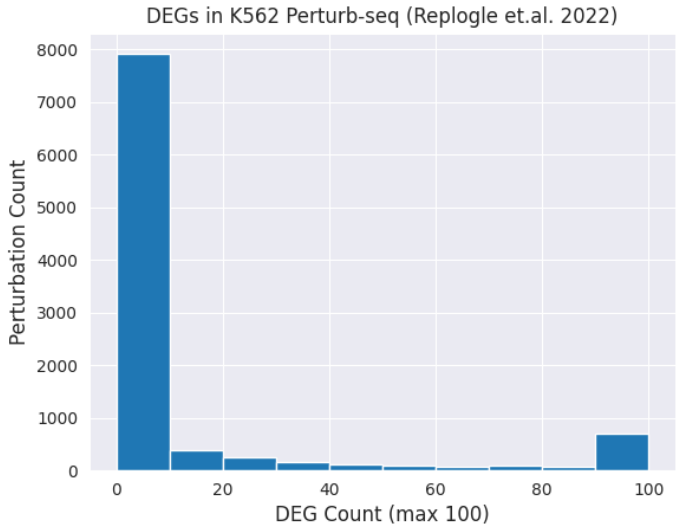

Figure A.3: Histogram of the number of DEGs per perturbation. Perturbations with more than 100 DEGs are captured in the last bin.

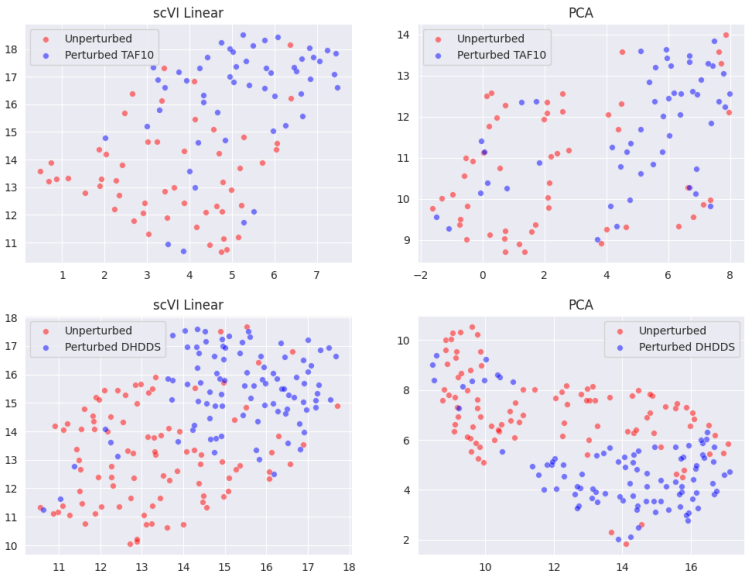

Figure A.4: UMAP visualisation of embeddings of perturbed (blue) and unperturbed (red) cells from the test set for the perturbation targets *TAF10* (top) and *DHDDS* (bottom), from PCA (left) and scVI Linear (right).

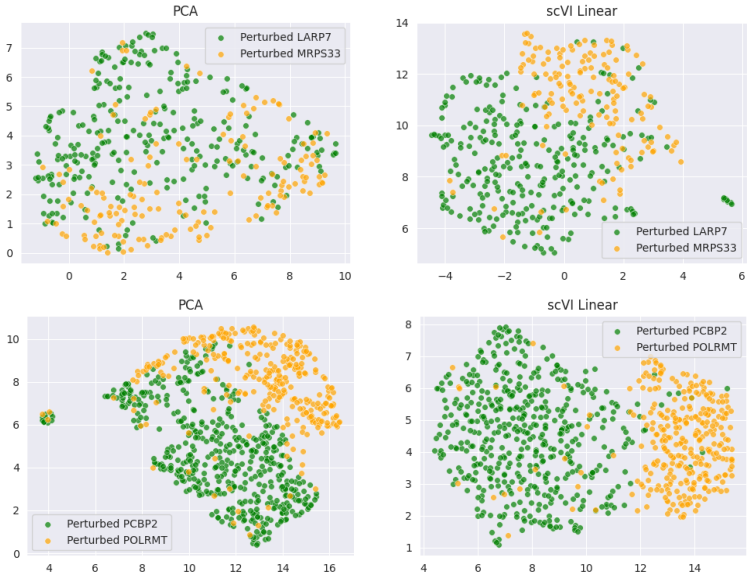

Figure A.5: UMAP visualisation of embeddings of perturbed cells from the test set with perturbation targets *LARP7* vs *MRPS33* (top) and *PCBP2* vs *POLRMT* (bottom), from PCA (left) and scVI Linear (right).

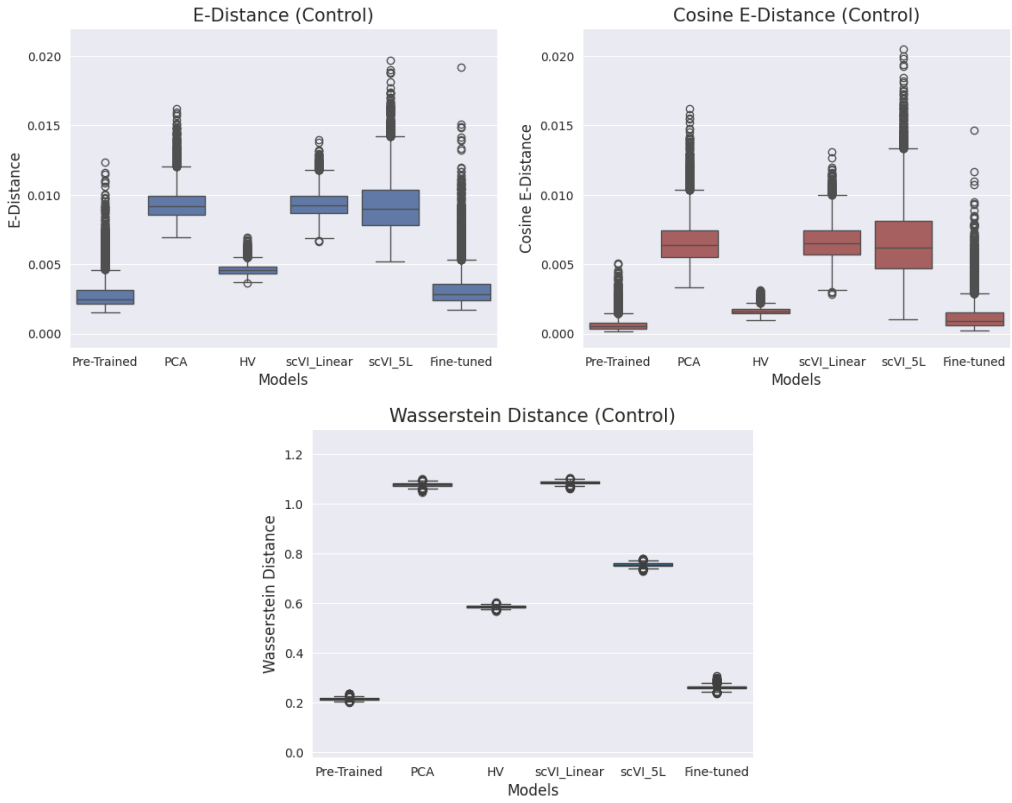

Figure A.6: E-distance (top left), Cosine E-distance (top right) and Wasserstein distance (bottom) control distributions of various models. Each data point represents the distance between the embeddings of two groups of randomly sampled unperturbed cells. In total, 10,000 random pairs of groups, each containing 300 cells, were sampled. The difference in control distributions of the models highlights the need to perform z-normalisation to provide fair comparisons across models.

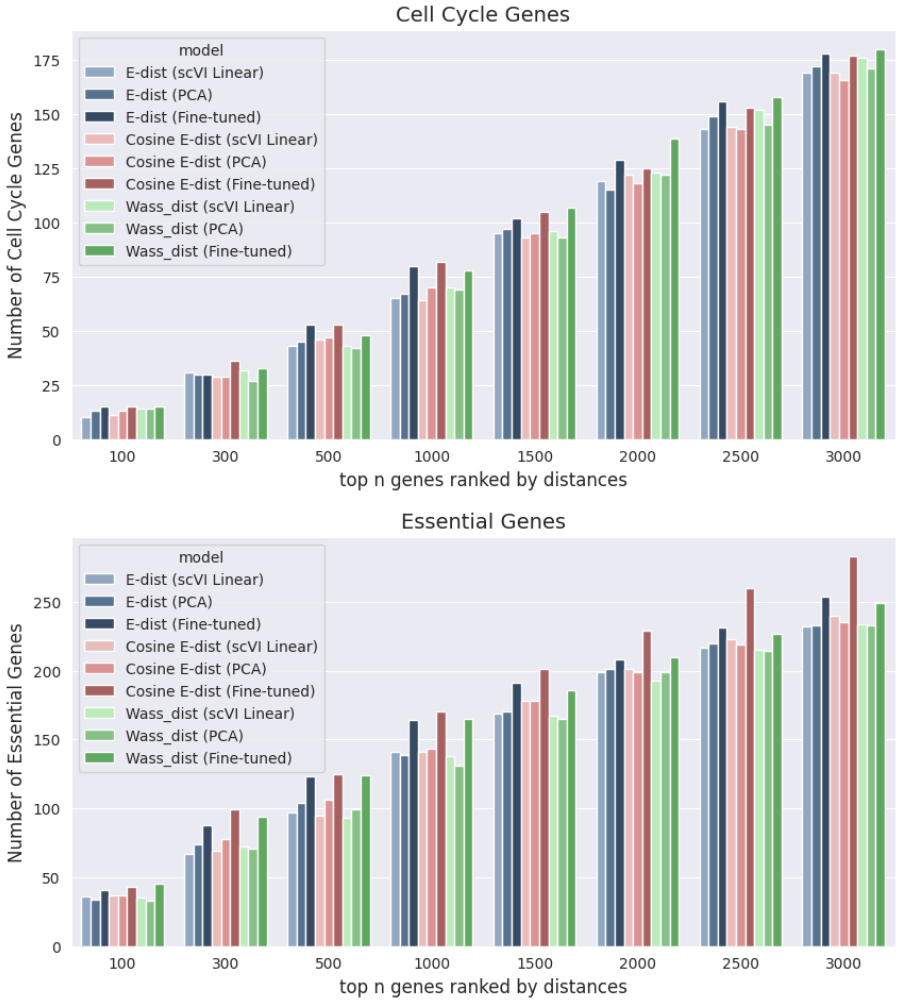

Figure A.7: Barchart showing the number of top n perturbations ranked by normalised E-distance (blue), normalised cosine E-distance (red) and Wasserstein distance (green) that are targeting either cell cycle genes (top) or essential genes (bottom). Comparing results from scVI Linear (lighter colours), PCA (median colours) and the contrastively fine-tuned model (darker colours).

Model	**Perturbed vs Unperturbed**			**Perturbation Pairs**		
	E-Dist	**Cosine E-Dist**	**Wass Dist**	**E-Dist**	**Cosine E-Dist**	**Wass Dist**
Fine-tuned	**84.14**	**74.10**	**29.18**	102.04	103.10	**37.79**
scVI Linear	76.09	68.73	11.76	**137.85**	**124.67**	15.58
scVI 5L	42.57	38.40	19.02	76.09	66.47	25.24
PCA	49.41	45.98	9.00	83.36	75.00	9.80
HV	48.27	43.13	5.77	76.02	71.31	10.41
Pre-Trained	10.73	8.54	9.70	17.30	14.24	14.73

Table 1: The medians (over all perturbations in the test set) of different normalized distances in the distributions of perturbed vs unperturbed and perturbation pairs across various models.

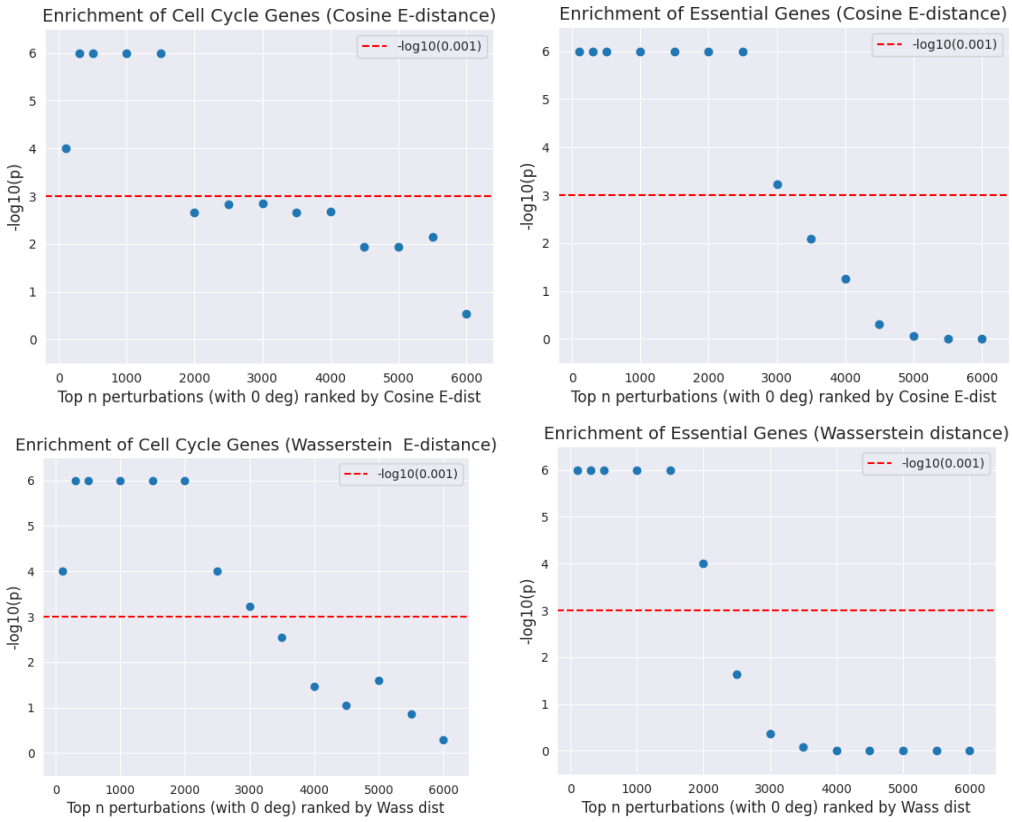

Figure A.8: Enrichment analysis of cell cycle (left) and essential genes (right) amongst perturbations with 0 DEGs ranked by normalised Cosine E-distance (top) and normalised Wasserstein distance (bottom) based on embeddings of the contrastive model. The minimum p-value is capped at $1e-6$ for the purpose of plotting.