DISSECTING AND STEERING CELL IDENTITY IN A SINGLE-CELL FOUNDATION MODEL USING SPARSE AUTOENCODERS

Nicolas Yanovsky[1,2]*, **Ariel Chernomoretz**[2,3]
[1]Universidad de Buenos Aires, [2]Fundacion Instituto Leloir, [3]Dto. de Fisica FCEN UBA, INFINA (CONICET)

ABSTRACT

Single-cell Foundation Models (scFMs) have demonstrated remarkable capability in learning cellular representations, yet their internal mechanisms remain largely opaque. In this work, we apply Sparse Autoencoders (SAEs) to the residual stream of AIDO.Cell, a transformer-based scFM, to decompose its latent space into interpretable biological features. We trained TopK SAEs on the 12th transformer layer using the PBMC3K dataset as a proof of concept. Using Gene Ontology enrichment to interpret features, we find that $\sim 64\%$ of trained SAE features achieve statistically significant biological annotations, compared to $\sim 34\%$ from the dense raw activations. Beyond interpretation, we demonstrate that these features can be used to functionally "steer" cell identity: amplifying and suppressing individual features (e.g., Viral Defense) drives symmetric changes in gene expression, with steered cell states aligning with the expected biological programs. Furthermore, we implement a contrastive steering method to automatically discover sparse feature combinations that drive CD4+ T cells towards a CD8+ T cell phenotype. Inspection of the selected features revealed that the model learned biologically relevant directions in latent space enabling cell-type steering. Our findings show that a scFM can learn a decomposable and manipulable model of cell biology, enabling interpretable in silico experiments.

1 INTRODUCTION

The development of Single-cell Foundation Models (scFMs) promises a unified understanding of cellular biology (Theodoris et al., 2023; Rosen et al., 2024). By pre-training on millions of transcriptomes, models such as scGPT (Cui et al., 2024) and scFoundation (Hao et al., 2024) have shown remarkable performance in tasks such as cell-type annotation and perturbation prediction. However, due to superposition (Elhage et al., 2022), these models compress biological complexity into lower-dimensional representations, resulting in entangled, polysemantic features that obscure the specific gene programs driving model behavior.

Mechanistic interpretability offers a solution by reverse-engineering model components into human-understandable concepts (Olah et al., 2020). Sparse Autoencoders (SAEs) have recently emerged as a powerful tool for this purpose, successfully decomposing dense neurons in Large Language Models (Templeton et al., 2024) and Protein Language Models (Adams et al., 2025) into sparse, monosemantic features. Recent work has begun to apply SAEs to scFMs (Pedrocchi et al., 2025; Schuster, 2025), yet these studies have often been constrained by their scFM architecture or focused on single-feature enrichments that lack strong causal validation.

In this work, we apply SAEs to AIDO.Cell-100M (Ho et al., 2024), a dense Transformer-based scFM that models the full 19,264-gene human transcriptome. We show that SAE trained features yield statistically significant Gene Ontology enrichments, which are more interpretable compared to those found with the raw activations. We also provide, via steering experiments, evidence that some of these features have causal influence on model's outputs, where amplifying or suppressing individual features drives targeted changes in gene expression. Finally, we implement a contrastive approach that discovers sparse feature combinations for cell-type reprogramming.

*correspondence: nyanovsky@dc.uba.ar

2 METHODS & RESULTS

2.1 SAE TRAINING AND INTERPRETATION

We trained a Top-K Sparse Autoencoder ($K = 32$, expansion factor $8\times$) on the residual stream (x) of the 12th transformer layer of AIDO.Cell (with a latent dimension of 640, resulting in 5120 features) using the PBMC3K dataset. This configuration balances feature granularity with computational efficiency, aligning with recent work in protein and single-cell modeling (Adams et al., 2025; Schuster, 2025). The encoder and decoder are defined as follows:

$$\hat{z} = TopK(W_{enc}(x - b_{pre}))$$

$$\hat{x} = W_{dec}\hat{z} + b_{pre}$$

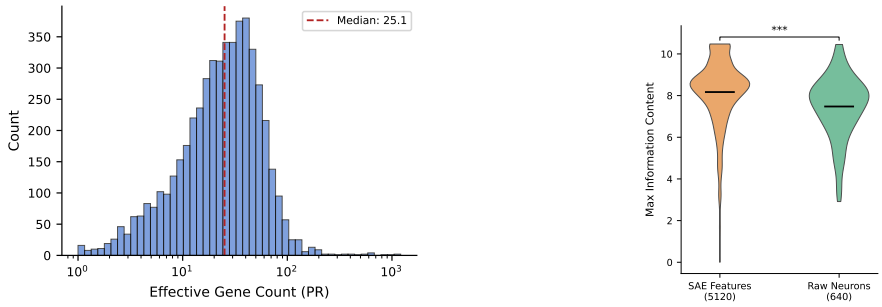
Where b_{pre} is a learned bias vector. SAEs were trained to minimize the reconstruction loss $||x - \hat{x}||^2$. To interpret the learned features, we compute a feature-gene activation matrix $\boldsymbol{F}$ (averaged across cells) and introduce a data-driven strategy to select top activating gene sets per feature. Standard enrichment analyses often rely on arbitrary Top-N gene sets; instead, we use the Participation Ratio (PR) to determine the biological footprint of each feature. We compute the PR for feature i from its squared activation vector (energy) across genes:

$$PR_i = \frac{(\sum_j \boldsymbol{F}_{ij}^2)^2}{\sum_j \boldsymbol{F}_{ij}^4}$$

Intuitively, this metric measures the "effective dimensionality" of the feature: if a feature activates N genes with equal strength, the PR is exactly N; if it relies on a single gene, the PR is 1. This provides a continuous, scale-invariant proxy for the number of active genes, allowing us to select the top PR genes to test for GO enrichment. Dynamically scaling our interpretation to the feature's scope allows us to discover features which activate from narrow markers to broad programs (Fig. 1a). We achieve statistically significant biological enrichments ($q < 0.05$, FDR-corrected) for 64% of features.

2.2 FEATURE SPACE ANALYSIS

Analysis of the feature space reveals biologically interpretable concepts ranging from broad cellular states (e.g., Feature 2685: Aerobic Respiration; Feature 1480: Cell Cycle Regulation) to cell-type specific programs (e.g., Feature 4590: Germinal Center Formation; Feature 3181: Type I Interferon Signaling). SAE features also show higher specificity than raw activation features, measured by their term's maximum Information Content (IC) (Fig 1b). A similar baseline was used using Independent Component Analysis (ICA) on the raw activations, with almost identical results.



(a) (b)

Figure 1: **a.** The PR distribution across all 5120 features. The histogram shows that features activate on gene sets of varying orders of magnitude. **b.** Maximum feature ICs of SAE vs raw features, showing SAE features are more specific by the Mann-Whitney U test.

Measuring overlap between feature's top activating gene sets shows these are highly orthogonal (Fig. S1), but constructing a feature graph where edges represent high overlap (> 0.2) reveals the existence of biologically coherent connected components (Fig S2a). These clusters group features that, while distinct, participate in common pathways (e.g., a Viral Defense cluster and a B-cell Signaling cluster).

2.3 Single Feature Steering

We validate causality by directly intervening on the residual stream x during the forward pass of AIDO.Cell. Following Templeton et al. (2024), we use the SAE's error term $e(x) = x - SAE(x)$ to preserve non-captured signal, and compute the steered residual as

$$x_{steered} = (\boldsymbol{\alpha} \odot \hat{z})W_{dec} + b_{pre} + e(x)$$

Where $\boldsymbol{\alpha}$ is a vector that determines steering strengths at the steered features, and $\odot$ is the element-wise product. That is, if we want to steer feature i by α, then $\boldsymbol{\alpha}$ is such that $\boldsymbol{\alpha}_{-i} = \mathbf{1}$, and $\boldsymbol{\alpha}_i = \alpha$. By scaling each gene's natural activation, instead of directly replacing it like in Templeton et al. (2024), we ensure keeping the steered state under biologically plausible conditions.

We perform experiments by selecting two features from different connected components; one corresponding to a "Viral Defense" state (feature 4367), and one corresponding to B-cell identity terms (Feature 3170). For each experiment, we test α values in the $[-5, 5]$ range and steer the full dataset. After steering, we perform GO enrichment ($q < 0.05$, FDR corrected) on the top 100 up and down-regulated differentially expressed genes (DEG) ($p < 0.05$, Bonferroni corrected) with respect to the unsteered model output.
For the "Viral Defense" feature, positive steering ($\alpha > 1$) significantly downregulates genes enriched in viral-defense related terms, while negative steering ($\alpha < 1$) upregulates them, suggesting the feature naturally encodes an anti-viral suppression state (note that the feature label is derived from its top enriched GO terms, not its causal direction).
The contrary can be seen for the "B-cell identity" feature, where positive steering results in upregulated genes enriched in B-cell related terms, while negative steering does the same for downregulated genes. This symmetric response (Tab. 1) confirms the feature encodes a continuous functional direction in feature space.

Table 1: Steering validation: top enriched GO terms show symmetric responses across opposite α values. Viral Defense acts as a repressor (terms appear in downregulated genes for $\alpha > 1$), while B-cell Identity acts as an activator (terms appear in upregulated genes for $\alpha > 1$).

Viral Defense (Feature 4367)

$\alpha = 5$ (down-genes)	$\alpha = -5$ (up-genes)
Defense response to virus	Defense response to virus
Type I interferon signaling	Type I interferon signaling
Neg. reg. viral genome replication	Neg. reg. viral process
Cytokine-mediated signaling	Antiviral innate immune resp.

B-cell Identity (Feature 3170)

$\alpha = 2$ (up-genes)	$\alpha = -2$ (down-genes)
B cell receptor signaling	B cell receptor signaling
B cell activation	B cell activation
Reg. of B cell proliferation	Reg. of B cell proliferation
Antigen receptor-mediated signaling	Reg. of B cell activation

2.4 Contrastive Steering

Finally, we introduce contrastive steering to identify trajectories for complex cell-state transitions. Using gradient descent, we find a sparse vector of steering strengths $\boldsymbol{\alpha}$ that minimizes the Euclidean distance between a set of steered cell embeddings $z_{steered}$ (by mean pooling the cell's final gene embeddings) and the target cell-type centroid μ_{target}. That is, instead of manually steering a predefined feature with a given strength α, we automatically find which features to steer at varying

strengths to achieve the objective:

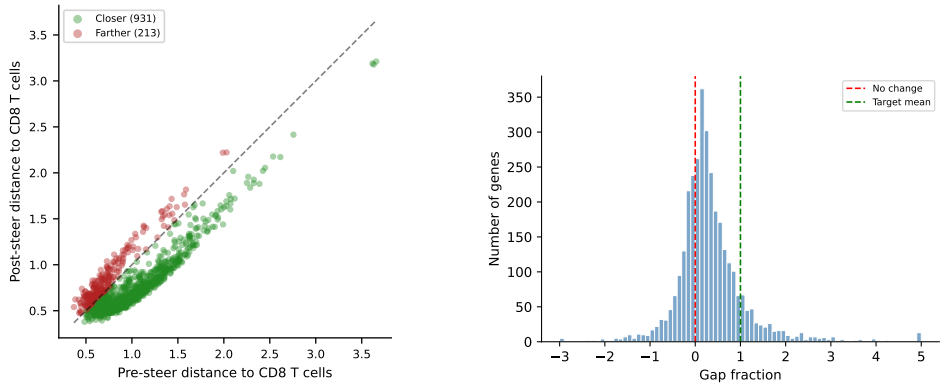
$$\min_{\boldsymbol{\alpha}} \|z_{steered}(\boldsymbol{\alpha}) - \mu_{target}\|^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

We applied this to identify features which would move CD4+ T cells into a CD8+ T cell phenotype. The experiment resulted in CD4+ cells shifting closer to the CD8+ centroid (Figs 2a, S3), even though distances closed were modest. Upon inspection of the resulting steering strengths $\boldsymbol{\alpha}$, we find 51 features such that $|\boldsymbol{\alpha}_i - 1| > 0.3$. The top 3 GO terms (ranked by IC) for each feature show that these include CD8+ T related terms such as regulation of CD8 T cell activation, antigen processing and presentation via MHC class I (in contrast with MHC class II, which is recognized by CD4 T cells), IL-12 and IL-18 response terms (which promote CD8 activity), and cytotoxic related terms. Nonetheless, these also include more generic terms corresponding to cell-cycle, ubiquitin/proteasome terms, RNA processing, among others. Still, we argue that the presence of the specific CD8 T terms are a good indication that steering occurred among a biologically realistic path.

We also assess steering results in the model's output expression space. For this, we take the set of DEG which are both differentially expressed between CD4s and CD8s and resulted differentially expressed from the steering process ($p < 0.05$, Bonferroni corrected). We compute the mean fraction of the logFC gap between the CD4 and CD8 expression closed by steering for each gene, that is, if for each gene i we define x_i^4, x_i^{4-s} x_i^8 as the gene's average output expression over CD4, steered CD4, and CD8 cells correspondingly, we compute the mean gap fraction for gene i as:

$$\frac{x_i^{4-s} - x_i^4}{x_i^8 - x_i^4}$$

. We find that 70% of genes move in the right direction for upregulated genes in CD8s (Fig 2b), while 65% do for downregulated genes (Fig S4a). Interestingly, we see that a positive effect ($> 50\%$ of genes moving in the right direction) is only achieved after including more than the first 100 DEG by effect size (S4b). For example, the CD8A and CD8B coreceptors remained almost intact by steering,



(a) (b)

Figure 2: **a.** Before and after steering distances from CD4 T cells to the CD8 T centroid, showing $\sim 81\%$ moved closer. **b.** Gap fraction distribution for upregulated genes in CD8 cells.

3 Conclusions

In this work, we show that Sparse Autoencoders can effectively decompose the dense activations of a full-transcriptome foundation model into interpretable biological programs. By introducing adaptive gene selection based on the Participation Ratio, we are able to capture distinct biological scales without the need for arbitrary thresholds. Furthermore, we show that these features are not merely correlative descriptors but can also be navigated to impact model outputs, either manually or through our contrastive optimization framework to reprogram cell identity in silico. By providing a mechanistic bridge between foundation model latents and biological gene sets, this approach serves as proof-of-concept for interpretable in-silico studies.

3.1 Limitations and Future Directions

Our study is primarily limited by the scale and scope of the PBMC3K dataset. While sufficient for methodological validation, its restriction to circulating immune cells means our feature dictionary likely under-represents tissue-resident or developmental programs. Furthermore, our interpretation relies only on Gene Ontology enrichment, which is inherently biased toward well-characterized pathways.

Another limitation factor is the absence of cross-model validation. We find that, given the compute resources available, AIDO.cell-100M was the only model that fit the project's requirements, being mainly full-transcriptome modeling across model layers (unlike scGPT or GeneFormer, which perform gene subsetting, or scFoundation, which only encodes non 0-expressed genes). We also found that scBERT's embedding quality was sub-par and not worth pursuing. The requirement for needing full-transcriptome modeling comes from feature interpretation, which is essentially a combinatorial problem (which sets of genes does this feature activate on?). Fixing a predefined set of genes would put a hard restriction on which features would be allowed to be discovered by the model. Not using full-transcriptome models would also make future downstream tasks nearly impossible. For example, if the task was to learn how to steer the model for reconstructing perturbed cell states, the same kind of gene subsetting would be needed in the SAE training dataset and the target dataset, making the problem much more difficult.

Future work will focus on scaling this pipeline to bigger and more diverse datasets, which would in theory make the SAEs more "universal", and expanding interpretability techniques beyond GO enrichment analysis. We also aim to further expand and analyze the contrastie steering experiment in a robust fashion. The fact that steering happens along plausible biological directions but seems to be dominated by changes in genes with lower effect sizes is a critical aspect to follow up on, needing to determine whether this is general to the steering process or an issue due to incomplete optimization.

The contrastive steering approach will work as a template for some possible downstream applications, e.g training MLP adapters on perturbation datasets to learn sparse feature combinations which result in the perturbed cell state, enabling interpretable perturbation prediction. We also aim to move beyond single-layer analysis to circuit discovery, mapping how features in early layers causally trigger downstream features in later layers.

REFERENCES

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, 2025.

Haotian Cui, Chloe Wang, Hassal Maan, Kuan Pang, Feng Ning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(7):1470–1480, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL `https://arxiv.org/abs/2209.10652`.

Mins Sheng 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*, 2024.

Nicholas Ho, Caleb Ellington, Hou Jinyu, Sohan Addagudi, Shentong Mo, Tianhua Tao, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, Le Song, and Eric. Xing. Scaling dense representations for single cell with transcriptome-scale context. *bioRxiv*, 2024. AIDO.Cell Foundation Model.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. URL `https://distill.pub/2020/circuits/zoom-in`.

Flavia Pedrocchi, Floria Barkmann, Amir Joudaki, and Valentina Boeva. Sparse autoencoders reveal interpretable features in single-cell foundation models. *bioRxiv*, 2025.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2024. doi: 10.1101/2023.11.28.568918. URL `https://www.biorxiv.org/content/early/2024/10/06/2023.11.28.568918`.

Viktoria Schuster. Can sparse autoencoders make sense of gene expression latent variable models? *Proceedings of Machine Learning Research*, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunnigham, Nicholas Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Edward Rees, Joshua Batson, Adam Jermyn, Shan CArter, Chris Olah, Tom Henighan, and Anthropic. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Anthropic Technical Report*, 2024.

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 T. Patrick Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

A Appendix

A.1 Training and evaluation details

All SAEs were trained on an NVIDIA RTX-3090 with 24gb VRAM, for 10 epochs on an 8-cell batch size (at a 640 hidden dimension per gene on $\sim$ 19k genes per cell), with a 0.001 learning rate using cosine annealing scheduling, and optimized with Adam.

Before obtaining the GO terms for each feature with the adapative gene selection method mentioned, we filter out genes which are either 0-expressed across more than 50% of the cells, or for which their mean expression is less than 0.01 (in log1p(CPM) space). We also apply the same kind of filtering before obtaining the top 100 up and downregulated DE genes in the steering experiments. The resulting gene set is used as the background set for all GO enrichments testing. This is done to prevent noise from very low expressed genes in the dataset, which AIDO.cell has problems reconstructing (Fig S2b).

Features picked for individual steering experiments were selected from their component due to them having the least number of annotated terms, to ensure the cleanest possible signal. The contrastive steering strength vector $\boldsymbol{\alpha}$ was trained for 10 epochs on a 4-cell batch size, with a 0.01 learning rate and sparsity factor λ, and optimized with Adam.

A.2 Supplementary Figures

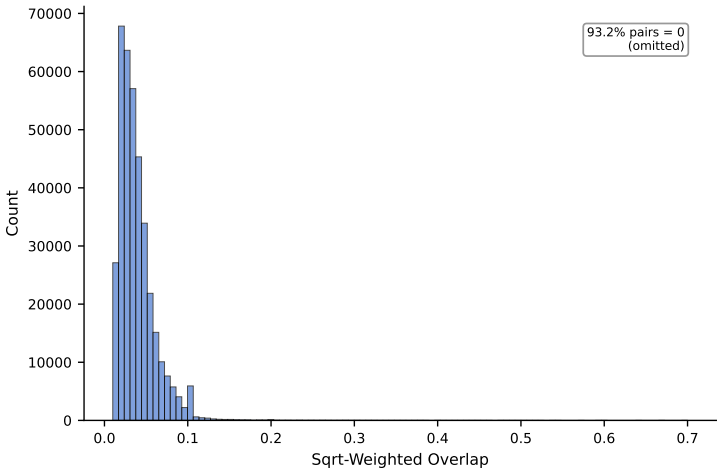
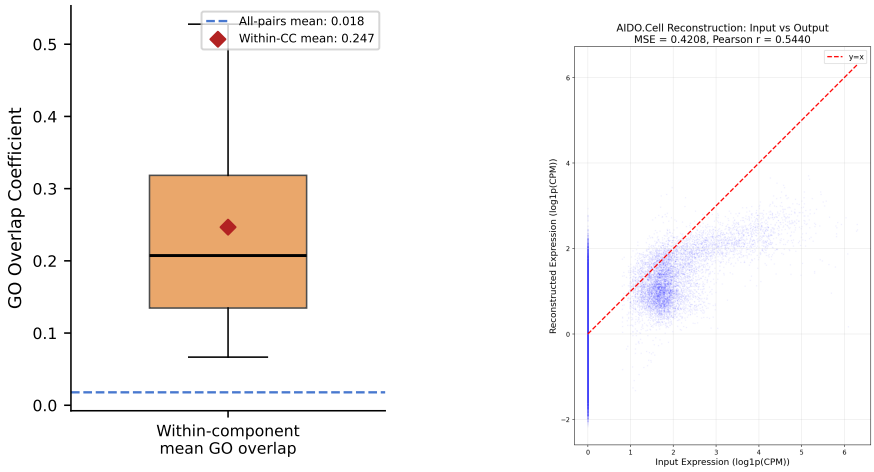


Figure S1: The gene overlap coeefficient distribution between all feature pairs. For two features i, j, if A_i, A_j are their corresponding top activating gene sets, then the coefficient is calculated as $\frac{|A_i \cap A_j|}{\sqrt{|A_i||A_j|}}$.



(a) (b)

Figure S2: **a.** The within connected component mean GO overlap distribution, which is considerably higher than the pairwise mean GO overlap **b.** Input vs output expression in AIDO.cell. The vertical line at 0 shows uncertainty in the reconstruction zero-expressed genes.

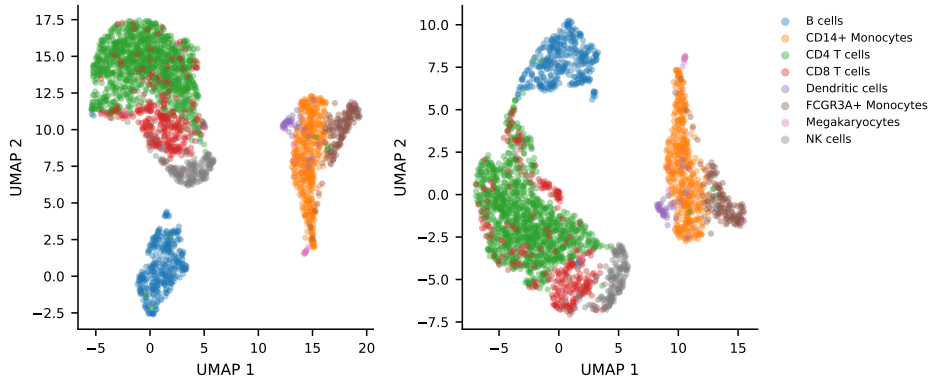
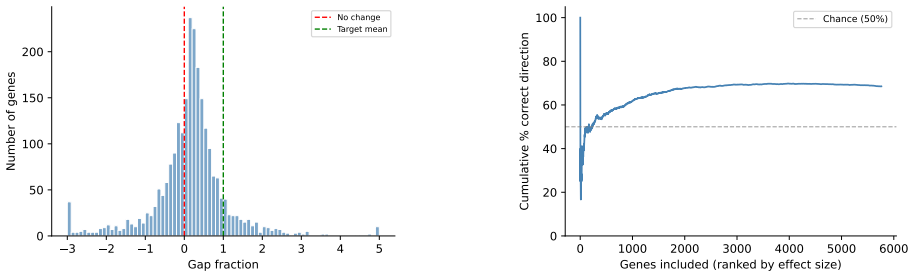


Figure S3: Left: UMAP of final cell embeddings from AIDO.cell, colored by cell-type. Right: UMAP of final cell embeddings from AIDO.cell, with CD4 T cell's being steered embeddings, which show higher overlap with CD8 T cells. This is merely illustrative, since distances between B cells and CD8 cells (both unsteered) are the same in both UMAPs.



(a) (b)

Figure S4: **a.** The gap fraction distributinon for downregulated genes in CD8s with respect to CD4s. **b.** Cumulative percentage of genes which changed expression in the right direction, ordered by difference between the mean CD8 and mean CD4 expression