Impacts of Dataset Imbalance on Single-Cell Foundation Models

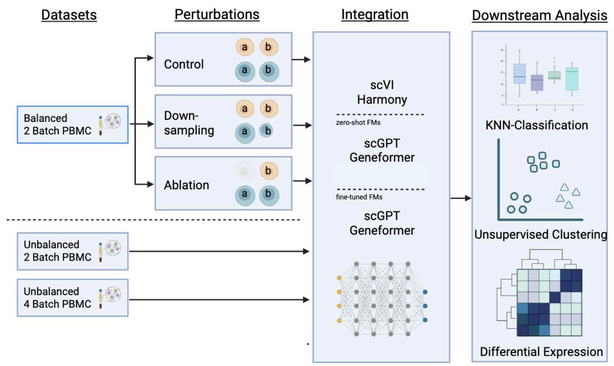
Izumi Ando[1,2], **Hassaan Maan**[*,3,4,5], **Kieran R. Campbell**[*,1,2,4,6,7,8]

*Jointly supervised this work, **1:** Lunenfeld-Tanenbaum Research Institute at Sinai Health, **2:** Department of Computer Science at the University of Toronto, **3:** Peter Munk Cardiac Centre at University Health Network, **4:** Vector Institute, **5:** Department of Medical Biophysics at the University of Toronto, **6:** Department of Molecular Genetics at the University of Toronto, **7:** Department of Statistical Sciences at the University of Toronto, **8:** Ontario Institute for Cancer Research

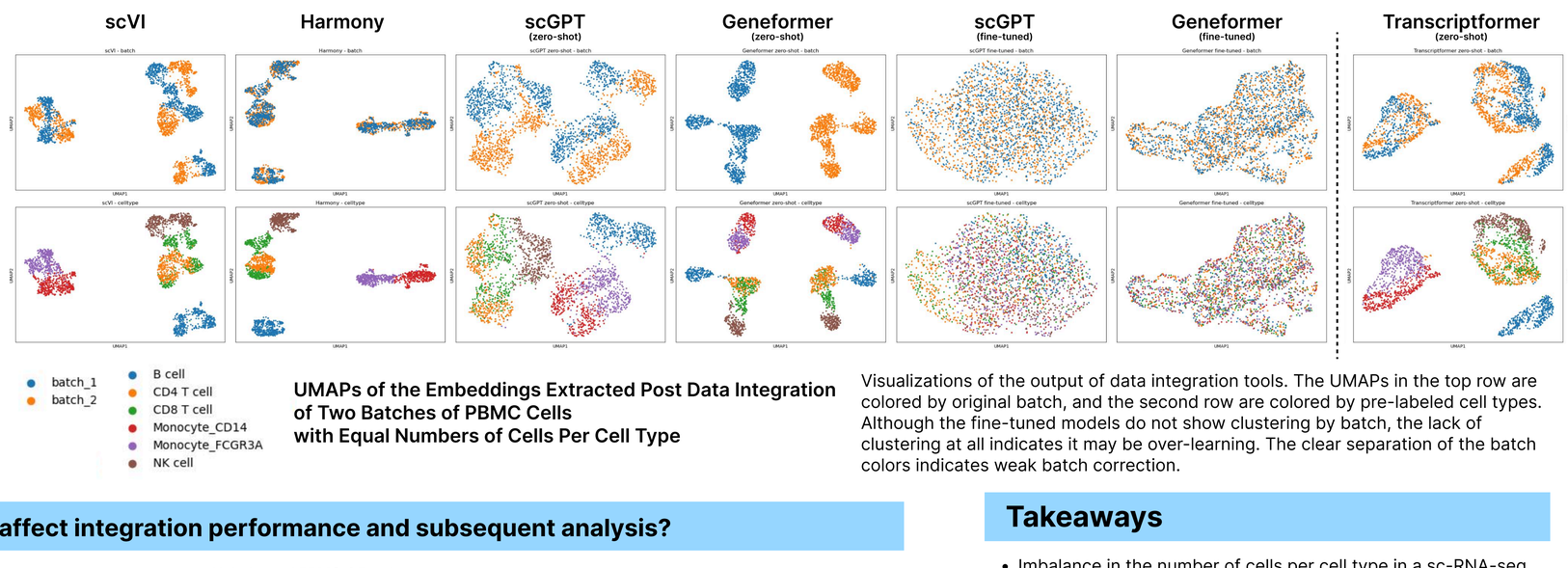
Pipeline

The goal of this project is to evaluate how novel single-cell foundation models perform compared to simpler models in handling imbalances in cell type proportions during data integration. Further, we examine the differences in biological signal that arise when imbalanced datasets are integrated using different tools.

To do so, we used the pipeline illustrated below. We employed perturbation experiments to study the impact of dataset imbalance and used inherently imbalanced datasets to study the potential differences in biological signal that arise from different integration methods. Foundation models were tested in both fine-tuned and zero-shot (non-fine-tuned) settings.



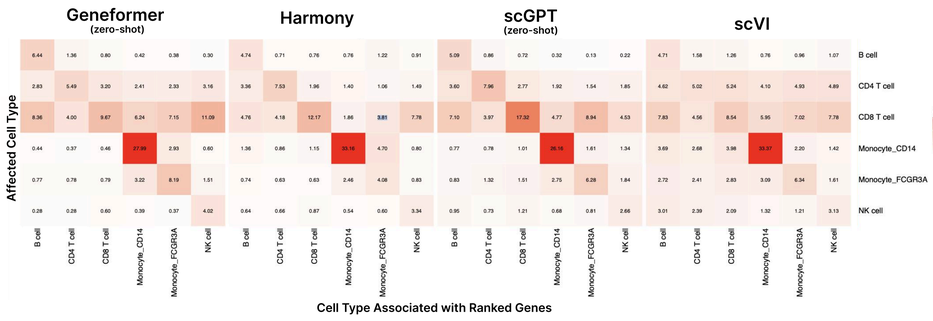
Visualizations of Integrated Embeddings



UMAPs of the Embeddings Extracted Post Data Integration of Two Batches of PBMC Cells with Equal Numbers of Cells Per Cell Type

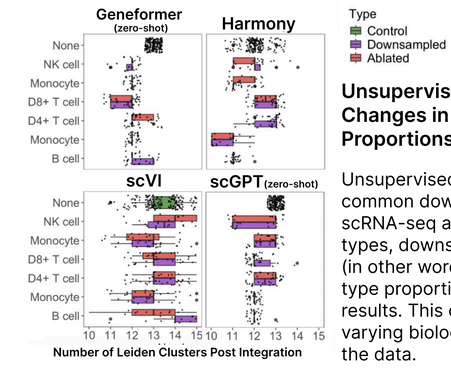
Visualizations of the output of data integration tools. The UMAPs in the top row are colored by original batch, and the second row are colored by pre-labeled cell types. Although the fine-tuned models do not show clustering by batch, the lack of clustering at all indicates it may be over-learning. The clear separation of the batch colors indicates weak batch correction.

Preliminary Results : How does imbalance in cell type proportions affect integration performance and subsequent analysis?



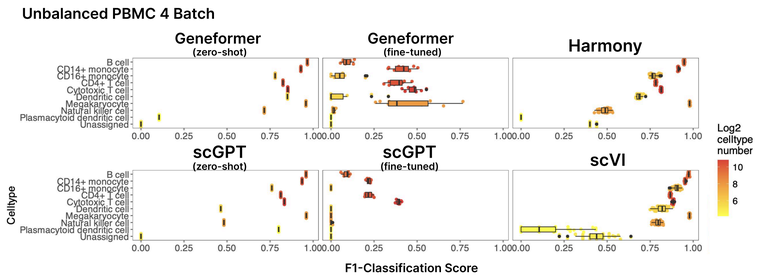
Differential Expression Analysis: Variance of Gene Rankings

It is important to understand whether imbalance in cell type proportions in integrated data sets affects biological signal in differential expression analysis. The top 10 marker genes for each cell type in the datasets being integrated were selected and the variance of those genes' rankings were computed after integration. This heat map shows that when CD14 monocytes are downsampled or ablated, their ranking can significantly vary regardless of the integration method.



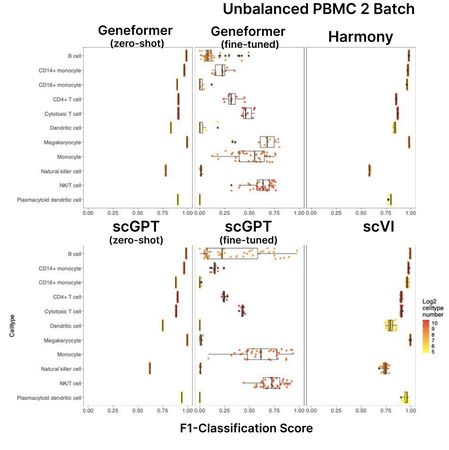
Unsupervised Clustering: Changes in Clustering When Proportions of Cell Types Shift

Unsupervised clustering is a common downstream task in scRNA-seq analysis. For most cell types, downsampling and ablation (in other words, an imbalance in cell type proportions) leads to varying results. This could potentially lead to varying biological interpretations of the data.



KNN Classification: Integration Performance

F1 scores from KNN Classification indicate integration performance. The two sets of boxplots above are from repeated integration attempts of the 2-batch unbalanced PBMC dataset and the 4-batch unbalanced PBMC dataset. The fine-tuned models show poor performance, while the other models display similar performance. In the 2-batch PBMC datasets, all models perform worse for classifying natural killer cells, while in the 4-batch PBMC dataset, plasmacytoid dendritic cells also show low scores agnostic to the integration method.



Takeaways

- Imbalance in the number of cells per cell type in a sc-RNA-seq dataset DOES affect the performance of data integration across all methods.
- Certain cell types affect integration performance, agnostic to integration method.
- Challenges in benchmarking sc-RNA-seq foundation models include the fact that most are not designed to be fine-tuned for data integration, and they each have very different interfaces.

Next Steps

- Incorporate Transcriptformer into the pipeline.
- Design a standard approach to testing fine-tuned performance of foundation models in regards to data integration.
- Incorporate more complex datasets into the pipeline. (ex: 8 batch PDAC dataset)

References

Allard, M., Putzeys, B., Schneider, R., Quiroga Londoño, M., & Klop, M. (2024). helicalAI/helical: V0.0.1-alpha (Version 0.0.1a0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.13135902

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., & Wang, B. (2024). scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods, 21(8), 1470–1480. https://doi.org/10.1038/s41592-024-02201-0

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods, 16(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nature Methods, 15(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

Maan, H., Zhang, L., Yu, C., Geuenich, M. J., Campbell, K. R., & Wang, B. (2024). Characterizing the impacts of dataset imbalance on single-cell data integration. Nature Biotechnology, 42(12), 1899–1908. https://doi.org/10.1038/s41587-023-02097-9

Pearce, J. D., Simmonds, S. E., Mahmoudabadi, G., Krishnan, L., Palla, G., Istrate, A.-M., Tarashansky, A., Nelson, B., Valenzuela, O., Li, D., Quake, S. R., & Karaletsos, T. (2025). A Cross-Species Generative Cell Atlas Across 1.5 Billion Years of Evolution: The TranscriptFormer Single-cell Model. bioRxiv preprint. https://doi.org/10.1101/2025.04.25.650731

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., & Ellinor, P. T. (2023). Transfer learning enables predictions in network biology. Nature, 618(7965), 616–624. https://doi.org/10.1038/s41586-023-06139-9