



scContrast: A contrastive learning based approach for encoding single-cell gene expression data

by
Winston Yuxiang Li

A Thesis submitted in partial fulfillment of the requirements for Honors
in the Department of Computer Science at Brown University

Providence, Rhode Island
April 2025

Acknowledgements

It is impossible to overstate my gratitude to my advisor, Professor Ritambhara Singh, for her endless generosity and tireless patience in guiding and supporting me through the short couple of years I spent at the Singh Lab. From her class, *Deep Learning in Genomics*, where my hitherto parallel passions in biochemistry and computer science were first intertwined, I was exposed to the exhilarating field of computational genomics. Such passions were allowed to flourish at the Singh Lab. Thank you, Professor Ritambhara, for providing me with this little arcadia, for all of your inspiration; for pushing me to challenge myself.

When I first expressed interest in Prof. Ritambhara's lab, amidst a salvo of potential projects, there was one that instantly caught my attention both by its boldness and by its scale: a universal, foundational model that would be able to handle every form of cellular data available. And this was but one of many brilliant ideas juggled by now Dr. Ghulam Murtaza. From Murtaza's bottomless well of knowledge, I was able to glimpse so many beautiful and creative ways of not just approaching but *thinking* of genomics; of biology and its intersections with computer science. It is without a doubt that I have been so very, very lucky to have been able to meet and work with you.

Finally, thank you to my family and friends who were always there to support me. It is your love and encouragement that keeps me motivated every day.

Contents

1	Introduction	4
2	Related Works	7
2.1	Contrastive Learning	7
2.2	Batch Integration	8
2.3	Foundational Models in Genomics	8
3	Methods	9
3.1	Augmentations Module	9
3.1.1	Cell-type Specific Augmentations	9
3.1.2	Cell-type Agnostic Augmentations	10
3.2	Cell Encoding Module	11
3.3	Contrastive Learning Module	11
3.4	Implementation details	11
4	Experimental Setup	13
4.1	Datasets	13
4.2	Baselines	13
4.3	Evaluation Metrics	14
5	Results	15
5.1	scContrast successfully integrates gene expression data from plate- and droplet-based sequencing platforms in unseen tissues	15
5.2	scContrast qualitatively exhibits biological conservation in scRNA-seq data collected from an unseen protocol following self-supervised fine-tuning	18
5.3	The cell shuffling augmentation is crucial for integrating across protocols . .	19
6	Discussion	22
6.1	Summary of Main Findings	22
6.2	Limitations	22
6.3	Future Work	23

Abstract

Single-cell RNA sequencing (scRNA-seq) captures gene expression at an individual cell resolution, which reveals critical insights into cellular diversity, disease processes, and developmental biology. However, a key challenge in scRNA-seq analysis is clustering similar cells across multiple batches, particularly when distinct sequencing protocols are used. In this work, we present scContrast, a semi-supervised contrastive learning method tailored for embedding scRNA-seq data from both plate- and droplet-based protocols into a universal representation space. By leveraging five simple augmentations, scContrast extracts biologically relevant signals from gene expression data while filtering out batch effects and technical artifacts. We trained scContrast on a subset of Tabula Muris tissues and evaluated its zero-shot performance on unseen tissues. Our results demonstrate that scContrast generalizes effectively to new tissues and outperforms the leading UCE approach in integrating scRNA-seq data from droplet- and plate-based sequencing protocols.

CHAPTER 1

Introduction

Single-cell RNA sequencing ([Jovic et al., 2022](#)) (scRNA-seq) enables researchers to profile gene expression at the level of individual cells, permitting insights into cellular heterogeneity ([Ximerakis et al., 2019](#)), developmental processes ([Qiu et al., 2022](#)), and disease mechanisms ([Mathys et al., 2019](#)). A typical scRNA-seq analysis begins by clustering similar cells and identifying key marker genes for each cluster to map out differentiation pathways or uncover disease mechanisms ([Su et al., 2022](#)). However, clustering cells based on the full gene expression profile is particularly challenging due to the high dimensionality, sparsity, and batch effects inherent to scRNA-seq data ([Bouland et al., 2023](#)).

Existing computational methods for clustering scRNA-seq data can be broadly divided into three categories. The first category includes unsupervised approaches, which typically involve a highly variable gene selection followed by dimensionality reduction methods such as Principal Component Analysis (PCA) to get low-dimensional cell-level representations that cluster similar cells together. While such approaches work well for individual datasets, these approaches struggle to integrate data from multiple batches because they cannot account for batch effects or technical artifacts ([Yu et al., 2024](#)). To resolve these challenges, a second category of methods that includes supervised approaches such as scVI ([Lopez et al., 2018](#)) requires batch labels to assign similar representations to similar cells. However, these supervised approaches need to be retrained for every new integration, making them computationally expensive and impractical for large-scale datasets. More recently, inspired by the success of large foundational models such as ChatGPT ([Brown et al., 2020](#)), LLAMA ([Touvron et al., 2023](#)), and SAM ([Kirillov et al., 2023](#)), researchers have developed similar foundational models for scRNA-seq ([Theodoris et al., 2023](#); [Yang et al., 2022](#); [Cui et al., 2024](#); [Rosen et al., 2023](#)). These foundational models are extremely large transformer-based architectures trained on atlas-scale datasets with a masked language modeling (MLM) objective. While these models promise to project arbitrary scRNA-seq data into a universal embedding space, it has been shown that these models struggle to effectively integrate data generated from different experimental protocols ([Kedzierska et al., 2023](#)).

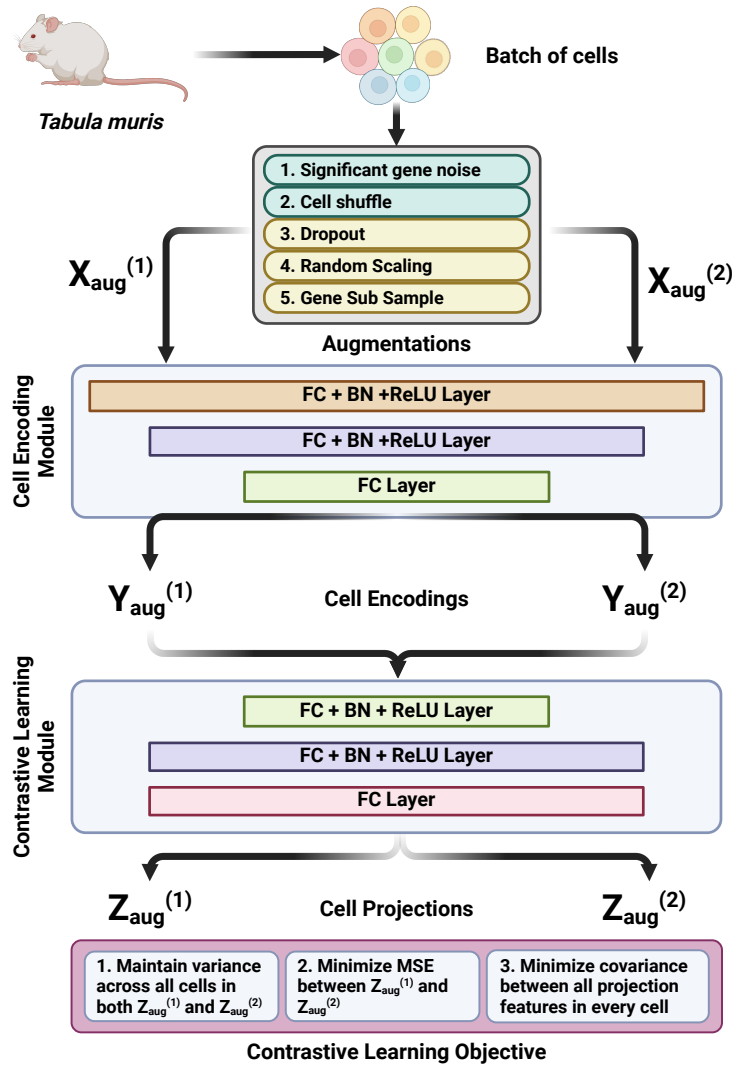


Figure 1.1: Overview of scContrast pipeline. First, given a batch of samples as a gene expression matrix, two augmented views are generated. Then, the views are encoded by a series of fully connected layers. Next, the encodings are projected into a higher dimension space by another series of fully connected layers. Finally, the projections are used in a contrastive learning objective that simultaneously maintain variance across samples, minimize MSE between augmented views, and minimize covariance between projection features.

This work introduces **scContrast**, summarized in Figure. 1.1, which is a semi-supervised contrastive learning approach designed to embed scRNA-seq data from different experimental protocols into a universal cell-level representation space. scContrast leverages a simple yet effective set of five augmentation functions, which simulate the technical artifacts, protocol biases, and batch effects characteristic of scRNA-seq assays. These augmentation functions

guide scContrast to extract biologically meaningful features of single-cell gene expression profiles.

We trained scContrast on 21 out of 25 tissue samples from the Tabula Muris dataset ([Consortium, 2018](#)), which includes data from multiple donor mice sequenced using both droplet- and plate-based sequencing platforms. To simulate a true zero-shot setting, we tested scContrast on the remaining four held-out tissues. Through our evaluations, we found that scContrast generalizes well to unseen tissue samples and outperforms the state-of-the-art UCE model in integrating batches from different sequencing protocols. Specifically, scContrast scores 0.18 and 0.73 on the kBET and PCR batch integration metrics, while UCE scores 0.02 and 0.39, respectively. Thus, our findings highlight scContrast’s superior zero-shot cross-protocol batch integration capabilities. However, scContrast performs slightly worse on KMeans ARI and ASW bio-conservation metrics compared to UCE, suggesting opportunities for improvement in the future.

CHAPTER 2

Related Works

2.1 Contrastive Learning

Major inspiration for our model’s contrastive learning approach comes from existing work originating from computer vision. The following methods all employ augmentations-based contrastive learning, albeit for image classification, which we have re-engineered to use gene-expression-specific augmentations.

simCLR. The method simCLR (Chen et al., 2020) generates two augmented views for each image in a batch. Pairs of augmented views that come from the same original image are thus called “positive pairs”, whereas those comprising augmented views derived from different images form “negative pairs.” An InfoNCE loss is then applied over all pairs within a batch. Beyond this training objective, simCLR also involved several engineering strategies such as large batch-sizes and the inclusion of a projection head prior to computing the InfoNCE loss.

Barlow Twins. The method Barlow Twins (Zbontar et al., 2021) generates an empirical cross-correlation matrix using the embeddings generated from two different distortions of an original image. The loss function then seeks to push this matrix towards the identity matrix; an *invariance term* tries to equate the diagonal elements of the feature cross-correlation matrix to 1, while a *redundancy reduction term* tries to equate the off-diagonal elements to 0, which de-correlates the different features of the embedding. Crucially, this loss function does not require negative samples and instead only focuses on the cross-correlations generated from the embedding features of each sample within a batch. Furthermore, its emphasis on *redundancy reduction* allows for larger embedding spaces by combating representation collapse.

VICReg. The method VICReg (Bardes et al., 2022) combines ideas from both simCLR and Barlow Twins with its titular, three regularization terms: variance, invariance, and covariance. Within a batch, embeddings are first generated via augmentation functions followed by encoding and projection. In the loss term, covariance across the embedding features are pushed to zero as in Barlow Twins, but the variance between different embedding samples is maintained as in the negative pairs of simCLR. Furthermore, the distance between embeddings derived from the same image is also minimized as in the positive pairs of simCLR. The mathematical formulation is given in 3.1, and this is the method we chose to explore furthest for scContrast.

2.2 Batch Integration

Below, we discuss a couple of common batch integration techniques. Note, both methods explicitly require batch labels and do not transfer well to new datasets unseen during training, thus needing to be re-trained.

scVI. scVI (Lopez et al., 2018) uses deep neural networks to model the observed gene expression of cells as a zero-inflated negative binomial distribution (ZINB) conditioned on the batch annotation plus additional Gaussians to capture residual variations. The expression of each cell is first encoded into a low-dimensional latent representation of normal random variables before being decoded by another neural network to generate a posterior estimate of the parameters associated with the distribution defining each gene for each cell.

scANVI. scANVI (Xu et al., 2021) extends scVI to use cell-type annotations to further fine-tune the model in a semi-supervised manner. Slotting in-between the latent embedding module and the ZINB parameterization of scVI, scANVI introduces an additional cell-type assignment objective, which allows the model to jointly optimize for batch harmonization and label prediction.

2.3 Foundational Models in Genomics

Several groups have sought to develop “foundational genomics models” in analogy to modern large language models. These models generally tend to employ a masked learning objective as its pre-training task. They are distinguished by their very large parameter and training data sizes.

scGPT and Geneformer. scGPT (Cui et al., 2024) applies an autoregressive approach for filling-in masked gene expressions. Geneformer (Theodoris et al., 2023) also employs a masked-language modeling approach, but the masks are filled in at once; i.e. it is not autoregressive. Moreover, Geneformer encodes gene expression by rank value, which the authors suggest helps the model be more robust to technical artifacts. Both models feature a transformer architecture and are pre-trained on over thirty million samples.

UCE. Universal Cell Embeddings (Rosen et al., 2023) project single cells into a unified latent space by sampling expressed genes and representing each gene via pre-trained protein language model. It employs a masked learning objective to predict masked gene tokens using remaining gene embeddings and learned cell embeddings. UCE is composed of 650 million parameters and is trained on a mega-scale atlas of over 36 million cells across 8 species.

CHAPTER 3

Methods

As shown in Figure 1.1, scContrast has three main components:

3.1 Augmentations Module

The augmentations module take in a batch of cells from the Tabula Muris dataset to produce two augmented views X_{aug}^1 and X_{aug}^2 in a semi-supervised fashion. For each sample, the corresponding augmented views form a positive pair. This module relies on five augmentation functions that aim to perturb the gene expression data in a manner that preserves each sample’s cellular identity. Such functions can be sub-divided into two categories. The first category of augmentation functions require cell-type annotations (highlighted in green in Figure 1.1), whereas the second category of augmentation functions does not rely on any annotations (highlighted with yellow color in the Figure 1.1). A visual summary of the augmentations is given in Figure 3.1

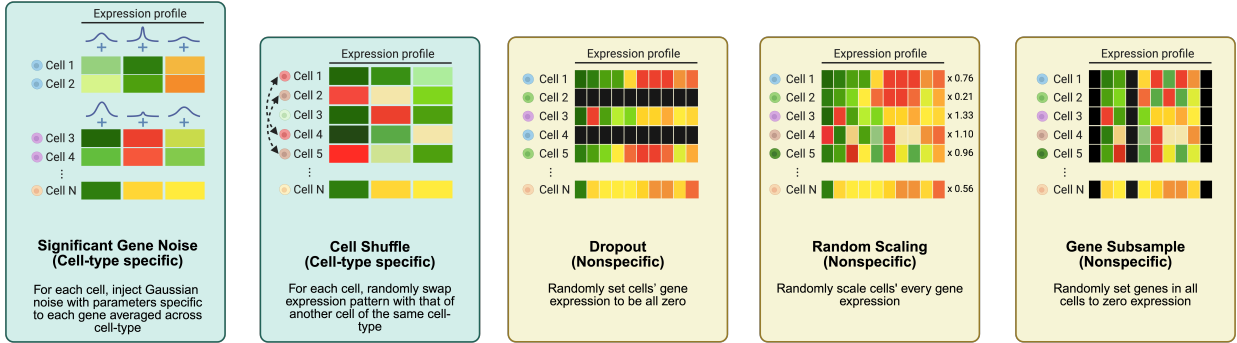


Figure 3.1: Summary of augmentations.

3.1.1 Cell-type Specific Augmentations

By leveraging the cell-type annotations of each sample, these cell-type specific augmentations ensure samples from the same cell-type maintain similar representations while challenging the model to distinguish between distinct cell-types.

Significant gene noise. This augmentation aims to noise the cell-type defining genes for each cell. “Gene significance” is determined using `sc.tl.rank_gene_groups`, which uses

a Wilcoxon rank-sum to calculate a p-value associated with how significantly each gene characterizes a cellular group, here defined as cell-types. For “most significant genes,” a threshold value of $p < 0.05$ was used to select relevant genes; for “least significant genes,” a threshold value of $p > 0.05$ was used, instead. From preliminary testing, we found that the “most significant genes” variant was more effective than its “least significant genes” counterpart in producing well integrated latent spaces.

After the significant genes are identified for each cell-type, a Gaussian across genes corresponding to the mean and variance associated with the pseudo-bulk of each cell-type is then added to the gene expression values of the relevant cell-types. This thus represents the “Gaussian noise injection” part of this augmentation.

To accelerate training, the pseudo-bulk representation of each cell-type is pre-computed and fed into the `Dataloader` as an explicit argument. Such pseudo-bulks are defined as the average and corresponding variances of the gene expressions across all cells for a particular cell-type, inclusive of both protocols; a separate μ and σ term is thus associated with each gene for each cell-type. The pseudo-bulk thus represents a “typical example” of a cell of that cell-type. This is motivated by the fact that different cells have inherently different gene expressions even outside of protocol biases, so tailoring augmentations to each cell-type permits greater control over augmentation strengths.

Cell shuffle. This augmentation is motivated by the idea that cells of the same cell-type should be considered similar regardless of which protocol they are measured by. Given a batch of cells, this augmentation loops through the unique cell-types represented in the batch and randomly swaps the gene-expression profile of cells of the same cell-type.

This augmentation proves to be very strong at overcoming batch effects (see Table. 5.2; however, the assumption that different samples of the same cells are interchangeable can be problematic if certain cell-type labels are very large, such as stem cells.

3.1.2 Cell-type Agnostic Augmentations

Dropout. This augmentation applies a zero-mask to the entire gene-expression profile of randomly selected cells. It is motivated by the idea that replicates of scRNA-seq assays will be performed on similar but never identical samples of cells, which will affect population membership. A uniform distribution is used to select cells to dropout, which means already rare cell-types are proportionally more affected than more common cell-types.

Random Scaling. This augmentation is inspired by the idea that, depending on assay conditions such as reagent sensitivity, overall expression patterns can vary. For example, suppose the DNA replication agents were partially denatured; then, measured expression during gene amplification will be reduced for all genes. As before, this augmentation is aimed to build the models robustness towards differing distribution characteristics of samples.

Gene Subsample. This augmentation can be thought of as the transpose of the dropout augmentation, where now genes are randomly selected to be zero-masked across all cells. Intuitively, this can represent genes that fail to be amplified. This also intensifies the inherent sparsity of scRNA-seq data, as there are now more zero-values in the expression matrix.

3.2 Cell Encoding Module

The core purpose of the Cell Encoding Module is to encode cells into a low-dimensional representation space that preserves biological semantic properties such as cell-type similarities, relationships, and developmental trajectories. Unlike existing scRNA-seq foundational models such as UCE (Rosen et al., 2023), Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), and scBERT (Yang et al., 2022) that have transformer blocks with self-attention, we rely on simple, fully connected neural networks to extract relationships between all genes in the cell. Specifically, we have two fully connected layers with batch normalization to stabilize the learning process and ReLU to capture non-linear relationships, followed by a single fully connected layer to produce cell encodings Y_{aug}^1 and Y_{aug}^2 . The entire Cell Encoding Module has 5.2 million parameters, which is 10100 times smaller than existing scRNA-seq foundational models (Theodoris et al., 2023; Rosen et al., 2023).

3.3 Contrastive Learning Module

The Contrastive Learning Module is responsible for comparing the two augmented encodings generated by the Cell Encoding Module. It is inspired by VICReg (Bardes et al., 2022) and aims to ensure **V**ariance, **I**nvariance, and **C**ovariance among each batch of projected sample encodings to further enrich latent space representation. We use the same encoder and projection head for each augmented view.

More specifically, outputs from the Cell Encoding Module are first projected into $Z_{aug}^{(1)}$ and $Z_{aug}^{(2)}$, respectively, using a lightweight, trainable projection head. Then, a loss comprising three terms are calculated from these projected augmentations as shown in the equation below, where X and Y are $Z_{aug}^{(1)}$ and $Z_{aug}^{(2)}$, respectively:

$$\mathcal{L}_{VICReg}(X, Y) = \underbrace{\mu \left(\ell_{\text{var}}(X) + \ell_{\text{var}}(Y) \right)}_{\text{variance}} + \underbrace{\lambda \|X - Y\|_2^2}_{\text{invariance}} + \underbrace{\nu \left(\ell_{\text{cov}}(X) + \ell_{\text{cov}}(Y) \right)}_{\text{covariance}},$$

where the variance term is $\ell_{\text{var}}(V) = \frac{1}{d} \sum_{j=1}^d \max\left(0, \gamma - \sqrt{\text{Var}(V_{:,j})} + \epsilon\right)$, the invariance term is $\|X - Y\|_2^2 = \frac{1}{N} \sum_{i=1}^N \|x_i - y_i\|_2^2$, and the covariance term is $\ell_{\text{cov}}(V) = \frac{1}{d} \sum_{i \neq j} [C(V)]_{i,j}^2$, such that $V \in \{X, Y\}$, γ is the target standard deviation, ϵ is a small constant (e.g., 10^{-4}) for numerical stability, and $C(V)$ is the empirical covariance matrix of embeddings V , which is squared and summed over all off-diagonal elements. Coefficients μ , λ , and ν are tunable hyperparameters that weigh each loss term.

3.4 Implementation details

We implemented the entire pipeline in Python==3.9.16 using Pytorch==2.5.1+cu118 and Pytorch-Lightning==2.4.0. scContrast produces cell encodings with 128 dimensions and

cell projections with 256 dimensions. We use the cell encodings Y for all of our evaluations and visualizations. We use cell projections Z_{aug}^1 and Z_{aug}^2 to calculate the VICReg ([Bardes et al., 2022](#)) contrastive loss with a batch size of 8192. We employ an AdamW optimizer with default parameters to calculate the gradients used to update all the weights.

CHAPTER 4

Experimental Setup

4.1 Datasets

We obtained scRNA-seq data from the Tabula Muris dataset, which is a comprehensive collection of single-cell transcriptomics data generated from both droplet-based (10x Genomics (Zheng et al., 2017)) and plate-based (Smart-Seq2 (Picelli et al., 2014)) sequencing platforms for *Mus musculus* (mouse). The Tabula Muris dataset contains 25 distinct tissues and has more than 350,000 cells. In order to evaluate the performance of scContrast on out-of-distribution data in a zero-shot setting, we randomly selected four of the twenty-five tissues for testing and kept the rest for training scContrast.

We preprocessed all datasets by first filtering out all cells with ‘UKWN’/NULL annotations. Next, we normalized gene expressions of each cell by scaling the total counts to 10000. Finally, we applied a log-transformation to minimize technical artifacts (Su et al., 2022). Note, while we do compute the highly variable genes (HVGs) associated with a dataset, we do not subset our data to just those HVGs as is commonly done in other methods such as scVI (Lopez et al., 2018); rather, we used the computed dispersions as the variance input for our Gaussian noise injection augmentations.

4.2 Baselines

In order to comprehensively compare the performance of scContrast against existing approaches, we include:

- (a) **PCA**: A fully unsupervised method that decomposes the cell-by-gene matrix into its principal components, producing low-dimensional cell representations.
- (b) **scVI** (Lopez et al., 2018): A supervised batch-integration approach that explicitly requires batch annotations. We train scVI with `n_layers=3` and `n_latent=128` to match the size and shape of the Cell Encoding Module of scContrast.
- (c) **UCE** (Rosen et al., 2023): A scRNA-seq foundational model with over 650 million parameters trained on an expansive corpus of scRNA-seq dataset containing over 36 million cells and spanning across eight diverse species. UCE can reportedly integrate scRNA-seq datasets with batch effects across unseen tissues and species without any finetuning. We use UCE with its default configurations without any finetuning to generate cell-level representations.

4.3 Evaluation Metrics

We evaluate the quality of cell encodings across bio-conservation and batch integration metrics defined by the `scib-metrics` package (Luecken et al., 2021), which is commonly used for benchmarking scRNA-seq models. For the sake of conciseness, we show two out of five metrics for each category. The rest of the scores can be found in the Supplementary (Table 6.1, Table 6.2).

We present the following four metrics:

- (a) **Bio-conservation metric ARI**: Calculates the Adjusted Rand Index from the KMeans-clustered embedding space of each model.
- (b) **Bio-conservation metric ASW**: Calculates the average silhouette width with respect to cell-type.
- (c) **Batch integration metric kBET**: Conducts the k -nearest-neighbor batch effect test, which rewards neighborhoods of embeddings having the same distribution of batch labels as that of the full dataset.
- (d) **Batch integration metric PCR**: Performs a principal component regression comparison by comparing the explained variance before and after integration.

CHAPTER 5

Results

5.1 scContrast successfully integrates gene expression data from plate- and droplet-based sequencing platforms in unseen tissues

scContrast is distinguished from existing approaches by its comparatively light model size, small training set, lack of explicit batch labels during training, and most importantly, its ability to generalize to unseen tissue samples in a zero-shot manner. For baselines, we choose PCA, scVI, and UCE to represent existing unsupervised, fully-supervised, and large-scale foundational model paradigms that are currently used for batch integration, respectively. For evaluation, we use four held-out tissues from Tabula Muris: skin, liver, pancreas, and limb muscle. These tissues feature diverse cellular landscapes ranging from organ-specific cells to common immune cells such (e.g. macrophages and resident variants). In order to visualize the quality of the cell representation spaces, we use UMAP ([McInnes et al., 2020](#)), which is a commonly employed visualization method. Moreover, we quantified these representations using ARI, ASW, kBET, and PCR.

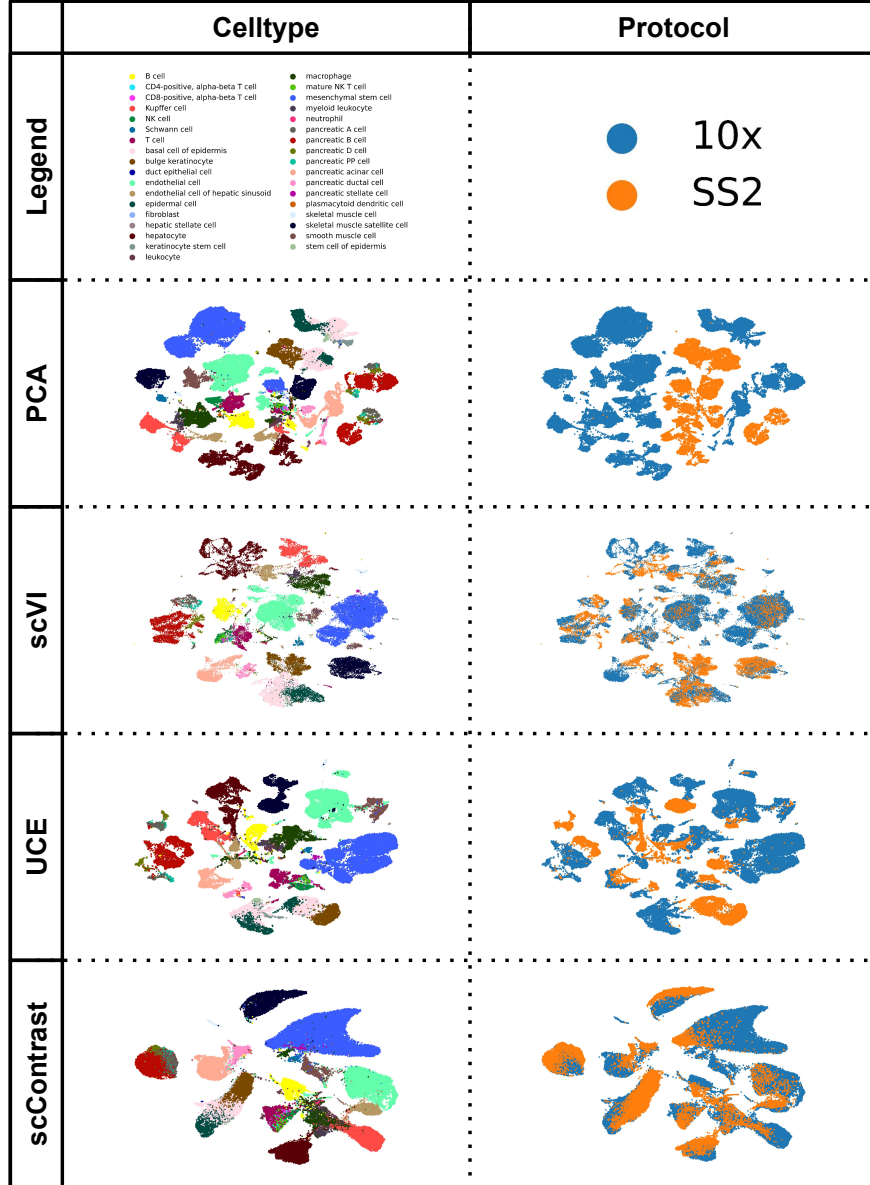


Figure 5.1: Qualitative comparisons of different models integrating scRNA-seq samples from plate (SS2)- and droplet (10x)-based sequencing platforms in unseen tissues in Tabula Muris

As shown in Figure 5.1, we find that both PCA and UCE struggle to integrate scRNA-seq data from droplet- and plate-based sequencing protocols. Specifically, from the UMAP visualizations, we observe that cells of the same cell-type but different sequencing protocol form very distinct and dense clusters in PCA; those of UCE tend to be closer to each other, but they still do not integrate to form homogeneous clusters. In contrast, both scVI and scContrast successfully integrate across sequencing protocols: the plate-based SS2-colored embeddings (orange) overlay atop the droplet-based 10X-colored embeddings (blue), highlighting that

representations of the same cell-type across both protocols homogeneously cluster together.

Model	Bio-Conservation		Batch Integration	
	KMeans ARI	ASW	kBET	PCR
PCA	0.42	<u>0.60</u>	0.00	0.00
scVI	0.40	0.56	0.25	0.93
UCE	0.51	0.65	0.02	0.39
scContrast	<u>0.43</u>	<u>0.60</u>	<u>0.18</u>	<u>0.73</u>

Table 5.1: Quantitative comparison of different models on bio-conservation and batch integration of unseen tissues from Tabula Muris. The best score is bolded and the second best scores are underlined.

Quantitatively, UCE performs best on the bio-conservation metrics, highlighting its capacity to cluster cells of the same type closer to each other in the cellular representation space. However, the batch integration metrics reflect findings in Figure. 5.1, showing that both PCA and UCE perform worse than scContrast and scVI in integrating samples from different sequencing protocols.

While these results show that our contrastive learning-based approach is promising in learning a universal cell representation space, there is potential to further refine our approach to learn better representation spaces.

5.2 scContrast qualitatively exhibits biological conservation in scRNA-seq data collected from an unseen protocol following self-supervised fine-tuning

Using the same scContrast model as before, which was previously trained on a subset of the Tabula Muris dataset, we then explored scContrast’s ability to be transferred to a fundamentally different dataset, HiRES (Liu et al., 2023), which is a simultaneous single-cell Hi-C and RNA-seq dataset of post-implantation mouse embryos. Most importantly, HiRES uses a different protocol compared to the droplet- and plate-based assays featured in the Tabula Muris dataset that scContrast was trained on.

We fine-tune the previously trained scContrast model on the new HiRES data in a self-supervised fashion; under the same VICReg training regime, this time we use only cell-type agnostic augmentations. Thus, unlike scVI which likewise requires re-training but also explicitly cell-type annotations, scContrast can instead be fine-tuned in a self-supervised manner as the only input for training are each cell’s gene expressions themselves.

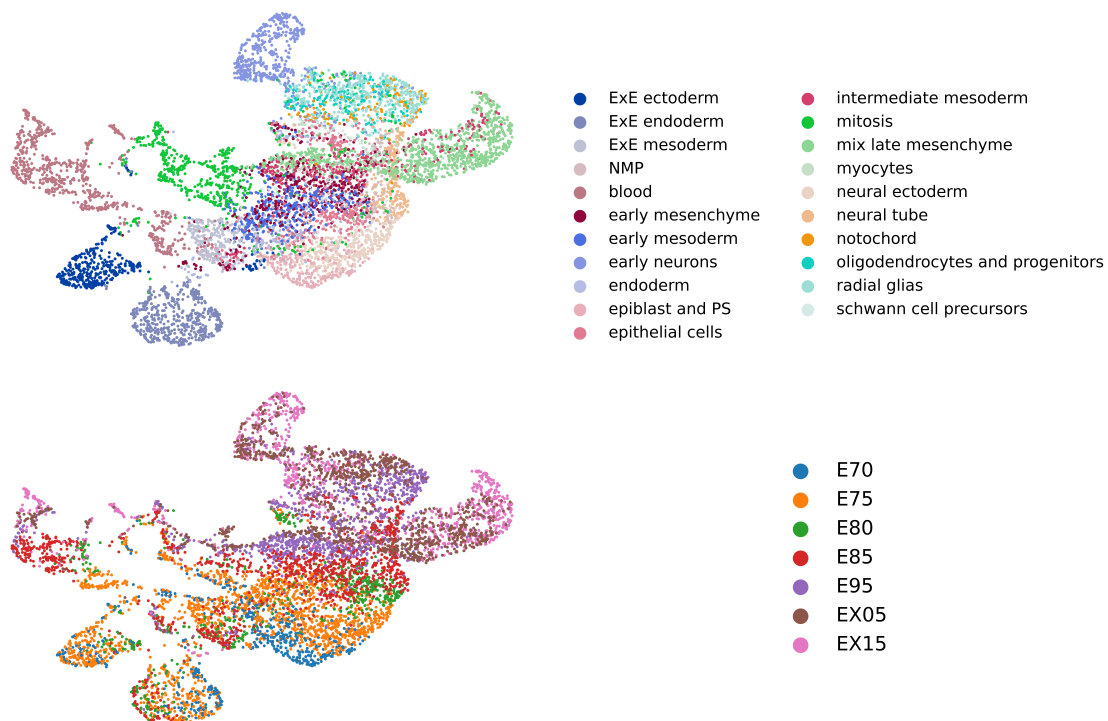


Figure 5.2: scContrast transfer performance on HiRES single-cell RNA expression data. **Top left:** UMAP colored on cell-type. **Top right:** legend of cell-types. **Bottom Left:** UMAP colored on developmental stage. **Bottom Right:** legend of developmental stages

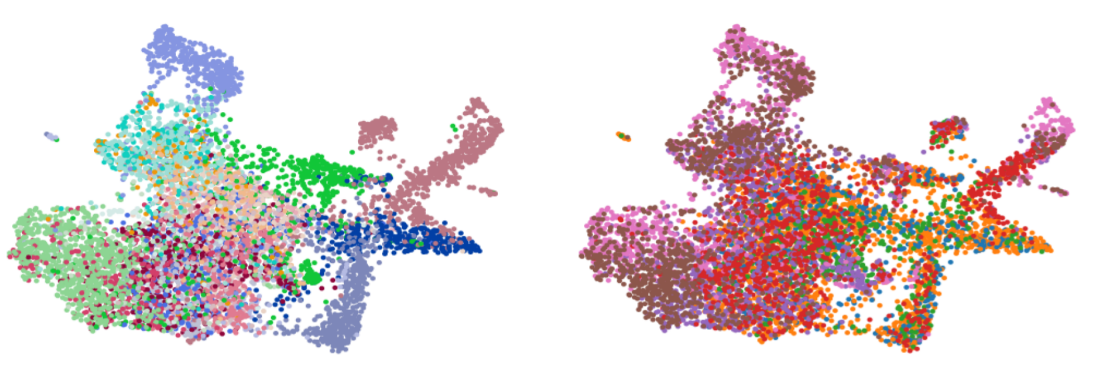


Figure 5.3: UCE zero-shot performance HiRES single-cell RNA expression data. **Left:** UMAP colored on cell-type. **Right:** UMAP colored on developmental stage.

From Figure 5.2, although we observe overall inhomogeneous clustering based on cell-type, there are certain cell-types that do form distinct clusters, namely blood cells, ExE ectoderm, ExE endoderm, ExE mesoderm, and mitosis cells. Furthermore, there is a general trend of subclustering based on developmental trajectories. This is especially present in blood cells (their subclustering follows from E70 to EX15 sequentially).

Nevertheless, the qualitatively inhomogeneous clusters suggest scContrast has room for improvement to more strongly distinguish between different cell-types provided they are truly dissimilar.

In comparison, Figure 5.3 qualitatively shows that UCE produces more inhomogeneous clusters, where cells of distinct cell-types tend to form overlapping clusters. Focusing on the blood cells, we see that there is some subclustering based on developmental trajectory, though only clearly for E85 and beyond samples.

5.3 The cell shuffling augmentation is crucial for integrating across protocols

To determine the contributions of each augmentation to overall latent representation quality, an ablations study was performed with results tabulated in Table 5.2. In general, removing cell shuffling expectedly had the greatest impact across all batch integration methods, though it did perform better on certain bio-conservation methods. Removing the significant genes noise injection also negatively impacted performance on batch integration methods, which makes sense given it weakens the contribution of the “positive-pair” dependent invariance term of the VICReg loss, as the model is no longer pushed to have softer understandings of cell-types.

Model	Bio-Conservation		Batch Integration	
	KMeans ARI	ASW Labels	kBET	PCR
All Augmentations	0.39	<u>0.60</u>	0.16	0.75
No Cell Shuffle	0.57	0.55	0.00	0.00
No Significant Genes	0.39	0.58	0.07	0.60
No Dropout	0.39	<u>0.60</u>	0.17	0.62
No Random Scaling	<u>0.44</u>	0.62	0.21	0.65
No Gene Subsample	0.37	0.59	<u>0.19</u>	<u>0.71</u>

Table 5.2: Quantitative comparison of scContrast with different augmentation functions ablated on bio-conservation and batch integration of unseen tissues from Tabula Muris.

The best scores are bolded and the second best scores are underlined.

Qualitatively, Figure 5.4 reaffirms the importance of the cell shuffling augmentation, for its loss leads to essentially the complete separation of the two protocols; there is minimal overlap between 10x and SS2 samples. Without the significant gene augmentation, there is also separation of the protocols to a lesser degree. Lastly, individual removal of the cell-type agnostic augmentations (dropout, random scaling, gene subsampling) tends to impact cluster homogeneity and separation.

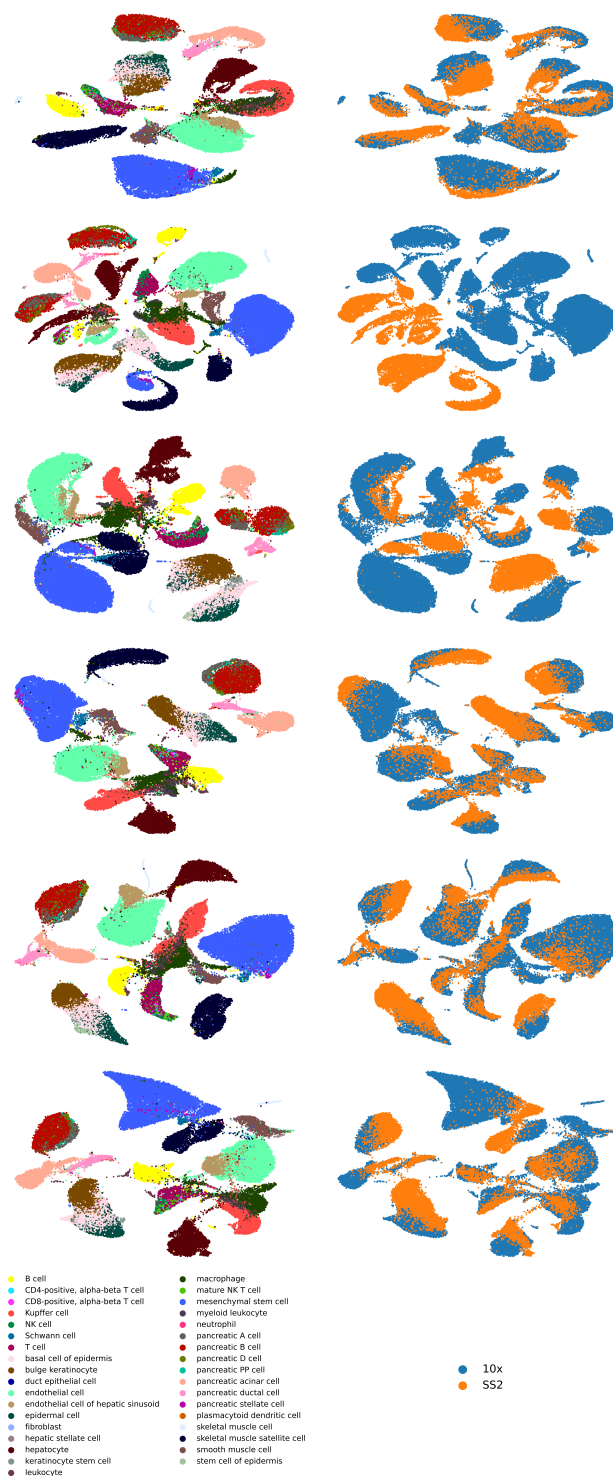


Figure 5.4: **Left:** UMAP colored on cell-type. **Right:** UMAP colored on tech. From top to bottom: all augmentations, no cell shuffling, no significant genes, no dropout, no random scaling, no gene subsampling, legend.

CHAPTER 6

Discussion

6.1 Summary of Main Findings

In this work, we have demonstrated a semi-supervised contrastive learning-based approach scContrast for encoding single-cell RNA-sequencing data into a universal cell representation space. We compared our model against PCA, scVI, and UCE, which exemplify current unsupervised, fully supervised, and large foundational model approaches. Results demonstrate that PCA understandably struggles to integrate the batches, since protocol-inherent gene expression biases likely produce spurious axes of variability, which is exacerbated by PCA’s reliance on covariances. More interestingly, UCE also struggles, which may reflect the limitations of a purely masked modeling approach. Finally, while scVI performs comparatively best at cross-protocol integration, it explicitly requires batch labels, which scContrast does not.

Additionally, scContrast features an interesting degree of biological conservation when fine-tuned on a dataset collected from an unseen protocol in a self-supervised manner. This manifests in subclustering based upon development trajectories, particularly in blood cells.

Lastly, ablation experiments suggest that the cell shuffling augmentation plays a major role in the successful integration of the different protocols. Without it, qualitative and quantitative results demonstrate a near complete separation of droplet-assayed samples from plate-assayed samples despite both being of the same cell-type.

6.2 Limitations

Although scContrast appears to perform reasonably well on batch integration metrics, quantitative and qualitative observations from Table 5.1 and Figure 5.1 suggest scContrast produces relatively diffuse clusters compared to the supervised method scVI. This is especially apparent in mesenchymal stem cells from the UMAP representations in Figure 5.1, where PCA and scVI tends to form two, distinct clusters, whereas scContrast coalesces all mesenchymal stem cell embeddings into a single, arrowhead-shaped cluster. We reason that this may be a drawback of relying on cell shuffling as the main protocol-integration augmentation, as random permutations of cells within a cell-type can destroy sub-cell-type hierarchical semantics.

Furthermore, our current bio-conservation seek only to separate clusters of cell-types from one another. This fails to capture richer biological information such as developmental

trajectories, or the fact that certain cell-types should be more similar to one another than others. Measuring such a degree of bio-conservation would require either more informed metrics (i.e. those that can refer to cell-type-interactions from an external atlas) or more informed datasets (i.e. they must include potentially a graph outlining cell-type-to-cell-type relationships). Such improved metrics can counter potential over-separation of representations.

6.3 Future Work

In the future, we aim to extend scContrast across three dimensions. First, we seek to develop better augmentation functions to closer simulate batch effects and technical artifacts to improve scContrast’s zero-shot capabilities. One possible augmentation would be to project cells to another protocol’s representation using a pseudo-bulk method. Another augmentation may involve simulating noise injections after a zero-inflated noise binomial distribution as in scVI (Lopez et al., 2018) instead of simple Gaussians.

Second, we plan to scale up our training datasets and explore deeper architectures to learn richer cell representations. Furthermore, we are also interested in the capability of scContrast to produce more consistent latent spaces across identical models trained with different seeds. Compared to methods like VAE, we reason that the “positive” and “negative pairs” in the contrastive learning objective anchors sample-to-sample relationships, which may influence more concordant latent codes across different training routines.

Lastly, we intend to investigate techniques like $\mathbb{X}$ -sample Contrastive Learning (Sobal et al., 2024) to explicitly impose hierarchical cell ontologies in the representation space.

Bibliography

- Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Singlecell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12(3), Mar 2022. doi: 10.1002/ctm2.694.
- Methodios Ximerakis, Scott L. Lipnick, Brendan T. Innes, Sean K. Simmons, Xian Adiconis, Danielle Dionne, Brittany A. Mayweather, Lan Nguyen, Zachary Niziolek, Ceren Ozek, and et al. Single-cell transcriptomic profiling of the aging mouse brain. *Nature Neuroscience*, 22(10):16961708, Sep 2019. doi: 10.1038/s41593-019-0491-3.
- Chengxiang Qiu, Junyue Cao, Beth K. Martin, Tony Li, Ian C. Welsh, Sanjay Srivatsan, Xingfan Huang, Diego Calderon, William Stafford Noble, Christine M. Disteché, and et al. Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nature Genetics*, 54(3):328341, Mar 2022. doi: 10.1038/s41588-022-01018-x.
- Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z. Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, and et al. Single-cell transcriptomic analysis of alzheimers disease. *Nature*, 570(7761):332337, May 2019. doi: 10.1038/s41586-019-1195-2.
- Min Su, Tao Pan, Qiu-Zhen Chen, Wei-Wei Zhou, Yi Gong, Gang Xu, Huan-Yu Yan, Si Li, Qiao-Zhen Shi, Ya Zhang, and et al. Data analysis guidelines for single-cell rna-seq in biomedical studies and clinical applications. *Military Medical Research*, 9(1), Dec 2022. doi: 10.1186/s40779-022-00434-8.
- Gerard A. Bouland, Ahmed Mahfouz, and Marcel J. Reinders. Consequences and opportunities arising due to sparser single-cell rna-seq datasets. *Genome Biology*, 24(1), Apr 2023. doi: 10.1186/s13059-023-02933-w.
- Ying Yu, Yuanbang Mai, Yuanting Zheng, and Leming Shi. Assessing and mitigating batch effects in large-scale omics studies. *Genome Biology*, 25(1), Oct 2024. doi: 10.1186/s13059-024-03401-9.
- Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):10531058, Nov 2018. doi: 10.1038/s41592-018-0229-2.

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothe Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollr, and Ross Girshick. Segment anything, 2023. URL <https://arxiv.org/abs/2304.02643>.
- 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 et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965): 616624, May 2023. doi: 10.1038/s41586-023-06139-9.
- 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):852866, Sep 2022. doi: 10.1038/s42256-022-00534-z.
- Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. Scgpt: Toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):14701480, Feb 2024. doi: 10.1038/s41592-024-02201-0.
- Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pages 2023–11, 2023.
- Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. *Assessing the limits of zero-Shot Foundation models in single-cell biology*, Oct 2023. doi: 10.1101/2023.10.16.561085.
- The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367372, Oct 2018. doi: 10.1038/s41586-018-0590-4.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL <https://arxiv.org/abs/2002.05709>.

- Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. URL <https://arxiv.org/abs/2103.03230>.
- Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. 2022. URL <https://arxiv.org/abs/2105.04906>.
- Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. Probabilistic harmonization and annotation of singlecell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17:e9620, 2021. doi: 10.15252/msb.20209620.
- Grace X. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, and et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), Jan 2017. doi: 10.1038/ncomms14049.
- Simone Picelli, Omid R Faridani, sa K Bjrkklund, Gsta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature Protocols*, 9(1): 171181, Jan 2014. doi: 10.1038/nprot.2014.006.
- Malte D. Luecken, M. Bttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colom-Tatch, and et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):4150, Dec 2021. doi: 10.1038/s41592-021-01336-8.
- Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL <https://arxiv.org/abs/1802.03426>.
- Zhiyuan Liu, Yujie Chen, Qimin Xia, Menghan Liu, Heming Xu, Yi Chi, Yujing Deng, and Dong Xing. Linking genome structures to functions by simultaneous single-cell hi-c and rna-seq. *Science*, 380:1070–1076, Jun 2023. doi: 10.1126/science.adg3797.
- Vlad Sobal, Mark Ibrahim, Randall Balestrieri, Vivien Cabannes, Diane Bouchacourt, Pietro Astolfi, Kyunghyun Cho, and Yann LeCun. X-sample contrastive loss: Improving contrastive learning with sample similarity graphs, 2024. URL <https://arxiv.org/abs/2407.18134>.

Supplementary

Supplementary metrics from `scib-metrics` are given in table 6.1 and table 6.2. Our scContrast model’s architecture is given here [here](#).

Model	Bio-Conservation				
	Isolated Labels	KMeans NMI	KMeans ARI	ASW Labels	cLISI kNN
PCA	0.67	<u>0.76</u>	0.42	<u>0.60</u>	1.00
scVI	0.67	<u>0.76</u>	0.40	0.56	1.00
UCE	0.67	0.80	0.51	0.65	1.00
scContrast	<u>0.62</u>	0.75	<u>0.43</u>	<u>0.60</u>	1.00

Table 6.1: Additional quantitative comparison of different models on bio-conservation metrics of representation spaces of cells from unseen tissues from Tabula Muris. The best scores are bolded and the second best scores are underlined.

Model	Batch Integration				
	ASW Batch	iLISI kNN	kBET	Graph Connectivity	PCR
PCA	0.54	0.00	0.00	0.76	0.00
scVI	0.85	0.02	0.25	0.92	0.93
UCE	0.73	0.00	0.02	0.79	0.39
scContrast	<u>0.77</u>	<u>0.01</u>	<u>0.18</u>	<u>0.80</u>	<u>0.73</u>

Table 6.2: Additional quantitative comparison of different models on batch integration metrics of representation spaces of cells from unseen tissues from Tabula Muris. The best score is bolded and the second best scores are underlined.

Model	Bio-Conservation				
	Isolated Labels	KMeans NMI	KMeans ARI	ASW Labels	cLISI kNN
All Augmentations	<u>0.61</u>	0.72	0.39	<u>0.60</u>	1.00
No Cell Shuffle	0.66	<u>0.73</u>	0.57	0.55	1.00
No Significant Genes	0.59	<u>0.73</u>	0.39	0.58	1.00
No Dropout	0.60	<u>0.73</u>	0.39	<u>0.60</u>	1.00
No Random Scaling	<u>0.61</u>	0.74	<u>0.44</u>	0.62	1.00
No Gene Subsample	<u>0.61</u>	0.72	0.37	0.59	1.00

Table 6.3: Additional quantitative comparison of scContrast with different augmentation functions ablated on bio-conservation metrics of representation spaces of cells from unseen tissues from Tabula Muris. The best scores are bolded and the second best scores are underlined.

Model	Batch Integration				
	ASW Batch	iLISI kNN	kBET	Graph Connectivity	PCR
All Augmentations	0.75	0.00	0.16	0.78	0.75
No Cell Shuffle	0.26	0.00	0.00	0.72	0.00
No Significant Genes	0.66	0.00	0.07	0.80	0.60
No Dropout	0.78	0.00	0.17	<u>0.79</u>	0.62
No Random Scaling	0.76	0.00	0.21	0.80	0.65
No Gene Subsample	<u>0.77</u>	0.00	<u>0.19</u>	<u>0.79</u>	<u>0.71</u>

Table 6.4: Additional quantitative comparison of scContrast with different augmentation functions ablated on batch integration metrics of representation spaces of cells from unseen tissues from Tabula Muris. The best score is bolded and the second best scores are underlined.

```

scRNASeqE_VICRegExpander(
  (encoder): scRNASeqEncoder(
    (encoder): Sequential(
      (0): Linear(in_features=20116, out_features=256, bias=True)
      (1): BatchNorm1d(256, eps=1e-05, momentum=0.1, affine=True,
        track_running_stats=True)
      (2): ReLU()
      (3): Linear(in_features=256, out_features=128, bias=True)
      (4): BatchNorm1d(128, eps=1e-05, momentum=0.1, affine=True,
        track_running_stats=True)
      (5): ReLU()
      (6): Linear(in_features=128, out_features=128, bias=True)
    )
  )
  (projector): scRNASeqProjectionHeadExpander(
    (projectionHead): Sequential(
      (0): Linear(in_features=128, out_features=256, bias=True)
      (1): BatchNorm1d(256, eps=1e-05, momentum=0.1, affine=True,
        track_running_stats=True)
      (2): ReLU()
      (3): Linear(in_features=256, out_features=256, bias=True)
      (4): BatchNorm1d(256, eps=1e-05, momentum=0.1, affine=True,
        track_running_stats=True)
      (5): ReLU()
      (6): Linear(in_features=256, out_features=256, bias=True)
    )
  )
  (vicreg_loss_fn): VICRegLoss()
)

```