Final Report: Single Cell Data Modeling

Aditi Garg and **Shruti Shelke**
Manning College of Information and Computer Sciences
University of Massachusetts Amherst

1 Introduction

1.1 Task Description

Insight into single cell genomics enables biological researchers to make strides in understanding the human body and treating diseases at an unprecedented scale. A necessary tool for such a detailed view of human tissues are reference atlases of organs, which are built by integrating biological datasets curated from diverse sources under varying experimental conditions. The differing conditions during the collection of each dataset significantly alter its properties by introducing challenging-to-discern non-biological batch effects that make integration of these datasets for large scale analysis a non-trivial task. (Sikkema et al., 2022)

In this project, we collaborate with the Chan Zuckerberg Initiative (CZI) (cha, 2023) and the Human Cell Atlas (hum) to compare various machine learning architectures that learn latent cell representations for two datasets - Integrated Lung Cell Atlas (ILCA) (lun) and Tabula Sapiens (TS) (cel). We then use these representations for cell-type annotation and analyze the results.

1.2 Motivation and Limitations of Existing Work

There has been steady ongoing research to perform integration of datasets and annotation for cell types, exponentially so as technological advances have allowed large scale data collection under diverse conditions. However, generalization of ML models to perform cell-type annotation across species and tissues remains a hurdle, more so with varying datasets due to the high dimensionality relations between cell type and genes as well as batch effects from different assay types.

While deep learning models such as MARS (Brbić et al., 2020) have been employed to perform cell annotation for mice for different organs, it has not been used on human datasets such as Tabula Sapiens (cel), nor has its performance been tested on single organ human datasets such as Integrated Lung Cell Atlas (lun). On the other hand, existing work has been used to generate localized atlases of lungs (Sikkema et al., 2022), but has not been translated to cross-tissue annotation. Moreover, existing works have low tolerance for counteracting batch effects that are ubiquitous in biological datasets, which in turn affects the model's performance. Furthermore, some transformer based models have been used for cell-type annotation such as scBERT (Wang et al., 2021) and scFormer (Cui et al., 2022). However, these models have not been evaluated on our datasets, which can be useful in uncovering important insights into our datasets.

In our research project, we intend to analyze and tweak models that learn effective universal cell representations for efficient performance on cell-type annotation across human tissues while increasing robustness to batch noise and retaining biological variances.

1.3 Approach

The approach we initially planned to follow was -

- Explore the datasets Integrated Lung Cell Atlas (ILCA) (lun) and Tabula Sapiens (TS) (cel).
- On the ILCA dataset, find latent representations using methods like PCA, scVI and scBERT and use different clustering metrics to study the effectiveness of these representations.
- Establish all the baselines for cell type annotation task on ILCA, i.e, PCA with logistic regression and random forest, scVI with logistic regression and random forest, and scBERT.
- Establish the same baselines for the Tabula Sapiens dataset.

- Use the representations to perform an additional downstream task of imputation (missing gene expression values in the datasets) and note performance.
- Tweak the existing baseline models to improve clustering metrics of the representations as well as performance on both the downstream tasks.

However, while establishing these baselines on our datasets for cell-type annotation, we noticed some inconsistencies in the datasets as well as some interesting observations in the latent representations clustering. Hence, we focused our attention on further analysing cell-type annotation for different train-test splits and obtaining clustering results on various features like gender, age, smoking habits etc instead of fine-tuning for an additional downstream task.

1.4 Challenges and Mitigations

Unlike ML datasets that tend to be quite large and are usually generated from a single experimental condition, biological data is collected in the form of small datasets generated under different conditions. Due to the nature of biological data, we encounter several challenges (Argelaguet et al., 2021) (cha, 2023):

- Data collected using different assays (instruments) have different statistical properties and assumptions leading to heterogenous data modalities. Combining these different likelihood models in a single inference framework is not a trivial statistical task.
- Overfitting is usually common for high dimensional biological data that has a large number of features and a small number of observations. Data redundancy is also a huge factor that contributes to overfitting.
- Information leakage is common during train/test split.
- Data sparsity - Single cell datasets tend to have a lot of missing data that needs to be handled appropriately.

Our mitigation strategies included:

- Utilizing large and diverse datasets to combat overfitting.
- Exploring transformer-based models that are equipped to handle data sparsity and are more robust to batch effects.
- Splitting the data into train-validation-test sets based on features like assay types and donors to avoid information leakage. For comparison, we also evaluated performance of the baselines on random split, which despite giving better performance, is not an ideal split because of leakage.

2 Related Work

There have been multiple efforts to find a way to obtain cell representations that are immune to non-biological variances.

Deployed by researchers at Stanford University, MARS (Brbić et al., 2020) uses a meta-learning approach to annotate and discover cell types across heterogeneous experiments. It employs deep neural networks to learn embedding functions for cell types from the cellxgene matrix as well as cell type landmarks in the embedded space which are shared across tissues. This model has been successfully implemented on a large mouse cell atlas. The challenge with MARS lies in its incapability of handling non batch corrected datasets.

ScVI (Lopez et al., 2018) is a generative framework that uses probabalistic modeling for analysis of single cell omics data. It performs downstream tasks such as dimensionality reduction and transfer learning while accounting for technical noise and biases and retaining biological effects. However, it does not have an interpretable latent space and is computationally expensive to train.

scBERT uses single cell RNA-sequence data to obtain contextual gene embeddings immune to batch effects. It has been tested on multiple datasets and is highly accurate for the downstream task of cell-type annotation including novel cell type discovery. However, it does not take into account additional gene metadata like its functionality, which can help improve the performance of all downstream tasks.

scFormer explores the same domain as scBERT but also takes into account gene metadata and passes it as input embeddings in addition to gene embeddings and gene expression embeddings. It also provides cell embeddings in addition to contextual gene embeddings.

Our project explored some of these methods on two datasets (Integrated Lung Cell Atlas and Tabula Sapiens) and performed a comparitive analysis of the obtained clustering and classification performances.

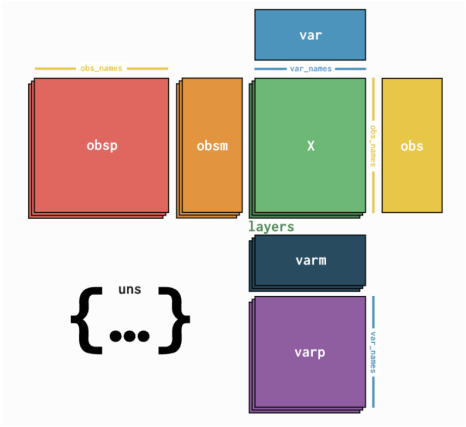


Figure 1: Structure of Annotated Data (Virshup et al., 2021)

3 Experiments

3.1 Datasets

The datasets that we work with are stored as annotated data (Virshup et al., 2021) whose structure is shown in Figure 1 where:

- **X:** This is a matrix with indexed rows (observations) and columns (features). For both our datasets, the rows are indexed by unique cell identifiers and the columns are indexed using gene names. Each entry represents the amount of expression of a gene in each cell and is real-valued. This is also called the *cellXgene matrix*. A section of this matrix for the ILCA dataset is shown in Figure 2.
- **obs:** These are 1-D annotations about the observations. For our datasets, some *obs* are *cell_type, assay, disease, sex* etc.
- **var:** These are 1-D annotations about features. For our datasets, some *var* are *highly_variable, feature_name* etc.
- **obsm, varm:** *obsm* are multidimensional annotations about observations. Examples include a dimensionality-reduced representation of *X* like *X_umap*. Similarly, *varm* are multidimensional annotations about features.
- **obsp, varp:** *obsp* are used to store relationships between observations. Examples include - distance or any other commutative relationship between cells. *varp* is defined in a similar way for features.
- **uns:** This includes unstructured data pertinent to the dataset.

The two publicly available datasets that we explore in our experiments are:

- **Integrated Lung Cell Atlas (ILCA):** This dataset consists of around 600,000 cells from 107 donors, and has been integrated and re-annotated by data-driven clustering. Its *cellXgene matrix* has dimensions **584884 × 28024**.
- **Tabula Sapiens (TS):** This dataset encompasses the human cell atlas of 24 tissues of 15 individuals, totaling to about 500,000 cells. We selected this dataset due to its expansive coverage of tissues and assay types. Its *cellXgene matrix* has dimensions **483152 × 58604**.

Since each of these datasets have a very high number of genes, we run all our methods on only the *highly variable genes* for both the datasets, which reduces the dimensionality of ILCA and TS to **584884 × 1996** and **483152 × 2432** respectively. Highly Variable Genes (HVGs) are genes that contribute to cell-to-cell differences in a mixed cell population. Both our datasets include information about a gene being highly variable or not as a boolean *var*.

3.2 Evaluation Metrics

In each of our methods, we obtain intermediate representations of all cells. These representations are evaluated in two ways -

- Qualitatively, we analyse the UMAPs of our latent representations. We colour each representation in two ways -
  - **By cell type:** Colouring by cell-type should show distinct clusters as we want the latent representations to effectively separate out various cell types.
  - **By non-biological variances:** Colouring by other non-biological variances among cells, like datasets or donors, should not show clustering as we want our representations to ignore these variances.
- Quantitatively, we use *rand score* which measures how close two clusters are to each other. If we cluster our representations using a clustering algorithm like leiden, we want this clustering to be representative of clusters formed using cell types, and not non-biological variances like datasets and donors. Hence, we find the rand scores for both - cell types and non-biological variances - for each of our methods and compare them.

	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938
GCGACCATCCCTAACC_SC22	0.000000	0.0	0.694030	0.000000	0.0	0.000000
P2_1_GCGCAACCAGTTAACC	0.000000	0.0	0.000000	0.000000	0.0	0.000000
GCTCTGTAGTGCTGCC_SC27	0.000000	0.0	0.000000	0.000000	0.0	0.000000
P2_8_TTAGGACGTTCAGGCC	0.000000	0.0	0.492366	0.492366	0.0	0.820822
CTTGATTGTCAGTTTG_T164	0.335701	0.0	0.124845	0.000000	0.0	0.000000

Figure 2: a section of the cellXgene matrix for the ILCA (lun)

For cell type annotation, we use the metrics - *accuracy, weighted f1 score* and *macro f1 score*.

3.3 Baselines

We have identified five baseline models for the cell-type annotation task.

- **PCA:** We use *Principal Component Analysis* for dimensionality reduction of our dataset. We then use the learned representations of the cells from PCA to perform clustering. We also perform cell-type classification by using simple machine learning algorithms of *logistic regression* and *random forest* on these representations.
- **scVI** (Lopez et al., 2018) (scV)**:** This model has been tested on mice as well as human brain datasets at different instances for cell-type annotation tasks. In this model, we don't have to worry about batch correction in our datasets as the model is equipped to handle such bias. The code is publicly available in a repository. Similar to PCA, we pair scVI with *logistic regression* and *random forest* to perform cell-type annotation on the latent representations learned from scVI.
- **scBERT** (Wang et al., 2021) (scB)**:** As shown in Figure 3, scBERT is a transformer-based model for learning latent representations through gene-gene interactions. The *performer encoder* (a variation of transformer, better suited for high-dimensional inputs) is pretrained on unlabeled data by first binning gene expression values in a fixed number of bins (to transform it to a classification problem) and then masking these binned gene expression values. In the code, the gene expression values are then converted to a 200-dimension vector, so the dimensions of the input and output matrices to the performer are ***num_of_cells X num_of_genes X 200***. The pre-trained model is available for download from the code repository.

  In the finetuning code, we use the pretrained performer, followed by a 1-D convolution, followed by 3 fully connected layers (the last one being the classification layer) and use the cross-entropy loss. To visualize the UMAPs and evaluate the clustering metrics for scBERT, we use the outputs from the second-last layer (the one preceding the classification layer), which gives a 128-dimension vector for each cell.

3.4 Updates since the midterm report

- Ran all the baselines on the second dataset, i.e, Tabula Sapiens.
- Evaluated performance of the models on random split.
- Analysed the results for incorrectly classified cell types and mapped them to shortcomings in the dataset as well as the train-validation-test split.
- Obtained clustering results on various features like gender and smoking habits for scBERT to analyse the impact of pretraining on the latent representations.

3.5 Results and Implications

- For ILCA, using train-test split by assay, we saw the following results:
  - In the UMAP for cell-type clustering (shown in figure 4), it seems that PCA and scBERT have discrete clusters while scVI has mixed clustering. The rand scores in table 1 corroborate that clustering by cell types is the best for scBERT. However, the rand score for cell-type clustering of PCA representations is pretty low despite clear clustering. This is because the number of clusters formed by running leiden on PCA embeddings are significantly more than the number of cell types.

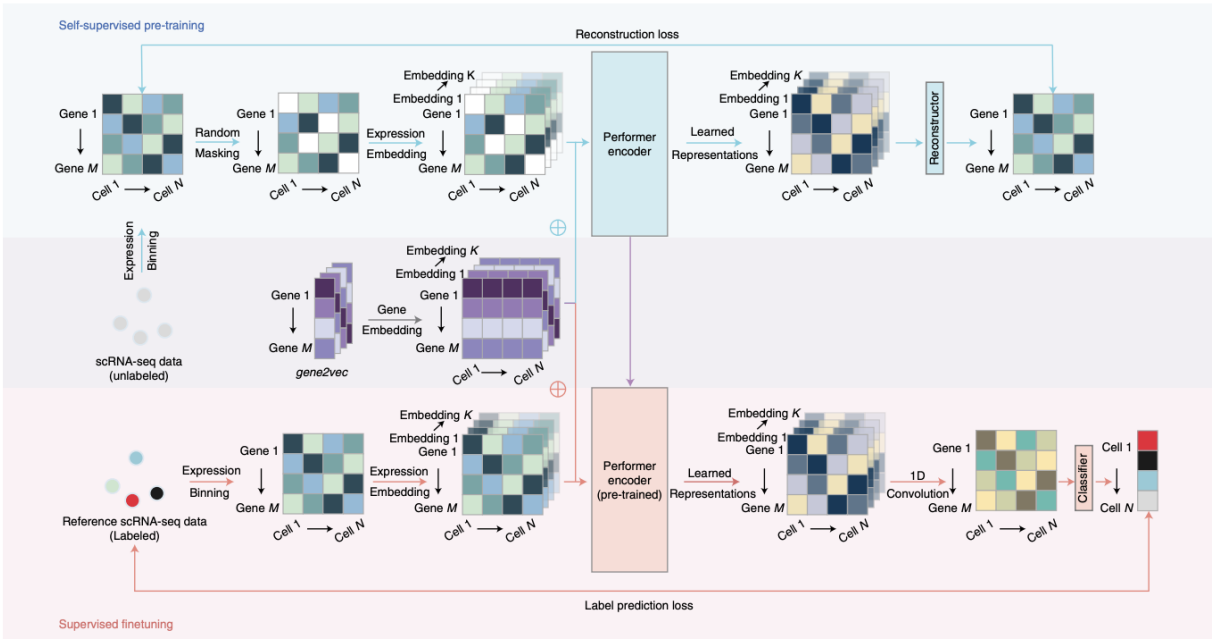


Figure 3: scBERT self-supervised pretraining on unlabeled data followed by supervised finetuning for cell-type classification(Wang et al., 2021)

(a) PCA (b) scVI (c) scBERT

Figure 4: ILCA clustering latent representations by cell-type

(a) PCA (b) scVI (c) scBERT

Figure 5: ILCA clustering latent representations by datasets

- The clustering UMAP for datasets as shown in figure 5 suggest high batch mixing for scVI and scBERT. Correspondingly, the *rand score* for dataset clustering in Table 1 is the lowest for scVI, while scBERT is a close second.
- The classification metrics in table 2 show that both *PCA + logistic regression* and *scBERT* show a similar accuracy and weighted f1, but scBERT shows a little better macro f1 score suggesting that it is performing well for the cell types that do not have a huge count in the dataset. Since ILCA is a well-integrated dataset, we see that the performance is good for a simple model like PCA + logistic regression.

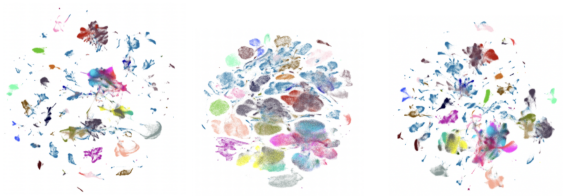
	PCA	**scVI**	**scBERT**
Cell Types	0.36	0.59	0.65
Datasets	0.21	0.10	0.145

Table 1: Rand scores for clustering ILCA latent representations

- For TS, using train-test split by donors, we saw the following results:

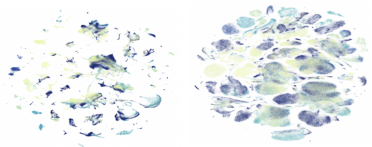
	PCA + Logistic Regression	PCA + Random Forest	scVI + Logistic Regression	scVI + Random Forest	scBERT
Accuracy	0.92	0.90	0.85	0.85	0.92
Macro F1	0.73	0.65	0.62	0.58	0.77
Weighted F1	0.92	0.90	0.86	0.85	0.92

Table 2: Performance of baselines on cell-type annotation task on ILCA (split by assays)



(a) PCA (b) scVI (c) scBERT

Figure 6: Tabula Sapiens clustering latent representations by cell-type



(a) PCA (b) scVI

(c) scBERT

Figure 7: Tabula Sapiens clustering latent representations by donors

	PCA	scVI	scBERT
Cell Types	0.45	0.38	0.52
Donors	0.05	0.04	0.05

Table 3: Rand scores for clustering Tabula Sapiens latent representations

  - The UMAP for cell-type clustering as shown in figure 6 suggests clear cell-type cluster boundaries for scBERT. The rand scores for cell-type clustering on TS as shown in table 3 prove the same.
  - In the UMAP for donor-clustering (figure 7, we see the maximum amount of mixing for scVI, which is also shown by the rand score for donors in table 3.
  - The classification metrics in table 4 suggest that complex models like scVI and scBERT show better performance than the simpler models. This is because TS is a more complex dataset and is not as well-integrated as ILCA.

- For the sake of completeness, we also ran all the methods (except scBERT) on a random split across the ILCA dataset. The classification metrics for this split are shown in table 5. We see that a random split outperforms the results from our split by assays. While most biological benchmarks use this split, it is not advisable because-
  - It leads to misleadingly high results because of data leakage.
  - It does not replicate real-world scenarios in which a finetuned model is used to infer results from new assays or new donors.

- For both the datasets, the primary reason for misclassified cell types was lack of support for the cell types in the dataset. However, there were some exceptions to this, which include:
  - The misclassified cell type had high support in the test set but was rare in the training set.
  - The cell was misclassified as the parent cell type. As an example, many instances of the *naive thymus-derived CD4-positive, alpha-beta T cell* were misclassified as their parent *effector CD4-positive, alpha-beta T cell*.
  - All examples for the misclassified cell type came from only one donor/assay in the training set, leading to the model not learning to mitigate differences arising from differences in donors/assays.

- We also analysed the differences in clustering across genders and smoking habits of donors:

	PCA + Logistic Regression	PCA + Random Forest	scVI + Logistic Regression	scVI + Random Forest	scBERT
Accuracy	0.55	0.56	0.59	0.56	0.59
Macro F1	0.24	0.21	0.27	0.26	0.24
Weighted F1	0 0.53	0.54	0.59	0.53	0.57

Table 4: Performance of baselines on cell-type annotation task on Tabula Sapiens (split by donors)

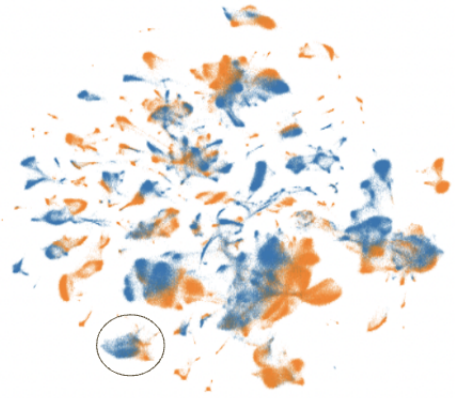
	PCA + Logistic Regression	PCA + Random Forest	scVI + Logistic Regression	scVI + Random Forest
Accuracy	0.94	0.95	0.86	0.86
Macro F1	0.93	0.89	0.80	0.74
Weighted F1	0.94	0.95	0.86	0.86

Table 5: Performance of baselines on cell-type annotation task on ILCA (random split)

- In figure 8, we see some differences in cluster embeddings across genders, most likely from pretraining. As shown in the figure, the highlighted cell-type *erythrocyte* forms two separate adjoining clusters for men and women, which is true biologically as well, i.e, this cell-type is significantly different for males and females. We also see no mixing across genders for cell types specific to genders, like mammary glands, which is expected.
- We also see differences in clustering across smoking habits of donors. In figure 9, cell types like *type II pneumocytes* are clustered as two adjoining clusters for smokers and non-smokers. Biologically as well, these cells are greatly affected by smoking habits of humans.

Both the above findings suggest that scBERT has a significant amount of knowledge preserved from its pretraining and can be used for multiple other classification tasks like novel cell-type discovery.

(a) Cell Types



(b) Genders

Figure 8: Tabula Sapiens latent representations coloured by cell types and genders, with cell type *erythrocyte* highlighted

3.6 Challenges and Future Work

One of the biggest challenges we faced was computational resource management as even our relatively smaller and simpler datasets are very large and high dimensional. Hence, storing and running our models was very time and space consuming. We also spent a considerable amount of time engaging with the genomics data to understand its nuances as we lacked domain knowledge in the field.

(a) Cell Types (b) Genders

Figure 9: Tabula Sapiens latent representations coloured by cell types and donors, with cell type *type II pneumocytes* highlighted

In terms of future work, there are several directions we plan to take to expand on our current research:

- Examine the effects of various kinds of train-test splits on our baselines. We have observed in our results how different splits give vastly different results and we believe testing that further would provide interesting insights.
- Run our models on larger and more complex datasets such as the entire cell-by-gene corpus in order to fully exploit the capabilities of these models.
- Alleviate parent-child level nomenclature discrepancies to improve performance on our cell-type annotation task. We can study ways to harmonize cell types across datasets as they might be annotated differently at a parent-child level.
- Gauge performance on novel cell type discovery and check if it is clustered away from existing cell types.
- Study the performance of our embeddings on other downstream tasks: Once we address issues for cell-type classification, we can test and finetune our models for other downstream tasks like imputation and translation between modalities.
- Explore advanced masking strategies for the transformer-based model - scBERT. In the transformer-based model, the masking strategy is sub-optimal, where only the non-zero values in the sparse cellxgene matrix are masked during training. Since the number of non-zero gene expressions in cell RNA sequence is very low, this leads to low utilization of single cell data for pre-training. We can explore methods to help us address this under-utilization.

4 Conclusion

To conclude, our research in this paper aimed to learn latent representations of biological cells that ignore non-biological variances and retain relevant gene-by-gene relationships in human cells. We tested the performance of our embeddings/latent representations on the downstream task of cell-type annotation for two datasets, namely the Integrated Lung Cell Atlas and Tabula Sapiens.

We modeled five baselines for learning these representations and then classifying them by cell type. These baselines were - *PCA + Logistic Regression, PCA + Random Forest, scVI + Logistic Regression, scVI + Random Forest*, and *scBERT* - on various splits. We also qualitatively studied clustering through UMAPs for various biological and non-biological factors. We observed that ILCA generally performed well with simple models like *PCA + Logistic Regression* while Tabula Sapiens performed better with the more complex models. We analyzed our results from both a computational and biological perspective. In the future, we hope

to improve the quality of our learned representations, test their effectiveness on other downstream tasks and expand on the same for larger datasets like the entire cell-by-gene corpus.

5 Acknowledgements

These ideas and roadmap would not have been possible without the continuous support of our industry mentor, Ivana Jelic and PhD mentor, Purva Pruthi. Ivana helped us get acquainted with genetic data and gave us the opportunity to play around with multiple CZI datasets. She was also really helpful in directing us to the resources for existing work in the domain.

Purva has been extremely supportive in helping us come up with a clear roadmap for the project. She has also helped us in getting comfortable with unknowns, which is a primary skillset to work in research.

References

Human cell atlas.

Lung cell atlas.

Scbert code.

scvi code.

Tabula sapiens dataset.

2023. Chan zuckerberg initiative.

Ricard Argelaguet, Anna Cuomo, Oliver Stegle, and John Marioni. 2021. Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39.

Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O Pisco, Russ B Altman, Spyros Darmanis, and Jure Leskovec. 2020. Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nature Methods*, 17(12):1200–1206.

Haotian Cui, Chloe Wang, Hassaan Maan, Nan Duan, and Bo Wang. 2022. scformer: A universal representation learning approach for single-cell data using transformers. *bioRxiv*.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. 2018. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058.

L Sikkema, D Strobl, L Zappia, E Madissoon, NS Markov, L Zaragosi, M Ansari, M Arguel, L Apperloo, C Bécavin, M Berg, E Chichelnitskiy, M Chung, A Collin, ACA Gay, B Hooshiar Kashani, M Jain, T Kapellos, TM Kole, C Mayr, M von Papen, L Peter, C Ramírez-Suástegui, J Schniering, C Taylor, T Walzthoeni, C Xu, LT Bui, C de Donno, L Dony, M Guo, AJ Gutierrez, L Heumos, N Huang, I Ibarra, N Jackson, P Kadur Lakshminarasimha Murthy, M Lotfollahi, T Tabib, C Talavera-Lopez, K Travaglini, A Wilbrey-Clark, KB Worlock, M Yoshida, , T Desai, O Eickelberg, C Falk, N Kaminski, M Krasnow, R Lafyatis, M Nikolíc, J Powell, J Rajagopal, O Rozenblatt-Rosen, MA Seibold, D Sheppard, D Shepherd, SA Teichmann, A Tsankov, J Whitsett, Y Xu, NE Banovich, P Barbry, TE Duong, KB Meyer, JA Kropski, D Pe'er, HB Schiller, PR Tata, JL Schultze, AV Misharin, MC Nawijn, MD Luecken, and F Theis. 2022. An integrated cell atlas of the human lung in health and disease. *bioRxiv*.

Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. 2021. Anndata: Annotated data.

Wenchuan Wang, Fan Yang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. 2021. scbert: a large-scale pretrained deep langurage model for cell type annotation of single-cell rna-seq data. *bioRxiv*.