
scARE: Attribution Regularization for Single Cell Representation Learning

Kemal Inecik^{1,2} Fabian Theis^{1,3}

Abstract

Single-cell data generation techniques have provided valuable insights into the intricate nature of cellular heterogeneity. However, effectively unraveling subtle variations within a specific gene set of interest, while mitigating the confounding presence of higher-order variability, remains challenging. To address this, we propose scARE, a novel end-to-end generative deep learning model, amplifies model sensitivity to a preselected subset of features while minimizing others. scARE incorporates an auxiliary attribution loss term during model training, which empowers researchers to manipulate the model’s behavior robustly and flexibly. In this study, we showcased scARE’s applicability in two concrete scenarios: uncovering subclusters associated with the expression patterns of two cellular pathway genes, and its ability to optimize the model training procedure by leveraging time-points metadata, resulting in improved downstream performance.

1. Introduction

Deep neural networks, particularly autoencoders, are extensively employed in integrating and analyzing single-cell data, demonstrating outstanding performance in tasks such as batch correction, dimension reduction, and perturbation modeling (Lopez et al., 2018; Inecik et al., 2022; Heumos et al., 2023). While biologically informed deep learning is an active research area (Lotfollahi et al., 2023; Qoku & Buettner, 2023; Conard et al., 2023; Janizek et al., 2023), a method is currently lacking that enables robust and flexible manipulation of an autoencoder model’s behavior in order to enhance the influence of a *freely* chosen subset of input features on the latent space or the reconstruction process. In this study, we present scARE, *single cell attribution*

regularization, an end-to-end generative model designed to amplify the model’s sensitivity to a predefined subset of features while concurrently reducing sensitivity to others. It employs a novel training procedure for single-cell data through an auxiliary *attribution loss* term, aimed at effectively unveiling the intricate cellular heterogeneity sought after, while minimizing the potential interference of confounding higher-order variations.

We demonstrate that scARE enables the formation of distinct subclusters for each cell-type in latent space, which differentiate from one another based on the expression patterns associated with the cellular pathway of interest. Enabling unrestricted selection of feature subsets, scARE provides the opportunity to investigate cellular heterogeneity in the expression patterns not only of genes from established biological pathways but also of custom gene lists of interest to researchers. In addition, feature selection based on their association with diseases (such as cancer, diabetes, or viral infections) could potentially facilitate the exploration of diverse cellular states that may be intricately linked to distinct responses and sensitivities to these diseases.

Moreover, scARE allows users to guide model training by utilizing the metadata of each data point, potentially leading to improved performance in downstream analyses. We show that scARE is able to steer model training to segregate clusters populated by cells from different time-points for each cell-type. While yielding outcomes similar to cVAE approaches that utilize additional information regarding cell-type labels during model training (Xu et al., 2021; De Donno et al., 2022), scARE achieves greater flexibility by employing binary attribution prior vectors instead of relying solely on such labels. Furthermore, we show that scARE adeptly learns a probabilistic representation of the data while concurrently addressing biological and technical factors, achieving performance on par with the baseline model.

2. Methods

2.1. scARE model architecture

We propose to incorporate an auxiliary loss term into a conditional variational autoencoder (cVAE) as the base model, where the conditions correspond to batch labels and are provided to both the generative and inference models. cVAEs have been shown to effectively handle the inherent spar-

¹Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany ²School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany ³Department of Mathematics, Technical University of Munich, Garching, Germany. Correspondence to: Fabian Theis <fabian.theis@helmholtz-muenchen.de>.

sity, noise, and variability in single-cell data while capturing latent representations that facilitate various downstream analyses such as clustering, data imputation, trajectory inference, and label transfer (Heumos et al., 2023). Both generative and inference models are optimized jointly by maximizing the evidence lower bound, ELBO, of the intractable marginal log-likelihood of data $\log p_\theta(x|c)$ given in Equation 1.

$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\phi(z|x,c)} \left[\log p_\theta(x|z,c) - \log \frac{q_\phi(z|x,c)}{p_\theta(z|c)} \right] \quad (1)$$

Here, $p_\theta(z|c)$ denotes a prior distribution of the latent variables z given condition c , $p_\theta(x|z,c)$ is the likelihood function of a data point x , $q_\phi(z|x,c)$ is an approximate posterior distribution, and θ and ϕ are respectively generative and inference model parameters. The training loss is defined as the negative ELBO, which can be understood as the sum of the reconstruction loss, $\log p_\theta(x|z,c)$, and KL divergence, $\log q_\phi(z|x,c) - \log p_\theta(z|c)$, where latent z is sampled using the reparameterization trick (Kingma et al., 2015; Kim et al., 2021) from the posterior.

A separate encoder for size factors is implemented, similar to scVI (Lopez et al., 2018), to enhance the model’s performance, interpretability in downstream analyses, and robustness to noise and heterogeneity. The generative model operates on the latent space to reconstruct observed data using a negative binomial loss function (Equation 2) specified by mean μ and inverse dispersion θ . KL divergence was calculated assuming Normal distribution in the latent space (Equation 3), with μ_l and σ_l estimated by the inference model.

$$\begin{aligned} L_R &= \text{NB}(x; \mu, \theta) \\ &= \frac{\Gamma(x + \theta)}{\Gamma(x + 1)\Gamma(\theta)} \left(\frac{\theta}{\theta + \mu} \right)^\theta \left(\frac{\mu}{\theta + \mu} \right)^x \end{aligned} \quad (2)$$

$$L_D = -\text{KL}(\mathcal{N}(\mu_l, \sigma_l) \| \mathcal{N}(0, 1)) \quad (3)$$

In this study, an auxiliary loss term, attribution loss, L_A , is introduced and incorporated into the base model. The modified model aims to have higher sensitivity to variations in a specific set of input feature values while maintaining its known effectiveness as a cVAE in handling the complex characteristics of single-cell data. The total loss function for the resulting model can be formulated as in Equation 4, where w_A , w_D , and w_R are hyperparameters that dictate the relative contributions of each loss component.

$$L_{\text{scARE}} = L_R w_R + L_D w_D + L_A w_A \quad (4)$$

2.2. Attribution loss

To compute the attribution loss, L_A , feature attribution methods expected gradients (EG) or integrated gradients (IG) are employed on individual minibatches as the first step. Given a differentiable function $F : \mathbb{R}^n \rightarrow \mathbb{R}$, where n is the number of input features, IG aims to attribute the difference in the function’s output between a given reference input x' ,

typically a zero vector, and the input of interest x to the individual features of x (Sundararajan et al., 2017; Chen et al., 2019). The attribution for the i th feature is defined in Equation 5, where the integral computes the average gradient along the straight-line path from x' to x in the input space.

$$\text{IG}_i(x, x') = (x_i - x') \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (5)$$

EG computes the expected contribution of each feature to the function’s output over a distribution of reference inputs (Erion et al., 2021). Given a distribution D over reference inputs x' , the EG attribution for the i th feature is defined in Equation 6. A simple notation is also in Equation 6 introducing noise distribution directly, using a variable ϵ representing a random perturbation drawn from a distribution D_ϵ . In this study, EG was preferred over IG on account of its robustness to reference inputs, ability to handle non-linearities and feature interactions, provision of distribution-aware explanations, and its superior efficiency with expedited model training process (Hesse et al., 2021; Erion et al., 2021).

$$\begin{aligned} \text{EG}_i(x) &= \mathbb{E}_{x' \sim D, \alpha \sim U(0,1)} \left[(x_i - x') \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \right] \\ &= \mathbb{E}_{\epsilon \sim D_\epsilon} \left[\frac{\partial F(x_i + \epsilon)}{\partial x_i} \right] \end{aligned} \quad (6)$$

The function F used in Equations 5 and 6 serves to consolidate the inference or generative processes into a single scalar, thereby enabling the regularization of either the entire model or exclusively the encoder during the training phase through attribution loss. While various forms of F have been implemented, including the simplest one being a summation operation, the selection of F remains a model hyperparameter to be tuned. After calculating attributions, the resulting vector undergoes a transformation to ensure non-negativity and scaling to reside within the interval $[0, 1]$. Several re-parameterization methods, such as sigmoid transformation or absolute value operation followed by min-max scaling, were experimented in this context. The selection of an effective re-parameterization method is a critical hyperparameter, requiring careful tuning due to its direct impact on model performance.

Prior to commencing the training procedure, a subset of input features must be chosen based on the objectives of subsequent downstream analysis. As previously mentioned, there are no predefined regulations or restrictions governing feature selection, granting the user absolute autonomy to make choices aligned with their analysis objectives. scARE requires a binary vector, called ground truth, GT , where features to sensitize are assigned 1 and the remaining features are assigned 0. Alternatively, features can be selected independently for individual data points, resulting in a binary GT matrix. In this study, we chose genes associated with GO:0043410 and GO:1901224 Gene Ontology (GO) terms as GT gene definitions in Section 3.1, which correspond to

positive regulation of MAPK and NF- κ B pathways respectively. The selection of these specific GO terms was solely intended to demonstrate the proof-of-concept of a biological application. However, it is noteworthy that comparable results were achieved when employing different GO terms (data not presented). On the other hand, in Section 3.2, the features are chosen by their established role as cell-type markers, or their ability to explain the variability across different time-points of each cell-type.

Lastly, the predefined GT vector is compared with the calculated attribution vector for each data point typically through MSE or MAE as in Equation 7, where n is number of features, g_i and a_i are respectively predefined GT and calculated attribution values. We designed attribution loss to penalize the discrepancy between expected attributions defined by GT vector with actual attribution scores, constraining the model to prioritize the variation in the chosen feature values while de-emphasizing the variation in other features. scARE also allows multiple GTs, where individual attribution losses are computed separately, potentially employing different functions. The total attribution loss is then aggregated, either by simple summation or squared-summation, incorporating a weight term for each attribution loss component.

$$L_{A_{\text{MSE}}} = \frac{1}{n} \sum_{i=1}^n (g_i - a_i)^2 \quad L_{A_{\text{MAE}}} = \frac{1}{n} \sum_{i=1}^n |g_i - a_i| \quad (7)$$

2.3. Feature-set subcluster founder score

We devised a novel metric, named feature-set subcluster founder score, *FSSF*, to investigate whether subclusters are formed within a given cell-type based on the expression patterns of GT genes. To calculate this metric, an experimental process is executed wherein the expression values for GT genes are swapped between cells belonging to two distinct subclusters of a cell-type, which have been previously identified by the Leiden algorithm in the latent space. The resulting modified cells, which were unseen during the model training phase, now carry the GT gene expression values from the cells of another latent subcluster. The FSSF score aims to quantify the degree to which these cells map in the boundaries of the subcluster from which the GT genes were sourced, as well as to the subcluster from which all other genes were sourced. The primary objective is to evaluate whether the model perceives all cells within these two subclusters as identical, with the sole differentiating feature being the patterns of GT gene expression. For each pair of Leiden subclusters within a cell-type, the FSSF score is formulated as $(n_g - n_o)/n_t$. Here, n_g and n_o represent the cell count that remapped to the subcluster where the GT genes belong and to the subcluster where all other genes belong, respectively, while n_t represents the total number of cells involved in the exchange. The FSSF score hence is a measure that indicates the relative impact of GT genes

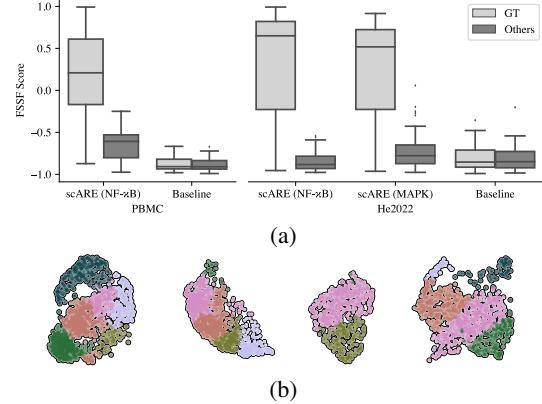


Figure 1. a) Improved FSSF scores upon inclusion of L_A . Each column correspond to FSSF scores for all subcluster pairs in cell-types after training. ‘Others’ represents mean score of 100 repetitions of FSSF calculations using randomly selected feature subsets of the same size as GT genes. b) Representative cell-type clusters from *He2022* with high FSSF score for each pair subclusters.

versus all other genes on the identity of a cell-type’s sub-clusters. A score close to 1 implies that GT genes dominate this identity, while a score near -1 signifies that all other genes exert a more significant influence.

2.4. Datasets

In this study, two scRNA-seq datasets were used for model training and subsequent analyses. The first dataset, named *PBMC*, was provided and processed by *scvi-tools* package (Gayoso et al., 2022; 10x Genomics, 2023), resulting in a dataset with 11990 cells and 3346 genes. The second dataset, named *He2022* (He et al., 2022), underwent quality checking and filtering for samples from 15 and 22 post-conception weeks, resulting in 8963 cells, followed by the selection of 2000 highly variable genes (HVGs) using *scanpy* (Wolf et al., 2018).

3. Results

3.1. Robust subclustering in cell-types associated with expression patterns of predefined gene subsets

In conventional single-cell analysis workflows, the discernibility of cellular dissimilarities with respect to a desired set of gene patterns within the latent space is often impeded by a confounding presence of higher-order variability that obfuscates the finer intricacies being pursued. In this study, we showed that incorporating an auxiliary loss term enables steering the model towards a particular set of genes. This leads to the formation of subclusters that correspond to distinct, yet possibly unexplored, cellular states associated with those genes. The prevalence of such subclusters is demonstrated for both datasets, as illustrated in Figure 1a, where FSSF scores approach 1 exclusively for GT genes when L_A is incorporated. It is worth noting that each of these subclusters exhibits unique expression patterns for GT

genes, as substantiated by autocorrelation matrices (data not presented).

The capabilities of the scARE model manifest clearly in a diverse range of potential applications. It could, for example, help pinpoint unique expression patterns of homeobox genes linked specifically to distinct time and tissue combinations within a large developmental dataset, deepening our understanding of the intricate dynamics of development. Additionally, ongoing research (data not presented) delves into the complex interplay between these cellular states and specific perturbations. Encouragingly, preliminary findings indicate that certain subclusters exhibit heightened responsiveness to particular perturbations, providing valuable insights into diverse cellular behaviors under varying conditions.

Importantly, the proposed approach maintains cellular homogeneity within individual cell-types, similar to the baseline model without attribution loss (further explained in Section 3.3), but it results in the reorganization of cells across different Leiden subclusters of a given cell-type (Figure 1b). Furthermore, although the relationship between the input and latent space is non-linear, the findings demonstrate that the inclusion of attribution loss leads to a rearrangement of cells in the latent space such that they display a gradient-like polarization in the total expression of GT genes (Figure 2).

3.2. Steering model training with time-points metadata for improved downstream analysis performance

Achieving optimal segregation among clusters that potentially delineate distinct phenotypes, such as diverse cell-types, time-points (ages), or organs, constitutes an important element in single-cell analysis workflows, including single-cell atlas building. (Luecken et al., 2022; Heumos et al., 2023). In this study, we showed scARE can reveal subtle changes in the transcriptomic landscape of various cell-types during early human development that are otherwise indistinguishable using conventional methods. scARE enables exploration of the developmental lung dataset, *He2022*, with enhanced performance, allowing for improved gene expression visualization and trajectory inference.

The top 40 genes that account for the variability between two time-points (post-conception week, PCW, 15 and 22) were calculated using mutual information scores for scARE model training. This calculation was performed for all cells, resulting in a GT vector, as well as for each cell-type inde-

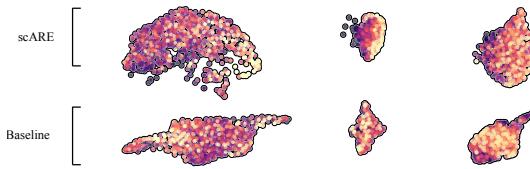


Figure 2. Representative cell-type clusters from *He2022*, colored for polarization in terms of total MAPK expression.

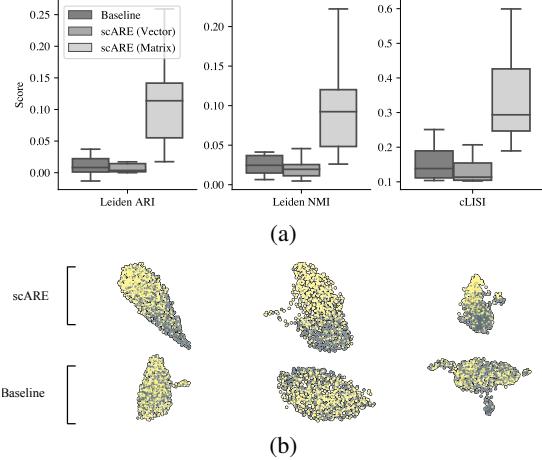


Figure 3. a) Improvement for cluster definition upon incorporation of attribution loss in model training. b) Representative cell-type clusters from *He2022* with segregated time-point subclusters by scARE training with cell-type specific time-point markers. Yellow and gray correspond to cells from 15 and 22 PCW respectively.

pendently, resulting in a GT matrix. The degree of overlap between time-point subclusters for each cell-type was observed to decrease substantially, leading to a more distinct delineation of cluster definitions (Figure 3). The improvement was observed when choosing cell-type specific GTs, which is anticipated given that the genes demarcating separate temporal instances of a cell-type are largely subject to the cell-type itself. Moreover, the initial findings (data not presented) on the HLCA dataset (Sikkema et al., 2023) suggest that incorporating an attribution loss into the model training, using a GT matrix based on cell-type markers, enhances some model performance metrics such as improved label transfer accuracy and lower prediction uncertainty.

3.3. Comparable performance to the baseline model

In an ideal scenario, scARE should maintain baseline performance, minimizing any significant compromise, while concurrently delivering substantial enhancements in the interpretability and explication of the model’s predictive outcomes. The incorporation of attribution loss was found to have no significant detrimental impact on the effectiveness of data integration, specifically in relation to batch removal and preservation of biological variation (Figure 4).

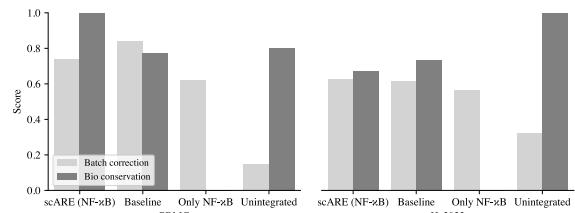


Figure 4. Data integration performance for batch correction and biological variation conservation, obtained by *scib* (Luecken et al., 2022). ‘Unintegrated’ is based on the top 50 PCA components of the input space, while ‘Only NF- κ B’ is based on the latent space of the baseline model using only these genes.

4. Discussion

The scARE model was shown to effectively improve its own responsiveness to variations in a designated set of feature values while upholding its acknowledged effectiveness as a cVAE for single-cell data. Its practical applications were highlighted in two concrete scenarios: identification of subclusters correlated with gene expression patterns of the predefined gene set, as well as refining the separation of clusters that are presumed to belong distinct groups. The incorporation of attribution loss into cVAE model training, as pioneered by scARE, is expected to aid future investigations into the integration of interpretability and generative modeling techniques.

As part of our future objectives, we aim to support our analysis by showcasing scARE on a broader range of datasets and exploring diverse feature subsets. Through these efforts, we seek to gain deeper insights into the variations among different subclusters and potentially correlate them with pertinent gene programs from the biological literature. Exploring the inclusion of multiple ground truths (e.g. GTs from NF- κ B and apoptosis pathways simultaneously) would introduce captivating dimensions to the analysis. While our existing implementation enables training with a combined loss, a method still needs to be devised to maintain their disjointness or orthogonality. Furthermore, the use cases of scARE are envisioned to be extended to select an optimal set of probed genes in targeted spatial transcriptomics (Kuemmerle et al., 2022) through the utilization of a zero GT vector and a loss function penalizing larger deviations, thereby imposing a constraint to employ the minimum number of features for reconstruction. As current implementation lacks multimodal functionality, a further objective is to extend scARE to incorporate multimodal single-cell datasets (Teichmann & Efremova, 2020).

5. Acknowledgement

Sincere gratitude is extended to the reviewers for their generous contribution of time and expertise in meticulously evaluating this work. Their insightful feedback and constructive suggestions have significantly elevated the caliber of this paper. Special thanks are also extended to Lisa Sikkema, Louis Kuemmerle, Adil Meric, Ignacio Ibarra, Ekin Gokce Cicek, and Soroor Hediyyeh-zadeh for their invaluable feedback during the writing process. Their meticulous attention to detail, thoughtful insights, and astute feedback have played an instrumental role in refining the coherence and effectiveness of this endeavor.

References

10x Genomics. 10x Genomics Single Cell Gene Expression Datasets. Retrieved from <https://support.10xgenomics.com/single-cell-gene-expression/datasets>, 2023. Accessed: 2023-05-12.

Chen, J., Wu, X., Rastogi, V., Liang, Y., and Jha, S. Robust attribution regularization. *Advances in Neural Information Processing Systems*, 32, 2019.

Conard, A. M., DenAdel, A., and Crawford, L. A spectrum of explainable and interpretable machine learning approaches for genomic studies. *Wiley Interdisciplinary Reviews: Computational Statistics*, pp. e1617, 2023.

De Donno, C., Hediyyeh-Zadeh, S., Wagenstetter, M., Moinfar, A. A., Zappia, L., Lotfollahi, M., and Theis, F. J. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *bioRxiv*, pp. 2022–11, 2022.

Erion, G., Janizek, J. D., Sturmels, P., Lundberg, S. M., and Lee, S.-I. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021.

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiolah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., et al. A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology*, 40(2):163–166, 2022.

He, P., Lim, K., Sun, D., Pett, J. P., Jeng, Q., Polanski, K., Dong, Z., Bolt, L., Richardson, L., Mamanova, L., et al. A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. *Cell*, 185(25):4841–4860, 2022.

Hesse, R., Schaub-Meyer, S., and Roth, S. Fast axiomatic attribution for neural networks. *Advances in Neural Information Processing Systems*, 34:19513–19524, 2021.

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücke, M. D., Strobl, D. C., Henao, J., Curion, F., et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, pp. 1–23, 2023.

Inecik, K., Uhlmann, A., Lotfollahi, M., and Theis, F. J. Multiepa: Multimodal compositional perturbation autoencoder. *bioRxiv*, pp. 2022–07, 2022.

Janizek, J. D., Spiro, A., Celik, S., Blue, B. W., Russell, J. C., Lee, T.-I., Kaeberlin, M., and Lee, S.-I. Pause: principled feature attribution for unsupervised gene expression analysis. *Genome Biology*, 24(1):81, 2023.

Kim, J., Kong, J., and Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pp. 5530–5540. PMLR, 2021.

- Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.
- Kuemmerle, L. B., Luecken, M. D., Firsova, A. B., Barros de Andrade e Sousa, L., Straßer, L., Heumos, L., Mekki, I. I., Mahbubani, K. T., Sountoulidis, A., Balassa, T., et al. Probe set selection for targeted spatial transcriptomics. *bioRxiv*, pp. 2022–08, 2022.
- Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.
- Lotfollahi, M., Rybakov, S., Hrovatin, K., Hediyyeh-Zadeh, S., Talavera-López, C., Misharin, A. V., and Theis, F. J. Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology*, 25(2):337–350, 2023.
- Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.
- Qoku, A. and Buettner, F. Encoding domain knowledge in multi-view latent variable models: A bayesian approach with structured sparsity. In *International Conference on Artificial Intelligence and Statistics*, pp. 11545–11562. PMLR, 2023.
- Sikkema, L., Ramírez-Suástequi, C., Strobl, D. C., Gillett, T. E., Zappia, L., Madissoon, E., Markov, N. S., Zaragozi, L.-E., Ji, Y., Ansari, M., et al. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, pp. 1–15, 2023.
- Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.
- Teichmann, S. and Efremova, M. Method of the year 2019: single-cell multimodal omics. *Nat. Methods*, 17(1):2020, 2020.
- Wolf, F. A., Angerer, P., and Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.
- Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., and Yosef, N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1):e9620, 2021.