Interpretable visualization of single cell data using Janus autoencoders

Gokul Gowri
Systems Biology
Harvard University
ggowri@g.harvard.edu

Philippa Richter
Computing and Mathematical Sciences
California Institute of Technology
prichter@caltech.edu

Xiao-Kang Lun*
Wyss Institute
Harvard University
xiaokang.lun@wyss.harvard.edu

Peng Yin*
Wyss Institute
Harvard University
py@hms.harvard.edu

Abstract

The emergence of single-cell transcriptomics and proteomics approaches has resulted in a wealth of high-dimensional data that are challenging to interpret. Dimensionality reduction methods, such as UMAP and t-SNE, project data points onto a low-dimensional space that preserves cellular similarities from the high-dimensional measurement space. However, the projected dimensions typically have no interpretable biological meaning, and the relationships between measured biomolecular features are obscured completely. These limitations can be overcome by finding embeddings in which each dimension is a function of a distinct and biologically meaningful set of features. Here, we introduce Janus autoencoders, a novel neural network architecture capable of finding such low-dimensional embeddings by jointly optimizing multiple distinct one-dimensional embeddings of a dataset. We demonstrate the utility of Janus autoencoders for (1) visualizing multiomic data such that modality-specific contributions to cell state can be deconvolved and (2) visualizing mass cytometry data such that cell cycle effects can be distinguished from "true" cell state differences. Our initial demonstrations indicate that Janus autoencoders can visually represent relationships between cellular states and their underlying cellular features in multiple biological contexts.

1 Introduction

The advances of sequencing technology enables profiling of whole genomes and/or transcriptomes at single-cell resolution [10, 6]. Single-cell proteomics technologies, such as mass cytometry (CyTOF), CITE-seq, and CODEX, use antibodies to detect proteins of interest, and allow for the simultaneous quantification of over 50 epitopes in millions of single cells [2, 14, 8]. Such high-dimensional measurements yield information on regulatory mechanisms of cellular behaviors and enable investigations of cellular and tissue heterogeneity. However, precise interpretations of highly multiplexed data remain challenging.

Visualization is an essential step in the analysis of high dimensional single cell data. Dimensionality reduction techniques allow visualization of -omics data in two-dimensional plots, interpretable by human inspection. Principal component analysis (PCA) is typically performed to linearly transform and project high-dimensional data to coordinates that preserve variation in cellular state. Recently

*Co-corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

popularized non-linear dimensionality reduction techniques, such as t-stochastic neighbour embedding (t-SNE) [11] and uniform manifold approximation and projection (UMAP) [1], aim to find a nonlinear mapping from a high-dimensional measurement space to a low-dimensional (typically 2D) space, in which similarities between data points are preserved.

Although UMAP and t-SNE are widely used in analyzing transcriptomics and proteomics data, a few inherent limitations of these type of methods prevent deep and precise biological interpretations. First, the low-dimensional embeddings identified by UMAP and t-SNE are determined by arbitrary nonlinear maps of the high-dimensional input space. As such, the dimensions of the low-dimensional embedding have no interpretable biological meaning. Second, while these methods preserve the similarity of single cells, the relationships between features are lost completely in the nonlinear mapping. Third, multi-omics analysis with UMAP or t-SNE involves intensive data normalization, preprocessing and integration [3, 9] which could potentially bias resulting visualizations.

One-SENSE, a related nonlinear dimensionality reduction approach, aims to find embeddings in which the dimensions have interpretable biological meaning and relationships between markers of interest are preserved [5]. This is done by splitting measured cellular features into biologically meaningful categories using prior knowledge, mapping each category of features to a single dimension using t-SNE or UMAP, and plotting the one dimensional embeddings together in a biaxial plot. This yields dimensions that each represent a distinct category of features, such that variation along an axis indicates the category of features responsible the variation.

However, one-SENSE visualizations typically lack well-defined clusters and clear distinctions of cell state [5, 12]. We hypothesize that this is due to the lack of robustness of reduction to a single dimension using UMAP and t-SNE. It has been shown that there are often many appreciably different yet equally faithful low dimensional embeddings of single cell data [4]. As there are likely to be several equivalently faithful one-dimensional embeddings, different pairs of one-dimensional embeddings may have very different qualities when jointly visualized in two-dimensional space. In one-SENSE, the two one-dimensional embeddings are optimized completely separately, so pairs of one-dimensional embeddings suitable for 2D visualization cannot be selected for.

In this work, we seek to visualize single cell data using a nonlinear dimensionality reduction method in which each reduced dimension represents a distinct, biologically meaningful set of features. In contrast with previous methods, we preserve cell state information by jointly optimizing multiple one-dimensional embeddings in a context-aware fashion. To perform this optimization, we introduce Janus autoencoders, a novel neural network architecture. We demonstrate the utility of Janus autoencoders for two tasks: (1) visualizing multiomic data such that modality-specific contributions to cell state can be deconvolved, and (2) visualizing mass cytometry data such that cell cycle effects can be distinguished from “true” cell state differences.

2 Results

2.1 Network architecture

To jointly optimize multiple one dimensional embeddings of high dimensional data, we have developed a topologically modified autoencoder that we refer to as a Janus ("two-faced") autoencoder (Fig 1a). While traditional autoencoders find low dimensional latent spaces in which each dimension is a function of all input dimensions, Janus autoencoders constrain the information used in each latent dimension by restricting the connectivity of the network. In particular, Janus autoencoders contain multiple disjoint encoder modules which map distinct sets of input dimensions to distinct latent dimensions, each of which can be used for a one-dimensional embedding. To ensure that these latent dimensions yield faithful joint embeddings as well as faithful one-dimensional embeddings, we use multiple decoder modules: one for each latent dimension which aims to reconstruct only the connected inputs, and one “joint” decoder which aims to reconstruct the full initial input based on all latent dimensions. A more specific description of the Janus architecture can be found in Appendix.

2.2 Decoupling modal contributions to cell state in single cell CITE-seq data

Janus autoencoders can be used naturally to analyze multiomics data in a modality-aware fashion, without intensive preprocessing or integration. With each measurement modality mapped to a different latent dimension, it is possible to infer modality-specific contributions to cell state based

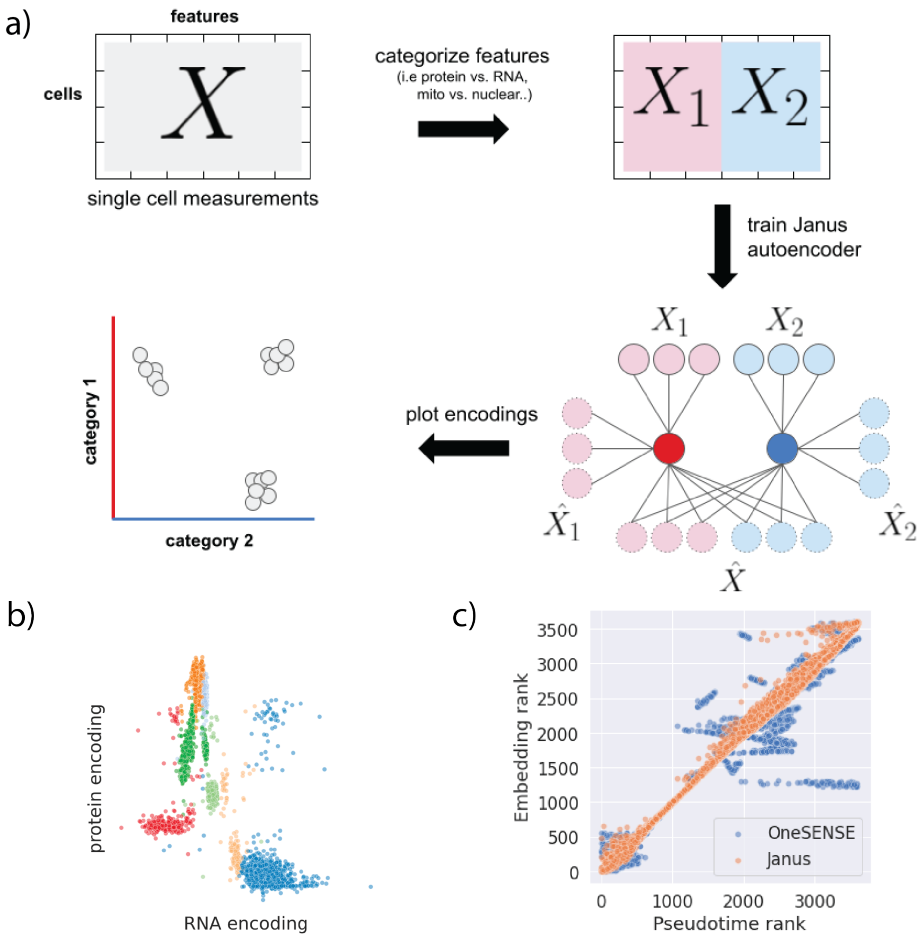


Figure 1: **(a)** Workflow of single cell data visualization using Janus autoencoders. First, features in high dimensional single cell data are split into biologically meaningful categories. Then, a Janus autoencoder is trained on the dataset with categorized features. The Janus autoencoder architecture contains two encoder modules (shown with inputs X_1, X_2) and three decoder modules (shown with outputs $\hat{X}, \hat{X}_1, \hat{X}_2$). This architecture jointly optimizes two latent dimensions, each of which represents a different category of features. Finally, the optimized latent space (shown in dark red and dark blue) is used to visualize the dataset. **(b)** A Janus autoencoder visualization of CITE-seq profiling of PBMCs. Each axis is an encoding of a distinct modality. Colors correspond to Leiden clusters. **(c)** Correspondence between 1D embeddings of cell cycle markers in mass cytometry data and pseudotime estimate based on phospho-RB value.

on visualization of the latent space. We demonstrate a proof-of-concept using a publically available CITE-seq [14] dataset simultaneously quantifying the transcripts and surface proteins of 5,000 peripheral blood mononuclear cells (PBMCs) [7].

Using Janus autoencoders, we visualize this data such that each axis corresponds to information from a single modality (Fig. 1b). Qualitatively, Leiden clusters are preserved, satisfying an initial sanity check for accurate embeddings. When a similar visualization is generated using oneSENSE, this sanity check is failed (Appendix Fig. 2b). Furthermore, the Janus visualization shows distinct populations of PBMCs within coarse Leiden clusters. For example, the cluster shown in green diverges into two sub-clusters along the RNA axis. Based on established cell type markers [15], we determined that this cluster corresponds to CD8+ T-cells, and the sub-clusters correspond to CD8+ memory and naive T-cells, differentiated by transcriptomic state (Appendix Fig 3).

It is well known mRNA levels in cells and tissues do not fully predict protein abundance [16]. In this initial demonstration, we show that Janus antoencoders could be used to decipher relationships between transcriptome and proteome in multi-omics data.

2.3 Regressing out cell cycle effects in mass cytometry data

Experimentally meaningful single cell variation is often obscured by confounding factors such as cell cycle stage and cell size [13]. Using Janus autoencoders, it is possible to map potential confounding features to a different latent dimension than other markers of interest, allowing one to visualize the effects of and regress out confounding features in single cell datasets.

We demonstrate a proof of concept using a mass cytometry dataset which profiles 4 proteins with prominent cell cycle dependence in addition to 31 intracellular phosphorylation sites in a monocyte cell line (THP1) [13]. As observed in prior work, non-cell-cycle aware analyses, such as visualization with 2D UMAP, erroneously indicate the presence of multiple cell clusters due to the differential cell cycle states of analyzed cells (Appendix Fig 4a).

By analyzing cell-cycle dependent markers independently from the phosphorylation panel, it is possible to deconvolve cell cycle effects from experimentally meaningful differences in cell state. When visualized with Janus autoencoders, it is clear that cell state is relatively homogeneous when standard phosphorylation targets are decoupled from cell-cycle dependent targets, with the exception of cells in M-phase (Appendix Fig 4c), in line with prior analyses [13]. Furthermore, the Janus autoencoder visualization faithfully reconstructs the pseudotime ordering of the cells in their cell cycle state (Fig 1c). In comparison to an analogous plot generated using oneSENSE, cell cycle trajectory is captured much more effectively: the Janus autoencoder has Spearman rank correlation of > 0.99 with a pseudotime estimate, compared to ~ 0.88 for the analogous oneSENSE plot.

This analysis indicates that Janus autoencoders enable directly analyzing the effects of confounding factors in mass cytometry data without intensive data correction or normalization.

3 Discussion

In this work, we have introduced Janus autoencoders, a novel neural network architecture for jointly learning multiple one dimensional embeddings of high dimensional data. Through initial demonstrations for modality-aware clustering in multiomic sequencing data and cell cycle normalization in mass cytometry data, we have shown the potential general utility of Janus autoencoders for interpretable visualization of single cell data. Compared to oneSENSE, the existing state-of-the-art, Janus autoencoders appear to more effectively capture cell type information, regress out cell cycle effects, and preserve pseudotime ordering.

Janus autoencoders are subject to some of the same intrinsic shortcomings as other single cell data visualization techniques. In particular, for most high-dimensional datasets, it is highly unclear if reduction to a low dimensional space is mathematically justifiable. In future work, we plan to develop general metrics of embedding quality that can serve as a heuristic for the validity of visualizations produced by Janus autoencoders.

4 Acknowledgements

We thank Alan Amin and Tatiana Brailovskaya for fruitful discussions. We thank anonymous reviewers for thoughtful criticism. We acknowledge the Caltech Summer Undergraduate Fellowship (SURF) Program and Wyss Institute Molecular Robotics Initiative (MRI) for funding. We thank the Jupyter Project for maintaining open-source computational tools.

References

[1] Etienne Becht et al. "Dimensionality reduction for visualizing single-cell data using UMAP". en. In: *Nat. Biotechnol.* (Dec. 2018).

[2] Bernd Bodenmiller et al. "Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators". en. In: *Nat. Biotechnol.* 30.9 (Sept. 2012), pp. 858–867.

[3] Danila Bredikhin, Ilia Kats, and Oliver Stegle. "MUON: multimodal omics analysis framework". en. In: *Genome Biol.* 23.1 (Feb. 2022), p. 42.

[4] Tara Chari, Joeyta Banerjee, and Lior Pachter. "The Specious Art of Single-Cell Genomics". en. Aug. 2021.

[5] Yang Cheng et al. "Categorical Analysis of Human T Cell Heterogeneity with One-Dimensional Soli-Expression by Nonlinear Stochastic Embedding". en. In: *J. Immunol.* 196.2 (Jan. 2016), pp. 924–932.

[6] Charles Gawad, Winston Koh, and Stephen R Quake. "Single-cell genome sequencing: current state of the science". en. In: *Nat. Rev. Genet.* 17.3 (Mar. 2016), pp. 175–188.

[7] 10x Genomics. *PBMC 5k from 10x genomics*.

[8] Yury Goltsev et al. "Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging". en. In: *Cell* 174.4 (Aug. 2018), 968–981.e15.

[9] Yuhan Hao et al. "Integrated analysis of multimodal single-cell data". en. In: *Cell* 184.13 (June 2021), 3573–3587.e29.

[10] Allon M Klein et al. "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells". en. In: *Cell* 161.5 (May 2015), pp. 1187–1201.

[11] Dmitry Kobak and Philipp Berens. "The art of using t-SNE for single-cell transcriptomics". en. In: *Nat. Commun.* 10.1 (Nov. 2019), p. 5416.

[12] Florian Mair et al. "A Targeted Multi-omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level". en. In: *Cell Rep.* 31.1 (Apr. 2020), p. 107499.

[13] Maria Anna Rapsomaniki et al. "CellCycleTRACER accounts for cell cycle and volume in mass cytometry data". en. In: *Nat. Commun.* 9.1 (Feb. 2018), p. 632.

[14] Marlon Stoeckius et al. "Simultaneous epitope and transcriptome measurement in single cells". en. In: *Nat. Methods* 14.9 (July 2017), pp. 865–868.

[15] Peter A Szabo et al. "Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease". en. In: *Nat. Commun.* 10.1 (Oct. 2019), p. 4706.

[16] Bing Zhang et al. "Proteogenomic characterization of human colon and rectal cancer". en. In: *Nature* 513.7518 (Sept. 2014), pp. 382–387.

5 Appendix

5.1 Network architecture details

Informally, for the case of a two dimensional latent space, Janus architectures contain two encoder modules computing

$$z_1 = E_1(x_1)$$
$$z_2 = E_2(x_2)$$

in which x_1, x_2 are two disjoint subsets of elements in the input vector x, and z_1, z_2 are the latent dimensions to which they are mapped. These two latent dimensions are decoded with three modules, computing

$$\hat{x_1} = D_1(z_1)$$
$$\hat{x_2} = D_2(z_2)$$
$$\hat{x} = D(z_1, z_2)$$

in which $|\hat{x}| = |x_1| + |x_2|$.

In the case of a Janus autoencoder with 2 latent dimensions, where the input X is split into disjoint categories X_1, X_2, with reconstructions $\hat{X}, \hat{X_1}, \hat{X_2}$, we use a weighted mean-squared error loss function

$$\mathcal{L}(\hat{x_1}, \hat{x_2}, \hat{x}, x_1, x_2, x) = w_1 MSE(\hat{x_1}, x_1) + w_2 MSE(\hat{x_2}, x_2) + w_3 MSE(\hat{x}, x)$$

where the decoder weights w_1, w_2, w_3 are user-defined parameters which must be tuned for each specific application.

In initial demonstrations, we have used three hidden layers for each of the encoders and decoders, generally following the rule of thumb that each successive hidden encoding layer should decrease in size by a factor of 2 until enforced to be 1 in the latent dimension, and the inverse for decoding layers.

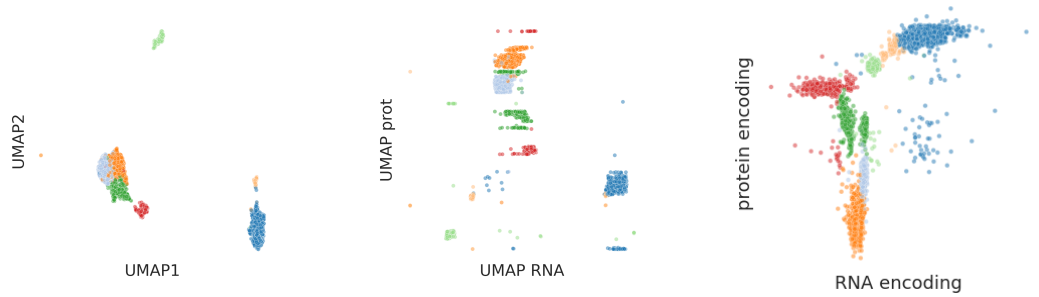
While this architecture is theoretically generalizable to larger latent spaces, we have thus far explored only 2 dimensional latent spaces.

5.2 CITE-seq analysis details

We used MUON [3] to preprocess data by filtering cells based on read count and percentage of mitochondrial reads, log transforming RNA counts, and selecting highly variable genes. In line with the MUON documentation, we performed multiplexed Leiden clustering. We generated a conventional 2D UMAP plot, preconditioned on 20 principal components (Fig 2a).

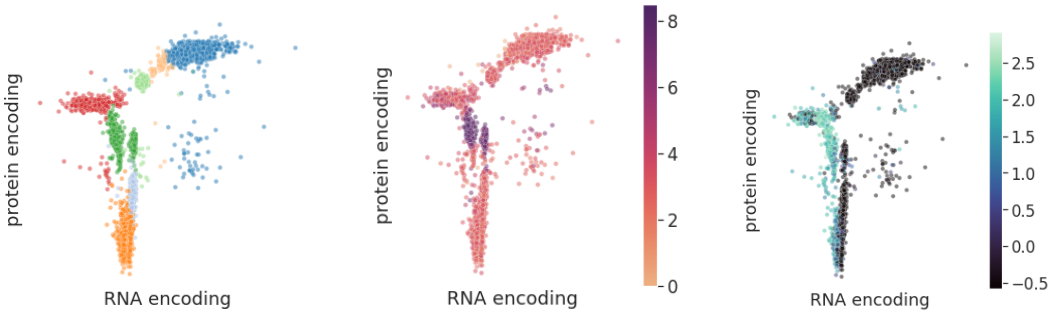
We find that a OneSENSE plot of the data partitioned by modality does not preserve Leiden clusters, indicating that cell type information is not faithfully represented (Fig 2b).

As seen in Fig , the Leiden cluster shown in green in Fig 2c contains cells with high CD8 protein expression. Through differential gene expression analysis, we found that the left subcluster has high expression of CCL5, a marker of memory T cells [15]. As such, we propose that the Janus autoencoder has differentiated CD8 positive cells into memory and naive cells, a finer visualization of cell type than shown in the 2D UMAP 2a. We note that while this Leiden cluster is split in the One-SENSE plot, it diverges on the protein axis rather than the RNA axis, and thus cannot indicate the same distinction in CD8 positive subtypes.



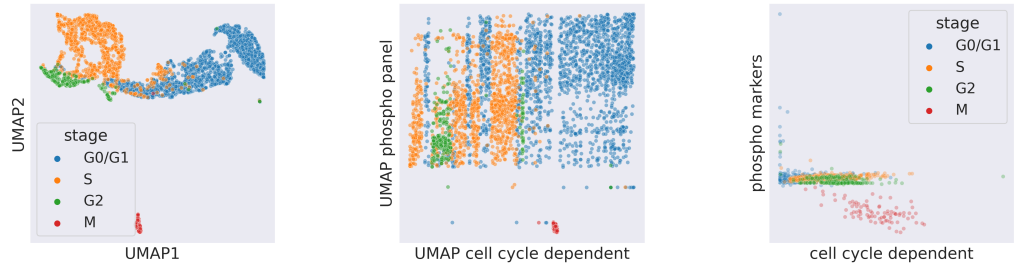
(a) 2D UMAP (b) oneSENSE (c) Janus autoencoding

Figure 2: Visualizations of CITE-seq profiling of PBMCs, colored by Leiden cluster. **(a)** A conventional 2D UMAP plot. **(b)** A one-SENSE plot, where the X and Y dimensions are 1D UMAP reductions of RNA and protein measurements respectively. **(c)** A Janus autoencoding visualization, where the X and Y dimensions are neural network encodings of RNA and protein measurements respectively.



(a) Colored by Leiden cluster (b) Colored by CD8 protein (c) Colored by CCL5 RNA

Figure 3: Specific feature values of cells in Janus autoencoding of CITE-seq PBMC dataset. **(a)** Janus autoencoding colored by Leiden cluster. **(b)** Janus autoencoding colored by CD8 surface protein expression. **(c)** Janus autoencoding colored by log-normalized CCL5 RNA expression.



(a) 2D UMAP (b) oneSENSE (c) Janus autoencoding

Figure 4: Visualization of mass cytometry profiling of THP1 cell line. **(a)** A conventional 2D UMAP plot, colored based on ground truth cell cycle stage annotations from [13]. **(b)** A one-SENSE plot, where the X and Y dimensions are 1D UMAP reductions of cell cycle markers and general protein measurements respectively. **(c)** A Janus autoencoding visualization, where the X and Y dimensions are neural network encodings of cell cycle markers and general protein measurements respectively.