AI for Genomics: from CNNs and LSTMs to Transformers

Nikolay Oskolkov, Group Leader (PI) at LIOS, Riga, Latvia

Physalia course, 09.09.2025

Session 2b: Autoencoder neural network applications for Cell Biology and Data Integration

@NikolayOskolkov

@oskolkov.bsky.social

Personal homepage:
https://nikolay-oskolkov.com

Image generated at www.canva.com

Session content

Topics we'll cover in this session:

1) Introduction to unsupervised machine learning

2) Autoencoder neural network architecture and principles

3) Introduction to advances in single cell biology

4) Autoencoders for dimension reduction in single cell biology

5) Autoencoders for integration of heterogeneous data

Unsupervised machine learning

Y = f (X), where X is input (data) and Y is output (response)

Y is present – supervised machine learning

Y is absent – unsupervised machine learning

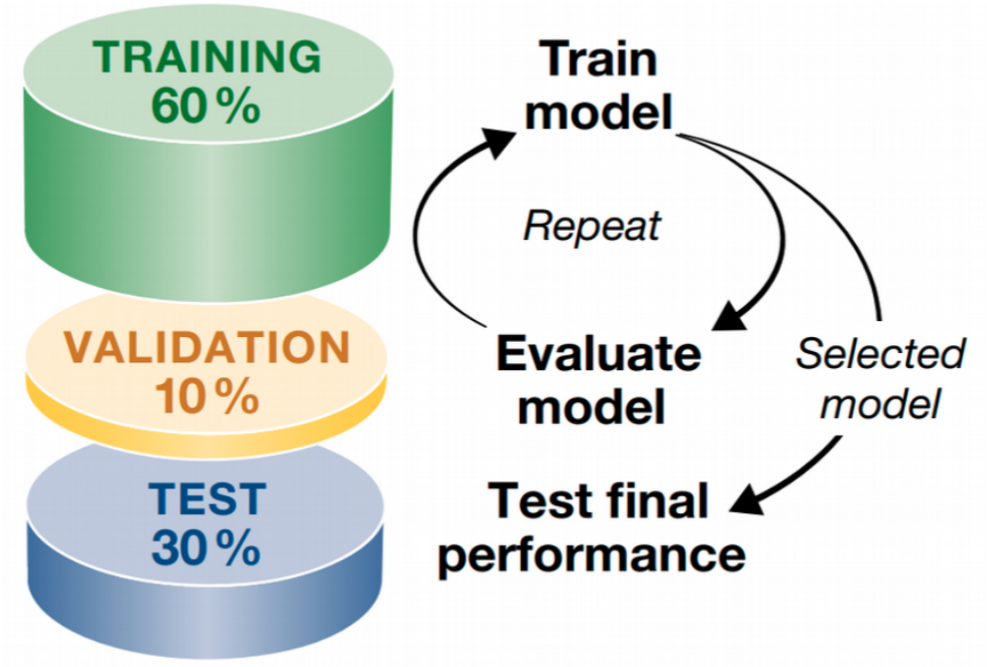
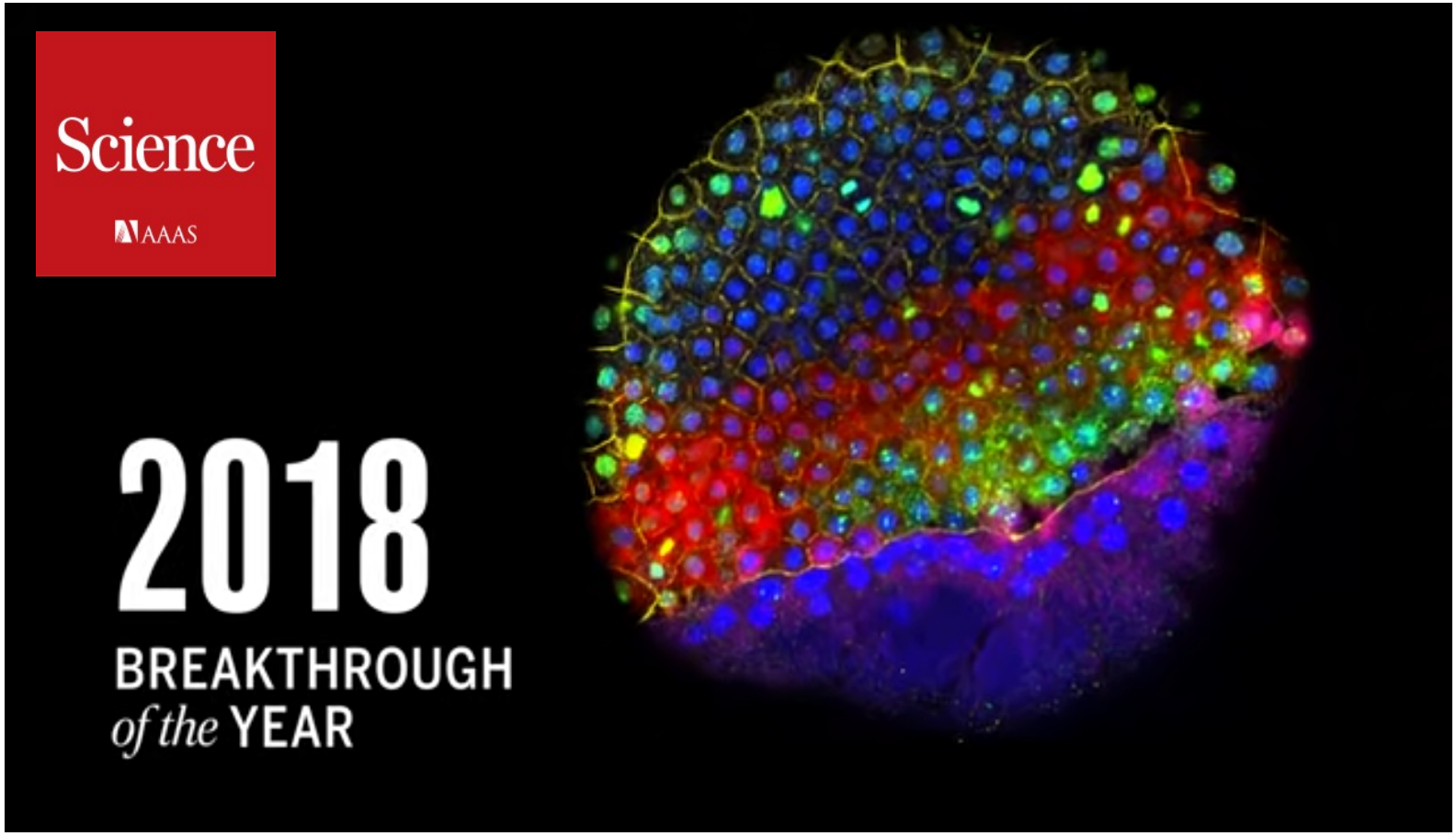


Image adapted from Angenmueller et al., Mol. Syst. Biol. 2016

Machine Learning typically involves five basic steps:

1. Split data set into **train**, **validation** and **test** subsets
2. Fit the model on the train subset
3. Validate your model on the validation subset
4. Repeat train - validation split many times and tune **hyperparameters**
5. Test the accuracy of the optimized model on the test subset.

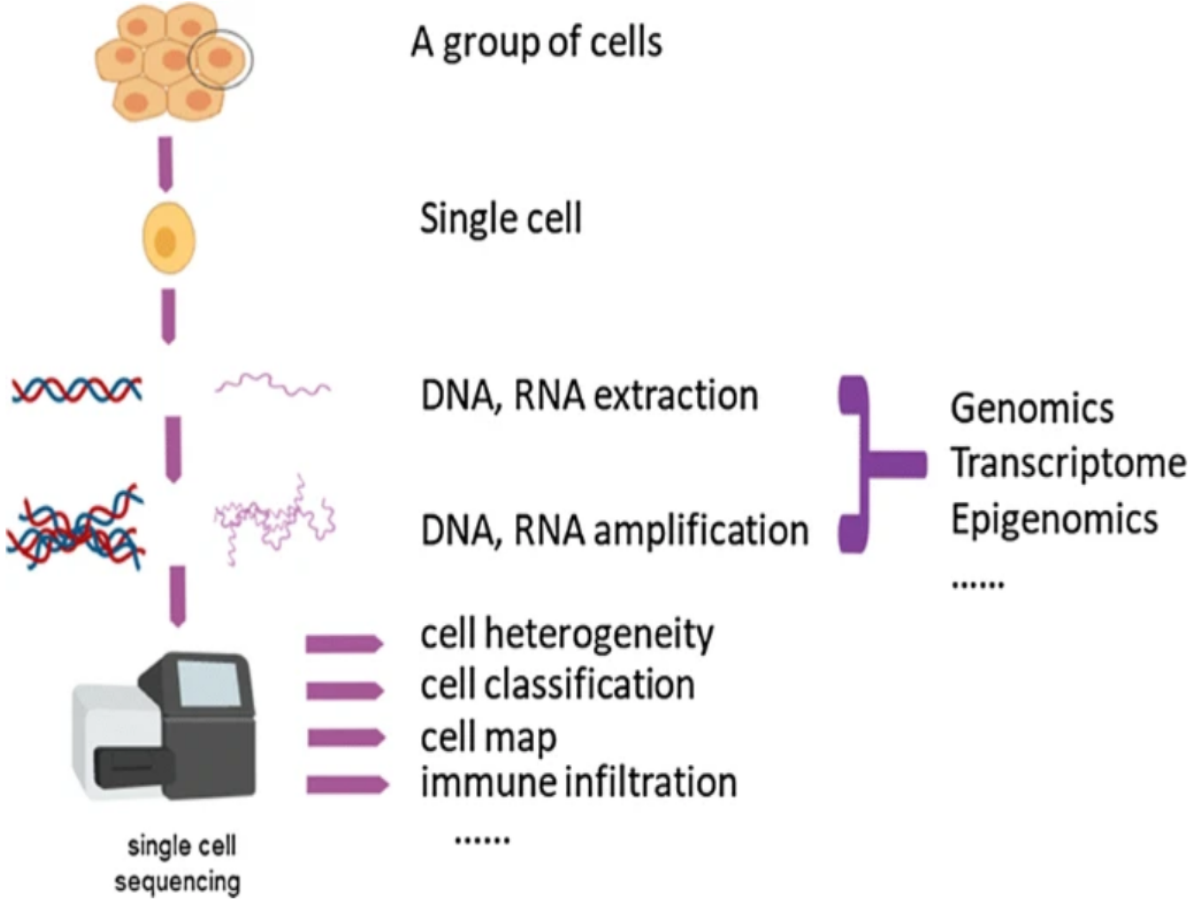
Recent advances in single cell biology

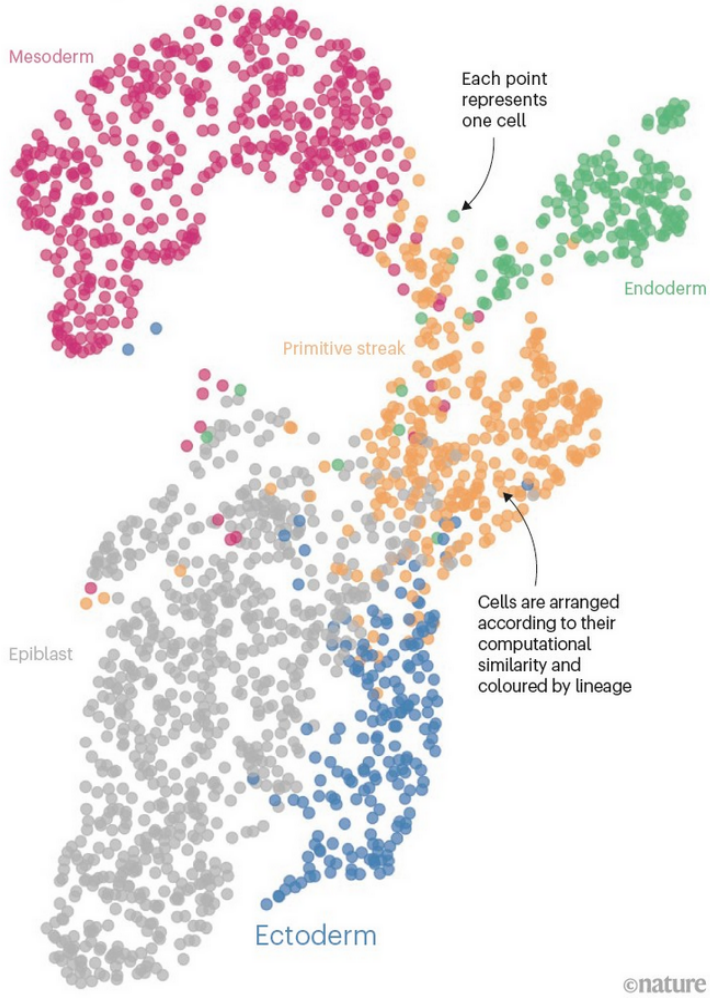


Nature method of the year: 2019 (single cell multi-omics), 2020 (spatial transcriptomics)

Image adapted from Science, 362, 6421, 2018

Principles and goals of single cell Omics





The principle of single-cell sequencing. It is a process of isolating a single cell for sequencing and studying cell heterogeneity, molecular mapping, immune infiltration and epigenetic changes

My experience with scRNAseq analysis

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-07582-3 OPEN

Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing

Michael Bartoschek[1], Nikolay Oskolkov[2], Matteo Bocci[1], John Lövrot[3], Christer Larsson[1], Mikael Sommarin[4], Chris D. Madsen[1], David Lindgren[1], Gyula Pekar[5], Göran Karlsson[4], Markus Ringnér[2], Jonas Bergh[3], Åsa Björklund[6] & Kristian Pietras[1]

Cancer-associated fibroblasts (CAFs) are a major constituent of the tumor microenvironment, although their origin and roles in shaping disease initiation, progression and treatment response remain unclear due to significant heterogeneity. Here, following a negative selection strategy combined with single-cell RNA sequencing of 768 transcriptomes of mesenchymal cells from a genetically engineered mouse model of breast cancer, we define three distinct subpopulations of CAFs. Validation at the transcriptional and protein level in several experimental models of cancer and human tumors reveal spatial separation of the CAF subclasses attributable to different origins, including the peri-vascular niche, the mammary fat pad and the transformed epithelium. Gene profiles for each CAF subtype correlate to distinctive functional programs and hold independent prognostic capability in clinical cohorts by association to metastatic disease. In conclusion, the improved resolution of the widely defined CAF population opens the possibility for biomarker-driven development of drugs for precision targeting of CAFs.

Oskolkov *et al. Skeletal Muscle* (2022) 12:16
https://doi.org/10.1186/s13395-022-00299-4

Skeletal Muscle

RESEARCH Open Access

High-throughput muscle fiber typing from RNA sequencing data

Nikolay Oskolkov[1,2], Malgorzata Santel[3], Hemang M. Parikh[4], Ola Ekström[1], Gray J. Camp[3], Eri Miyamoto-Mikami[5], Kristoffer Ström[1,6], Bilal Ahmad Mir[1], Dmytro Kryvokhyzha[1], Mikko Lehtovirta[1,7], Hiroyuki Kobayashi[8], Ryo Kakigi[9], Hisashi Naito[5], Karl-Fredrik Eriksson[1], Björn Nystedt[10], Noriyuki Fuku[5], Barbara Treutlein[3], Svante Pääbo[3,11] and Ola Hansson[1,7*]

Abstract

Background: Skeletal muscle fiber type distribution has implications for human health, muscle function, and performance. This knowledge has been gathered using labor-intensive and costly methodology that limited these studies. Here, we present a method based on muscle tissue RNA sequencing data (totRNAseq) to estimate the distribution of skeletal muscle fiber types from frozen human samples, allowing for a larger number of individuals to be tested.

Methods: By using single-nuclei RNA sequencing (snRNAseq) data as a reference, cluster expression signatures were produced by averaging gene expression of cluster gene markers and then applying these to totRNAseq data and inferring muscle fiber nuclei type via linear matrix decomposition. This estimate was then compared with fiber type distribution measured by ATPase staining or myosin heavy chain protein isoform distribution of 62 muscle samples in two independent cohorts ($n = 39$ and 22).

Results: The correlation between the sequencing-based method and the other two were $r_{ATPas} = 0.44$ [0.13–0.67], [95% CI], and $r_{myosin} = 0.83$ [0.61–0.93], with $p = 5.70 \times 10^{-3}$ and 2.00×10^{-6}, respectively. The deconvolution inference of fiber type composition was accurate even for very low totRNAseq sequencing depths, i.e., down to an average of ~ 10,000 paired-end reads.

Conclusions: This new method (https://github.com/OlaHanssonLab/PredictFiberType) consequently allows for measurement of fiber type distribution of a larger number of samples using totRNAseq in a cost and labor-efficient way. It is now feasible to study the association between fiber type distribution and e.g. health outcomes in large well-powered studies.

Introduction

Our bodies constitute to ~ 30–40% of the skeletal muscle, and it is the most abundant form of the three types of muscle, the others being smooth and cardiac. The skeletal muscle is composed of different fiber types (i.e., muscle cell types), and the relative proportions of these types vary among the muscles, locations within the muscles, individuals, and the sex of individuals [1–4]. The oxidative and glycolytic potential and the contractile properties differ considerably between fiber types, with the mitochondria-rich slow-twitch fibers (type I) having higher oxidative capacity, and fast-twitch fibers (type IIa and type IIx) having higher glycolytic capacity [1]. The proportions also change as people age, with type II fibers being preferentially affected by sarcopenia [5]. Exercising the skeletal muscle is a major site for catabolic metabolism of the blood glucose and lipids and the metabolic characteristics of this tissue influence both the

*Correspondence: Ola.Hansson@med.lu.se

[1] Department of Clinical Sciences, Lund University, Malmö, Sweden
Full list of author information is available at the end of the article

Growing sample sizes in single cell

PERSPECTIVE

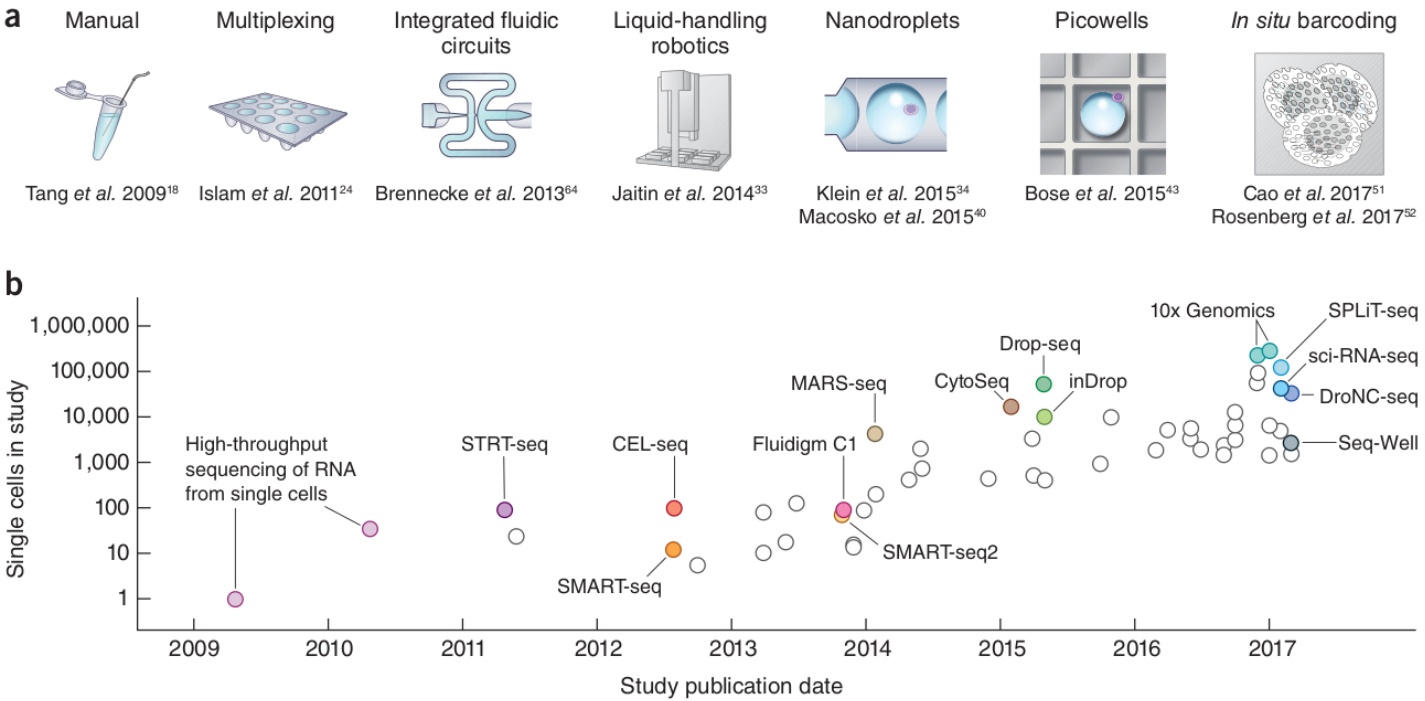


Figure 1 | Scaling of scRNA-seq experiments. (**a**) Key technologies that have allowed jumps in experimental scale. A jump to ~100 cells was enabled by sample multiplexing, and then a jump to ~1,000 cells was achieved by large-scale studies using integrated fluidic circuits, followed by a jump to several thousands of cells with liquid-handling robotics. Further orders-of-magnitude increases bringing the number of cells assayed into the tens of thousands were enabled by random capture technologies using nanodroplets and picowell technologies. Recent studies have used *in situ* barcoding to inexpensively reach th next order of magnitude of hundreds of thousands of cells. (**b**) Cell numbers reported in representative publications by publication date. Key technologies are indicated. A full table with corresponding numbers is available as **Supplementary Table 1**.

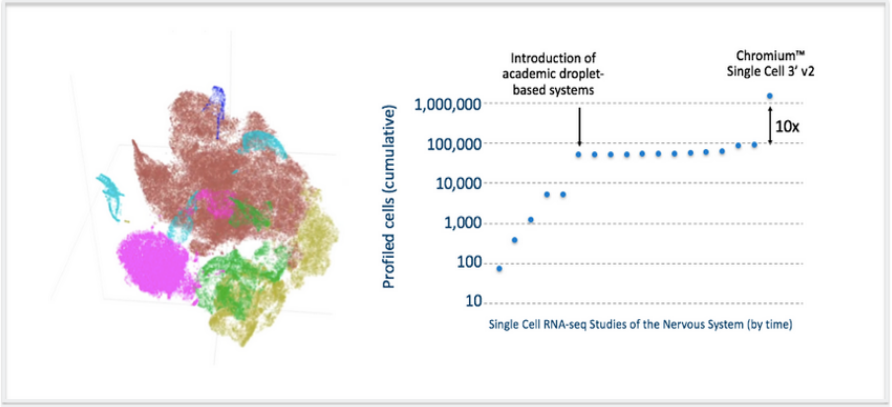
V. Svensson et al., Nature Protocols 2018

Big Data in single cell genomics

CAREERS BLOG 10X UNIVERSITY

10X GENOMICS SOLUTIONS & PRODUCTS RESEARCH & APPLICATIONS EDUCATION & RESOURC

« Back to Blog ‹ Newer Article Older Article ›



Our 1.3 million single cell dataset is ready to download

0 KUDOS

POSTED BY: grace-10x, on Feb 21, 2017 at 2:28 PM

At ASHG last year, we announced our 1.3 Million Brain Cell Dataset, which is, to date, the largest dataset published in the single cell RNA-sequencing (scRNA-seq) field. Using the Chromium™ Single Cell 3' Solution (v2 Chemistry), we were able to sequence and profile 1,308,421 individual cells from embryonic mice brains. Read more in our application note Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium™ Single Cell 3' Solution.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Single cells make big data: New challenges and opportunities in transcriptomics

Philipp Angerer[1], Lukas Simon[1], Sophie Tritschler[1], F. Alexander Wolf[1], David Fischer[1] and Fabian J. Theis[1,2]

Abstract

Recent technological advances have enabled unprecedented insight into transcriptomics at the level of single cells. Single cell transcriptomics enables the measurement of transcriptomic information of thousands of single cells in a single experiment. The volume and complexity of resulting data make it a paradigm of big data. Consequently, the field is presented with new scientific and, in particular, analytical challenges where currently no scalable solutions exist. At the same time, exciting opportunities arise from increased resolution of single-cell RNA sequencing data and improved statistical power of ever growing datasets. Big single cell RNA sequencing data promises valuable insights into cellular heterogeneity which may significantly improve our understanding of biology and human disease. This review focuses on single cell transcriptomics and highlights the inherent opportunities and challenges in the context of big data analytics.

Addresses

[1] Institute of Computational Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany
[2] Department of Mathematics, Technical University of Munich, Garching, Germany

Corresponding author: Theis, Fabian J. Institute of Computational Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany. (fabian.theis@helmholtz-muenchen.de)

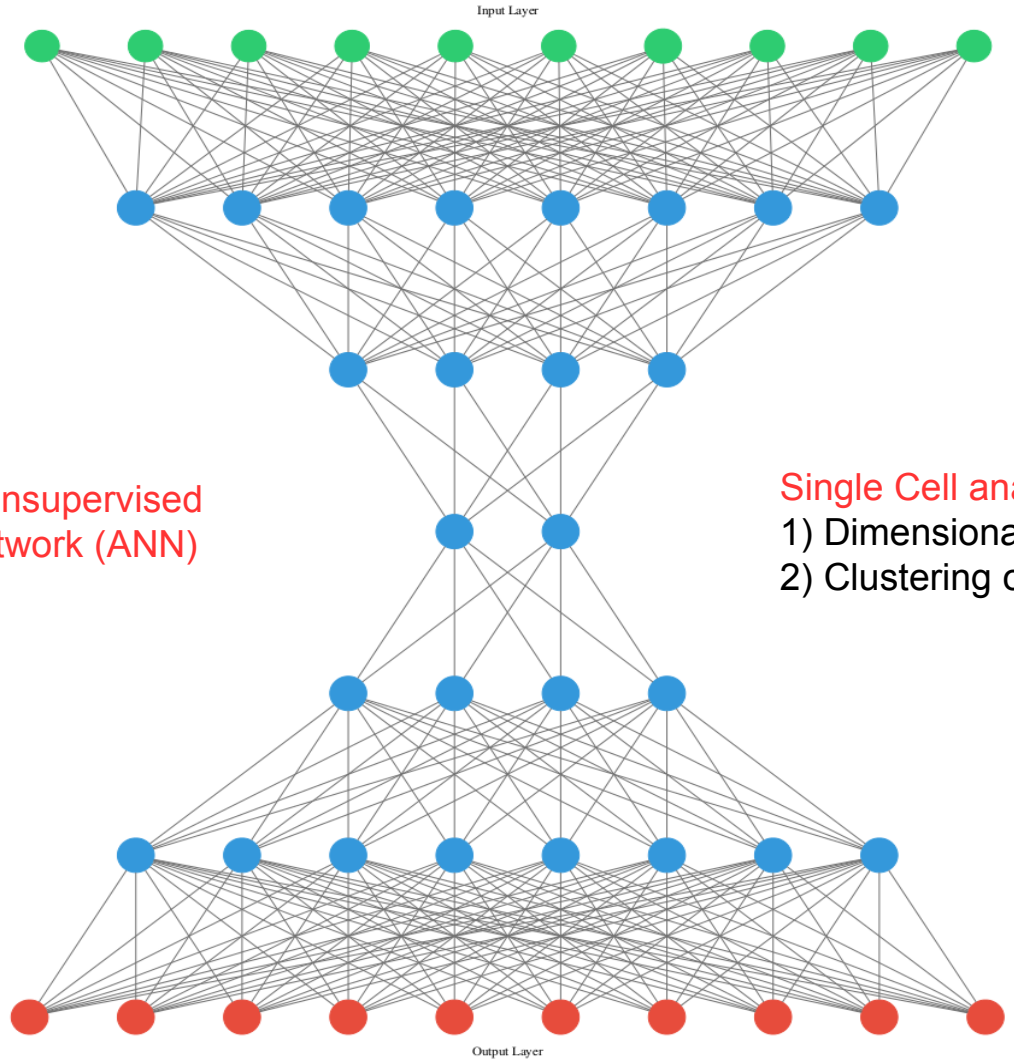
Current Opinion in Systems Biology 2017, 4:85–91

This review comes from a themed issue on **Big data acquisition and**

volume — the amount of data, velocity — the required processing speed, veracity — trustworthiness and availability, and variety — necessary model complexity [2]. The traditional scientific big data field is astronomy because of the huge *volume* of image data produced by telescopes with a high daily *velocity* [3]. Big data has also reached biology, mainly driven through the advent of next generation sequencing technology. For biologists, assessing *veracity* through statistical means is nothing new.

Recent technological advances now allow the profiling of single cells at a *variety* of omic layers (genomes, epigenomes, transcriptomes and proteomes) at an unprecedented level of resolution [4]. Single cell transcriptomics (SCT) entails the profiling of all messenger RNAs present in a single cell and constitutes the most widely-used sc profiling technology [4]. Unlike bulk RNA-seq profiling where sequencing libraries are generated from thousands of cells, scRNA-seq technologies isolate single cells and generate cell-specific sequencing libraries (e.g. Fluidigm [5]) mark RNA content with a cell-specific molecular barcode [6–9]. Both approaches generate gene expression estimates at the single cell level [10]. SCTenables, for the first time, the measurement of the transcriptomic information of thousands, and up to millions of single cells, in a single experiment [7]. The complexity of SCT data coupled with the massive volume inherent to next generation sequencing data makes it a paradigm of big data.

Autoencoder for single cell analysis

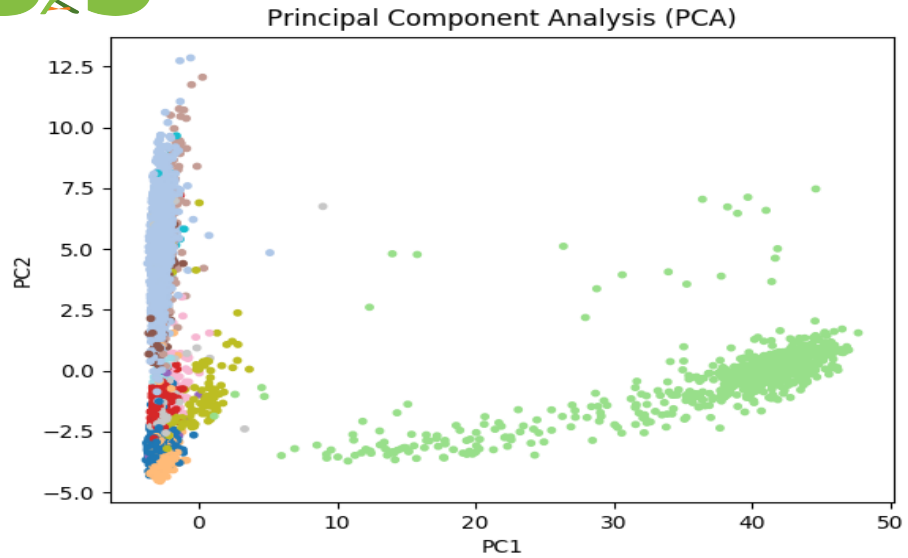


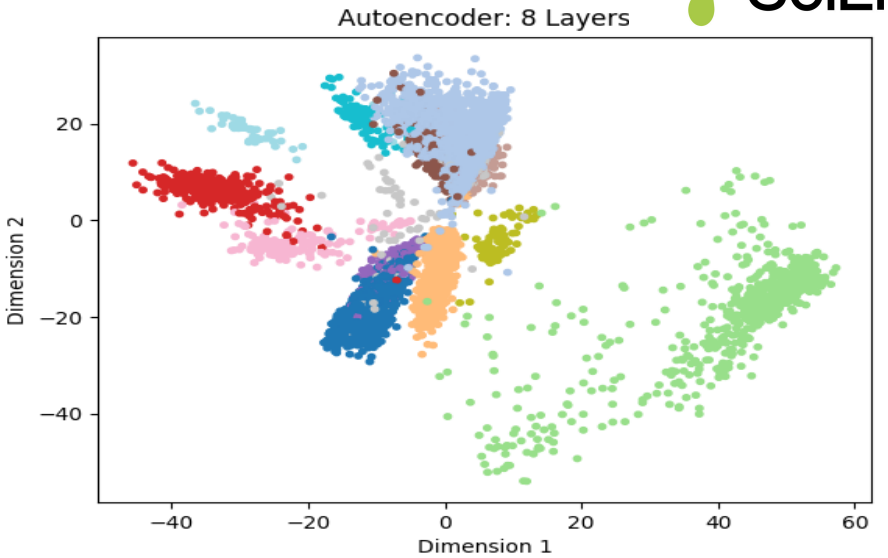
Autoencoder is an unsupervised Artificial Neural Network (ANN)

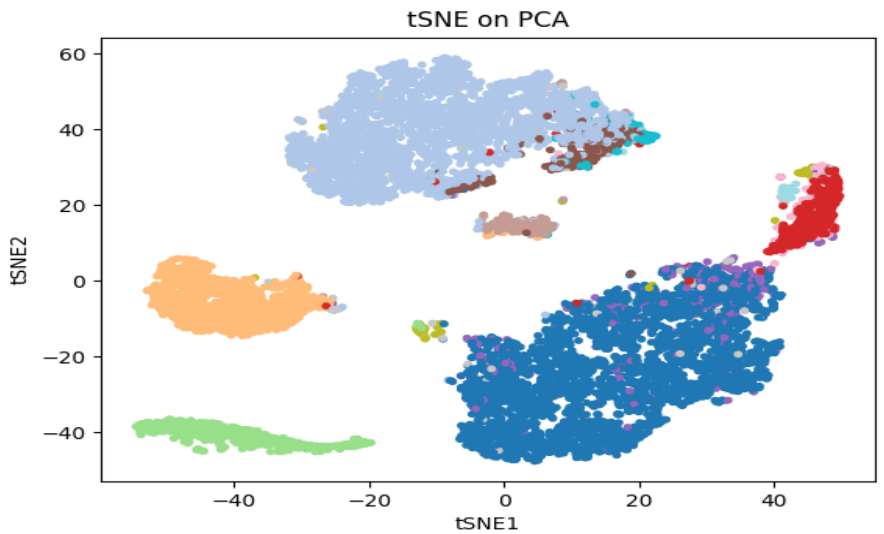
Single Cell analysis is unsupervised

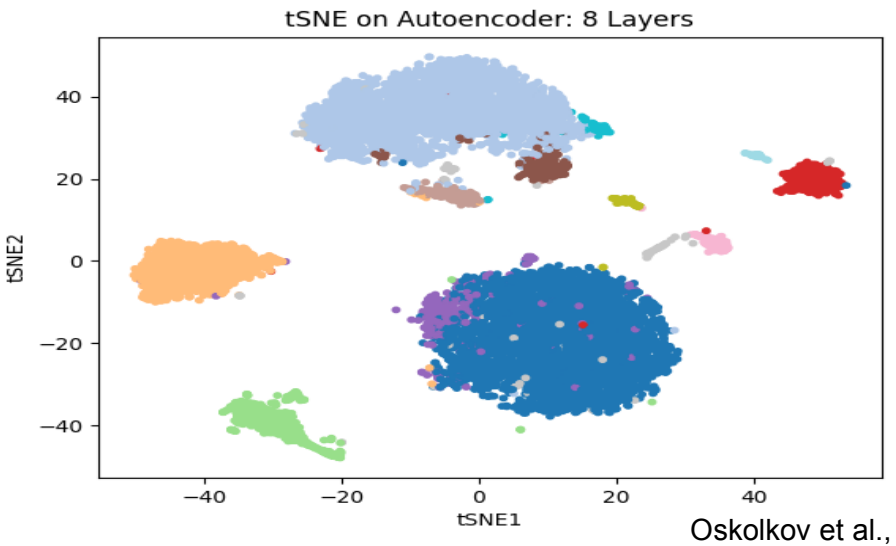
1) Dimensionality reduction: visualization
2) Clustering of cells: discover cell populations

Application of autoencoder for single cell



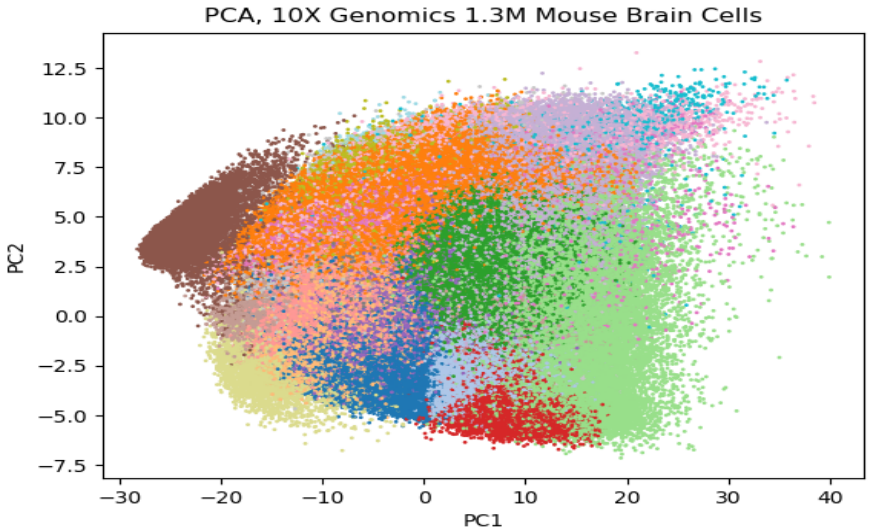


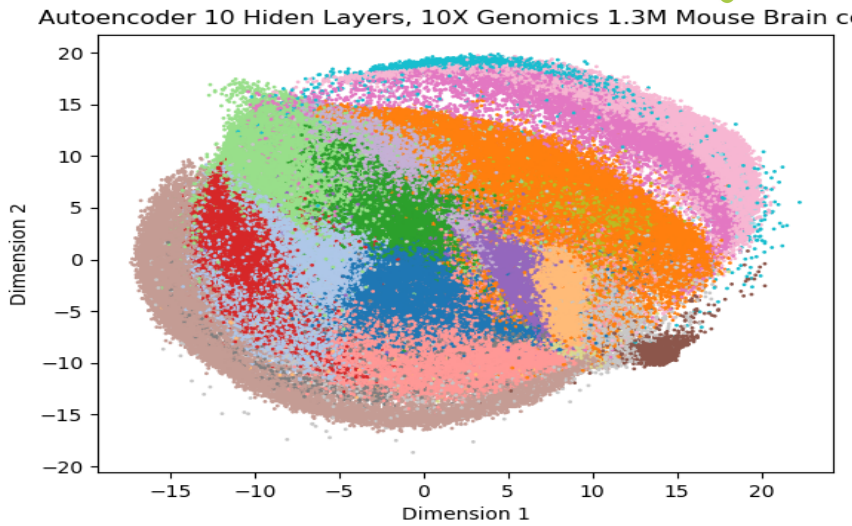


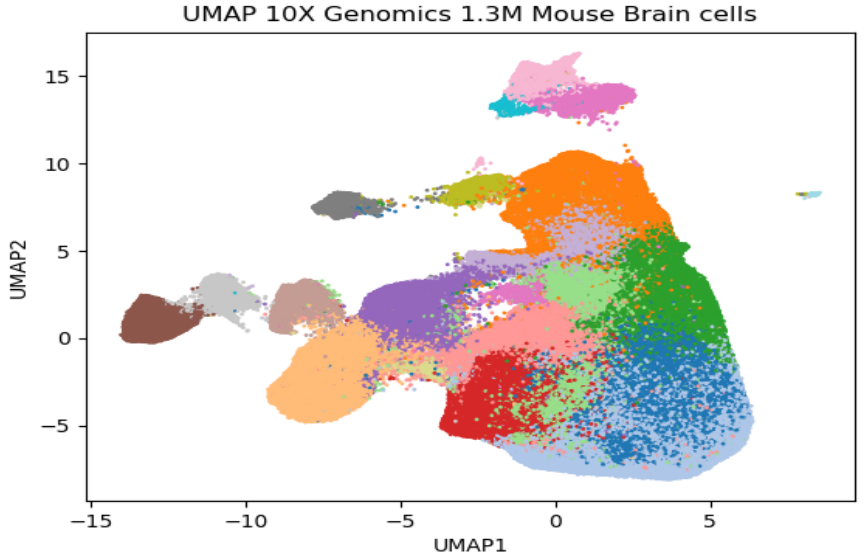


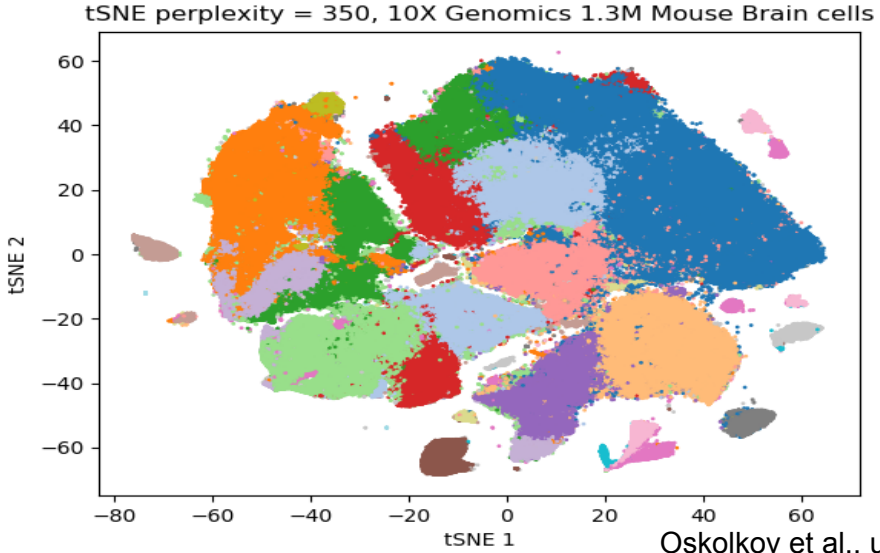
Oskolkov et al., unpublished

10X Genomics Mouse Brain: scRNAseq, 1.3M cells



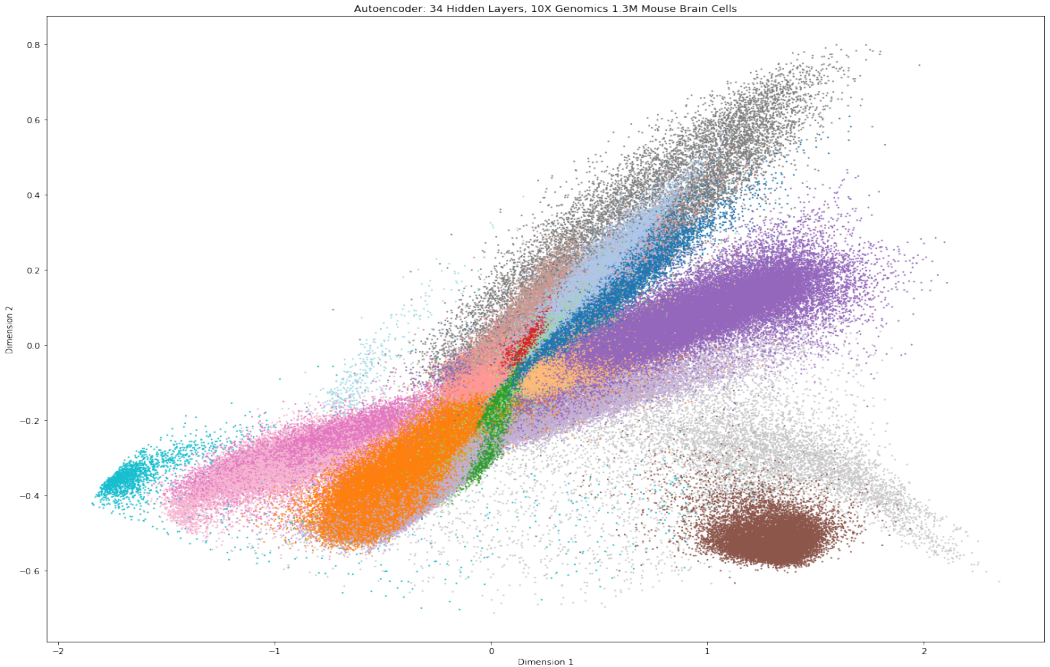




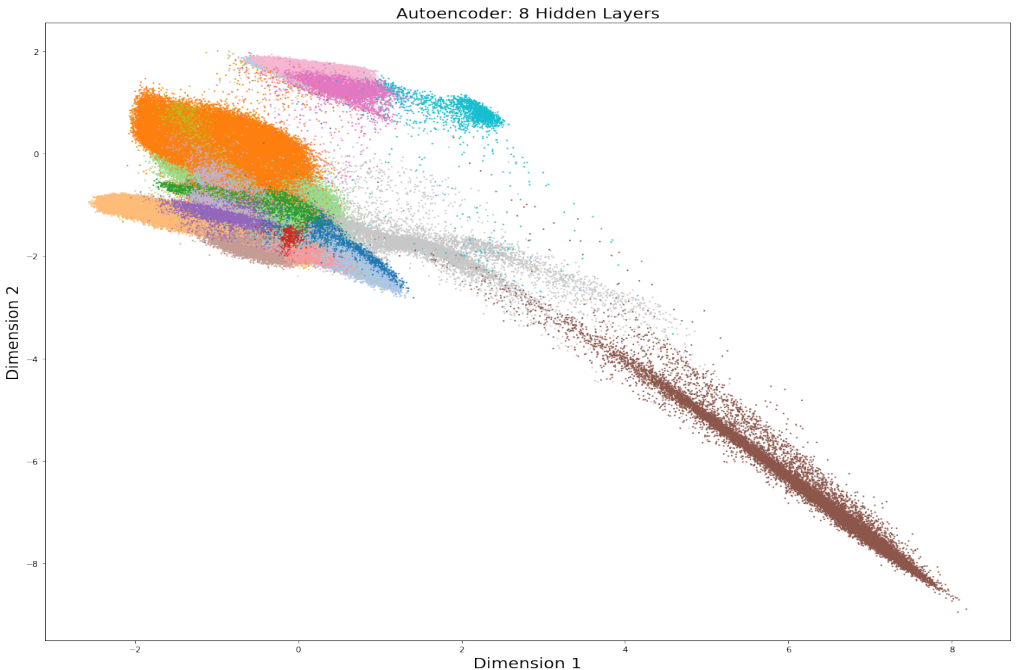


Oskolkov et al., unpublished

10X Genomics mouse brain 1.3M cells: applying Autoencoders with Tensorflow and Keras



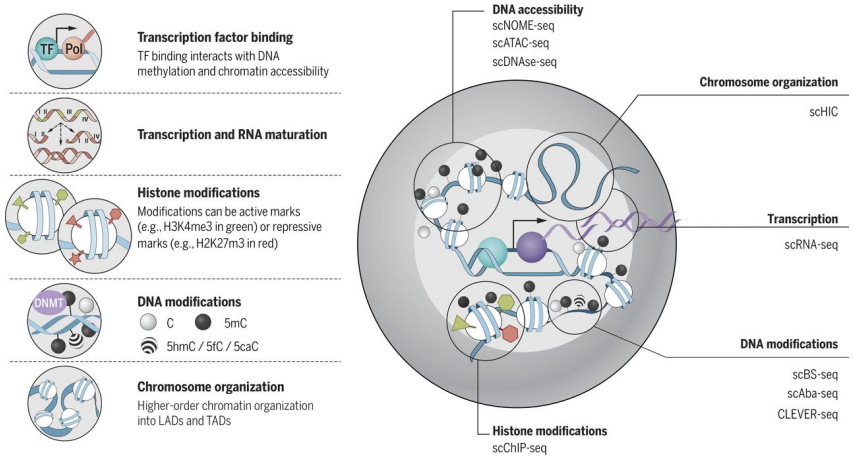
Autoencoders can be promising for non-linear data pre-processing, the bottleneck can potentially be fed to tSNE / UMAP



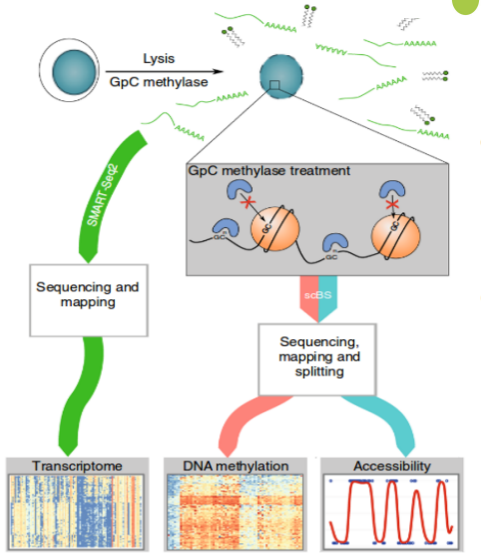
Autoencoders themselves are perhaps not optimal for visualization of scOmics

Oskolkov et al., unpublished

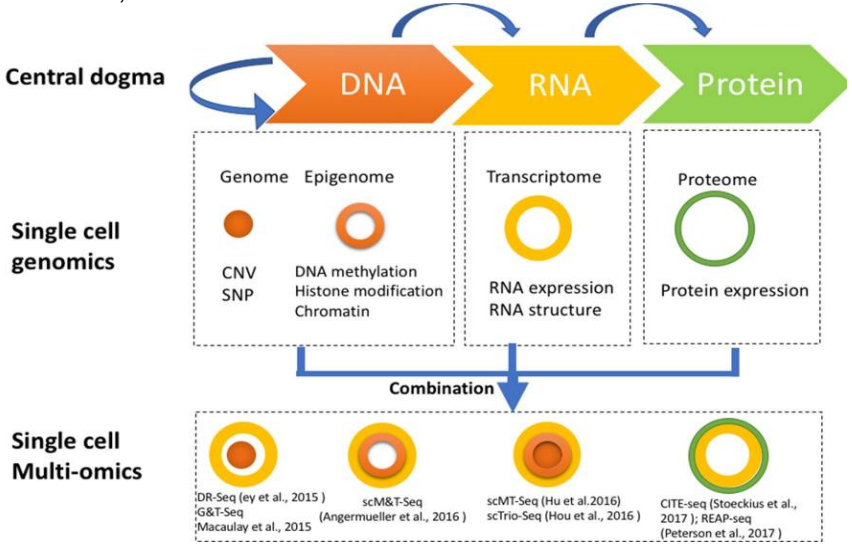
Multi-Omics in Single Cell Genomics



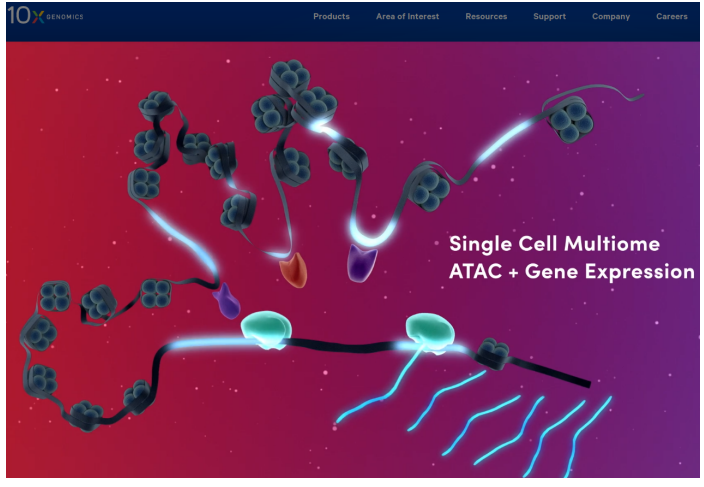
Kelsey et al., 2017, Science 358, 69-75



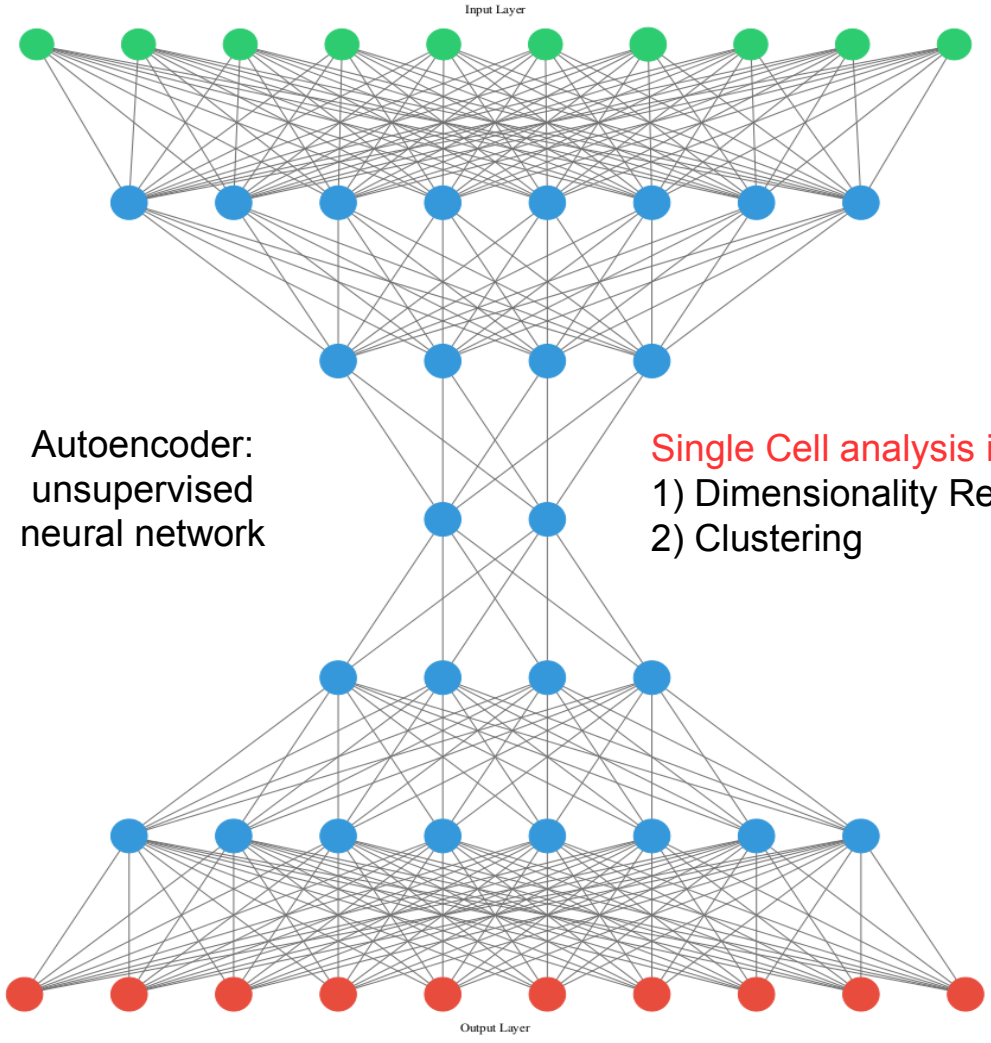
Clark et al., 2018, Nature Communications 9, 781



Hu et al., 2018, Frontier in Cell and Developmental Biology 6, 1-13

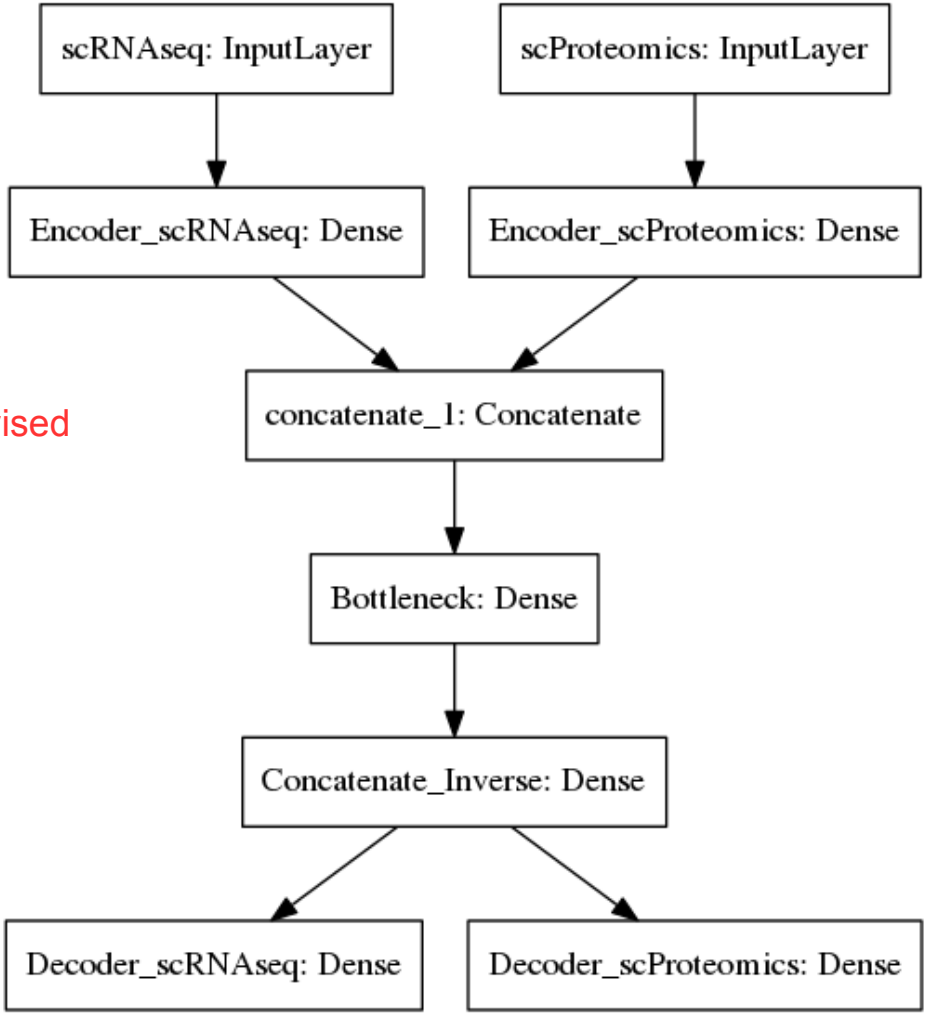


Autoencoder for data integration

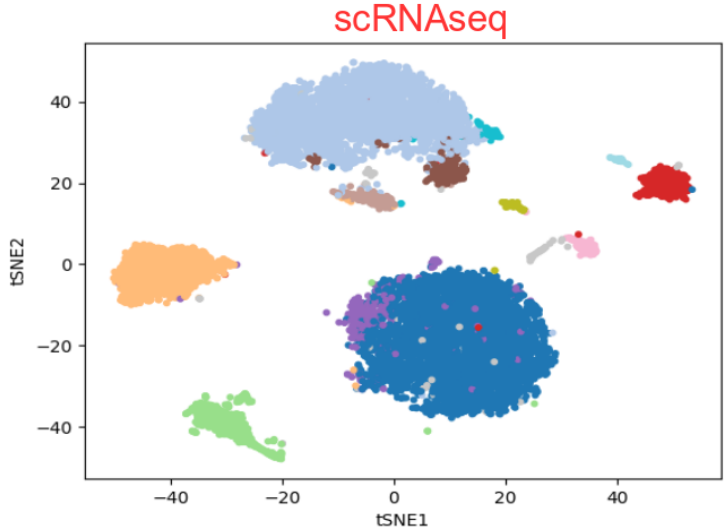


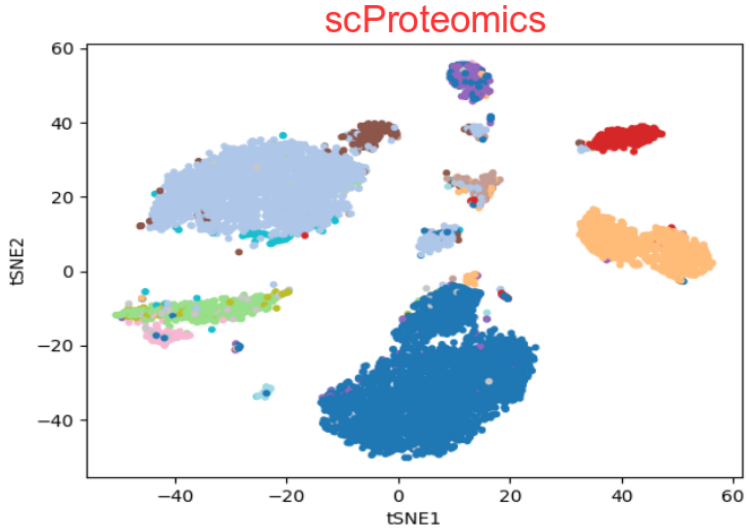
Autoencoder:
unsupervised
neural network

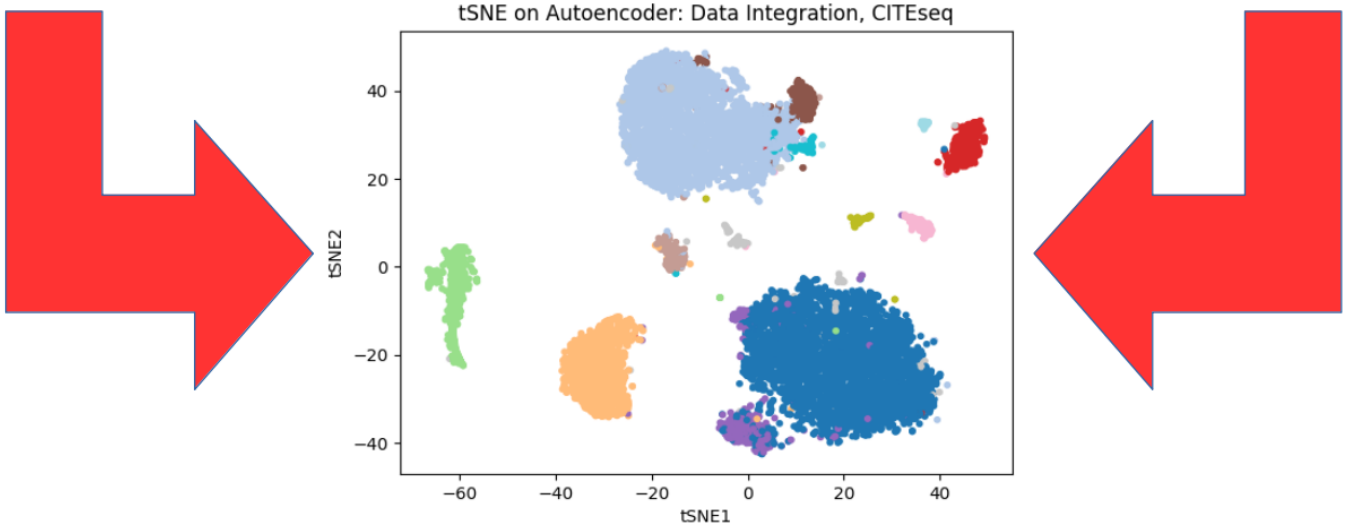
Single Cell analysis is unsupervised
1) Dimensionality Reduction
2) Clustering



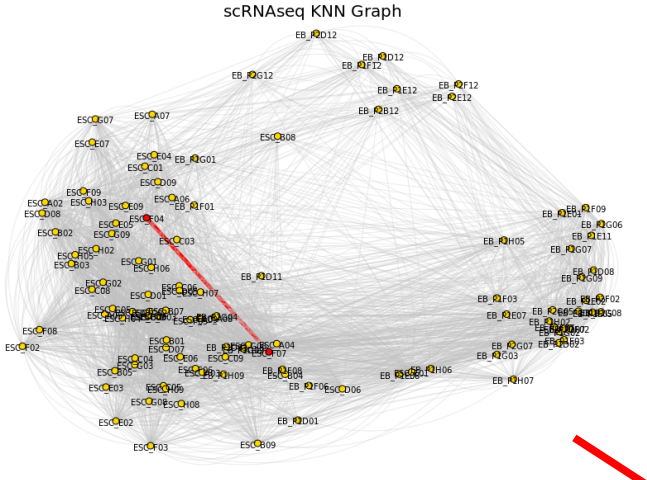
Autoencoder for data integration: CITEseq



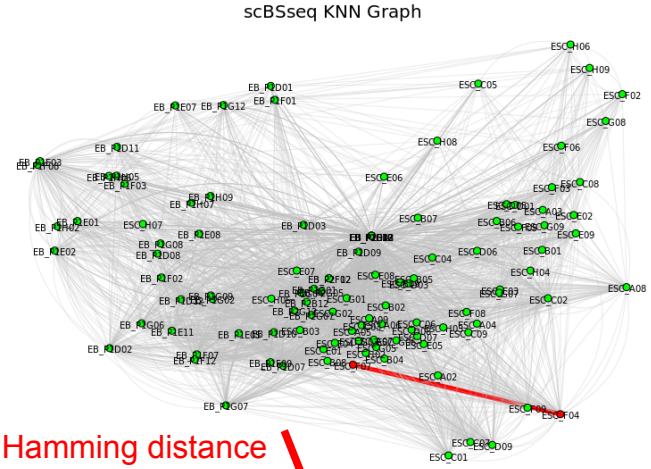




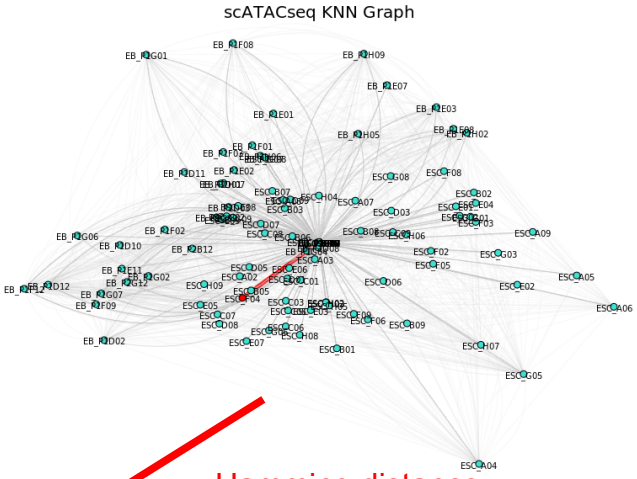
Graph Intersection Method



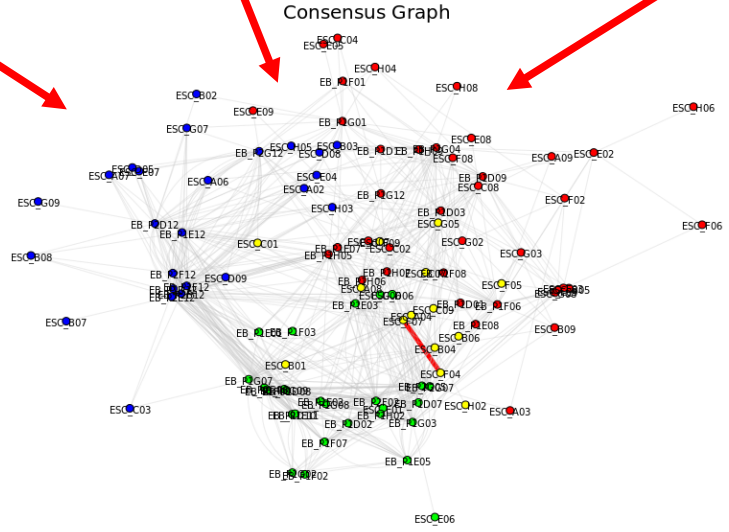
Euclidean distance



Hamming distance



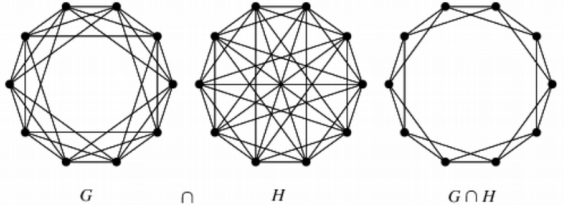
Hamming distance



Keep edges consistently present across the Omics

Graph Intersection

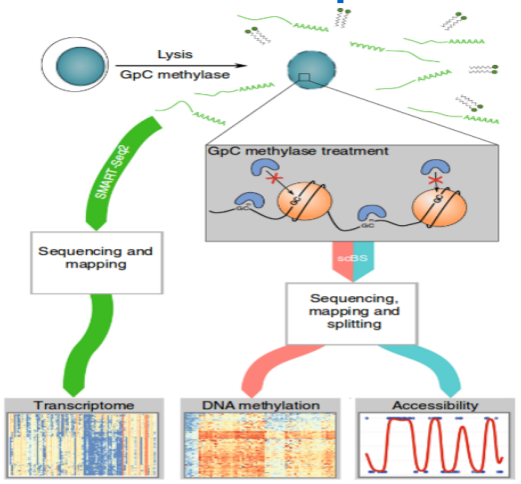
DOWNLOAD Wolfram Notebook



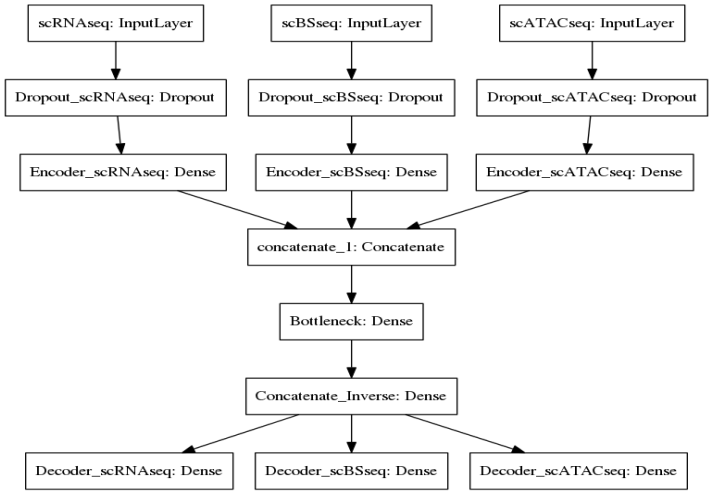
$G \quad \cap \quad H \qquad G \cap H$

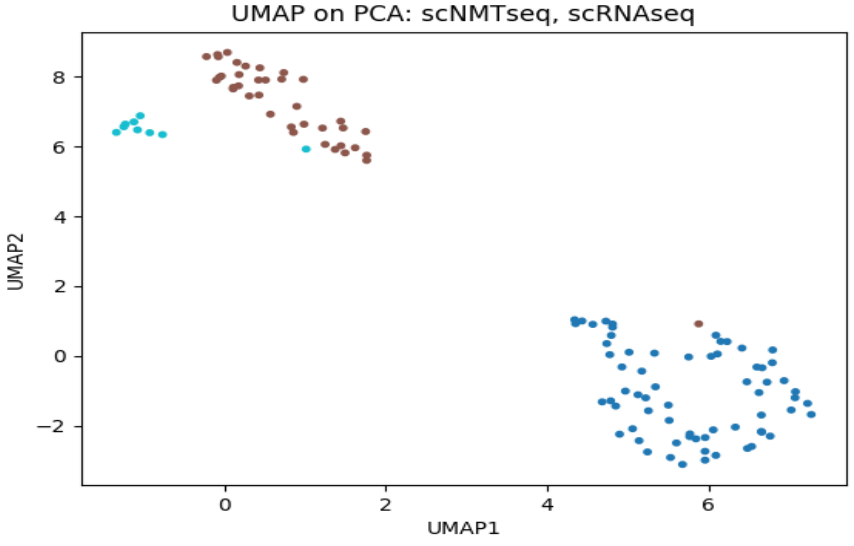
Let S be a set and $F = \{S_1, \ldots, S_p\}$ a nonempty family of distinct nonempty subsets of S whose union is $\bigcup_{i=1}^{p} S_i = S$. The intersection graph of F is denoted $\Omega(F)$ and defined by $V(\Omega(F)) = F$, with S_i and S_j adjacent whenever $i \neq j$ and $S_i \cap S_j \neq \emptyset$. Then a graph G is an intersection graph on S if there exists a family F of subsets for which G and $\Omega(F)$ are isomorphic graphs (Harary 1994, p. 19). Graph intersections can be computed in the Wolfram Language using `GraphIntersection[g, h]`.

scNMTseq: scRNAseq + scBSseq + scATACseq

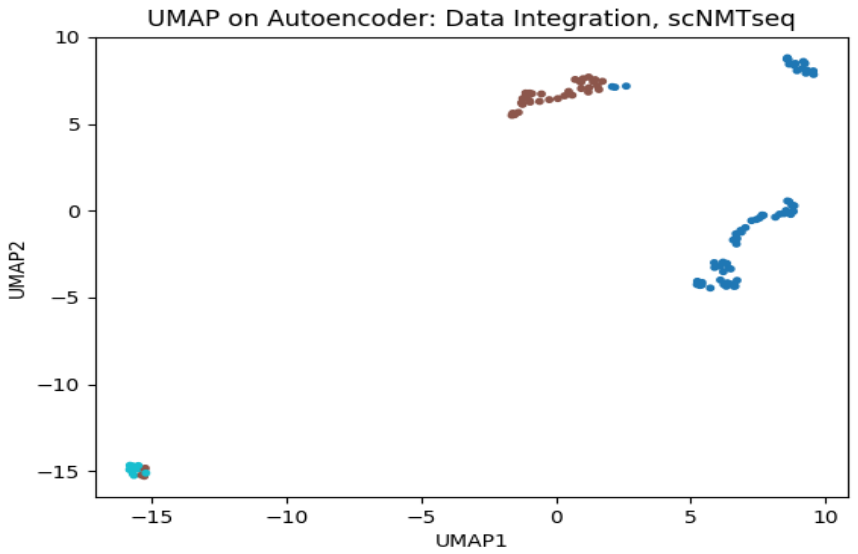


scNMTseq: Clark et al., 2018, Nature Communications 9, 781





From Single
To
multi-Omics



Session summary

Take home messages of the session:

1) Unsupervised ANNs are useful for nonlinear data exploration

2) Autoencoders are data-compression-type neural networks

3) Single cell data is Life Science-specific Big Data type

4) Autoencoders are used for dimension reduction in single cell

5) Autoencoders are common for single cell data integration

National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

Lunds Universitet