Dimension Reduction for Single Cell Data

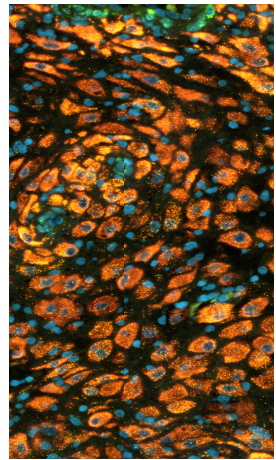
Franck Picard*
*Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon

franck.picard@ens-lyon.fr

Outline

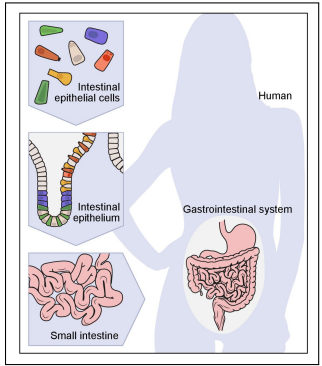
Cell biology revolution

- The cell has been discovered in the 17th century
- Cells are the basic unit of structure and function in living organisms
- Physiology emerges as the meta-cellular science (interaction between cells)



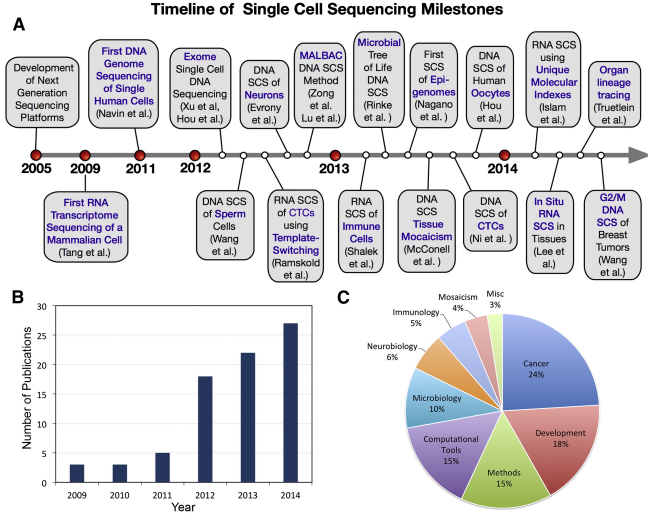
Main Biological Context

- Decypher cell diversity among living tissues
- Impossible before ∼2010 due to technical limitations
- Single Cell genomics: measure genomic features (DNA variations, RNA, Epigenome) at the single cell resolution



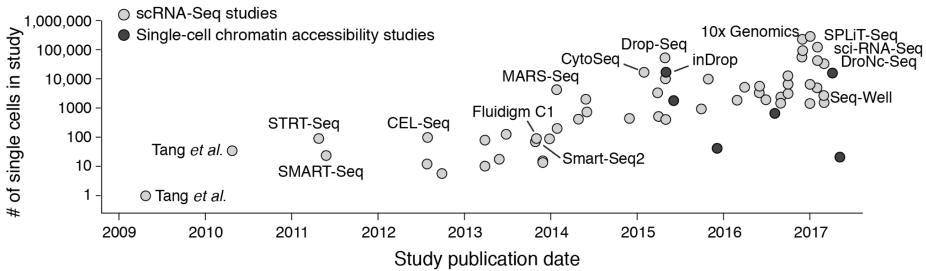
From the Human Cell Atlas [6]

A timeline: technologies



[9]

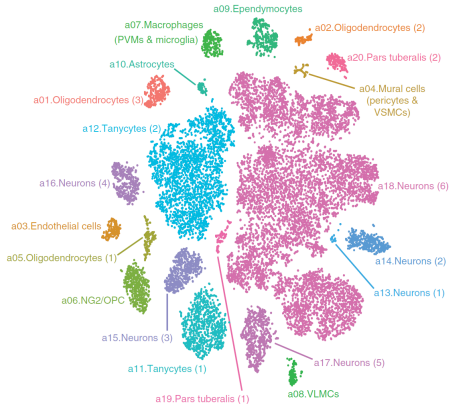
A timeline: produced data



[6]

Cell biology goes genome-wide

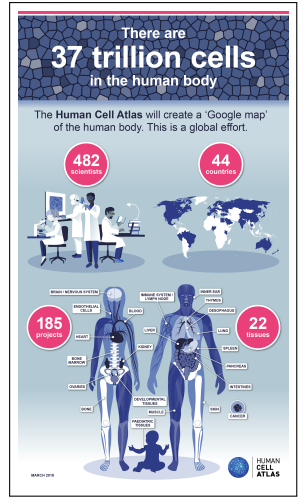
- Classify cells into distinct cell types
- Shape, location, interactions, function
- Recent technological breakthroughs allow the molecular characterization of cells



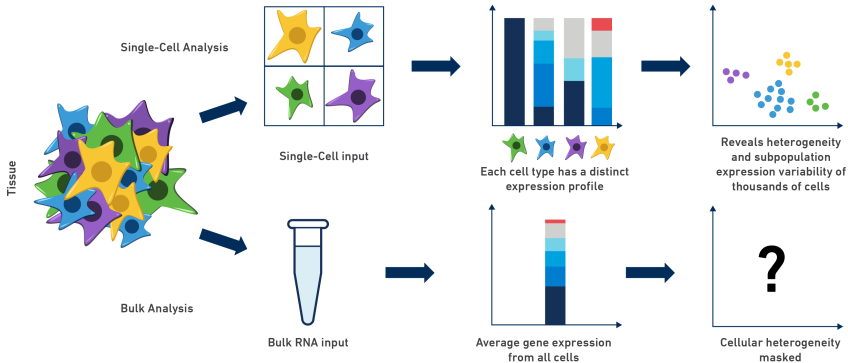
[1]

The human cell Atlas project

- comprehensive reference catalog of all human cells
- use stable properties, transient features, locations and abundances.
- describe each human cell by a defined set of molecular markers
- based on DNA variations, RNA, Epigenome at the single-cell resolution



Single-Cell from a statistician's perspective



From 10X Genomics

High-dimensional count data

x_{ij} = expression of gene j in cell i

$$\mathbf{X}_{n\times p} = \begin{bmatrix} & & & & & & \\ & & & x_{ij} & & & \\ & & & & & & \end{bmatrix} \begin{matrix} 1 \\ \vdots \\ n \end{matrix} \Big\} \text{cells}$$

$$\underbrace{1 \quad \dots \quad \dots \quad \dots \quad \dots \quad p}_{\text{genes}}$$

- **High dimension:** n grows but $\ll p$ & **Big Data:** n and p grow
- **Count data** with dropouts

Machine Learning Challenges for Single-Cell data analysis

- Dimension Reduction / Visualization
- Clustering cell-type discovery (non supervised and semi supervised)
- Datasets alignments for non-matched samples
- Catch cells-ecosystems behaviors
- Simulation of fake data
- Data integration
- Statistical Testing (compare gene expressions)

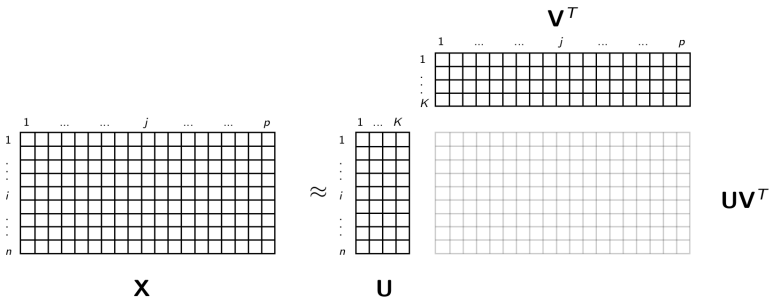
Outline

Matrix factorization: $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$

Cells: $\mathbf{U} \in \mathbb{R}^{n \times K}$

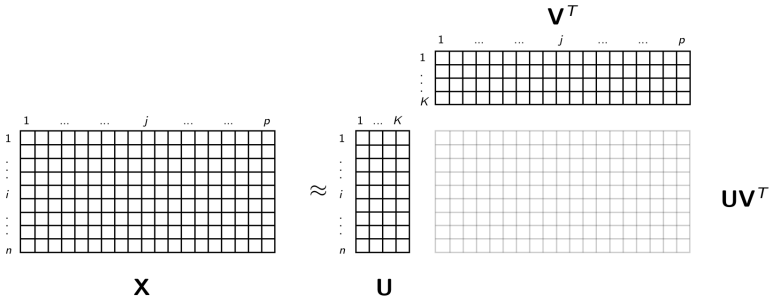
Genes: $\mathbf{V} \in \mathbb{R}^{p \times K}$

} Low dimensional representation

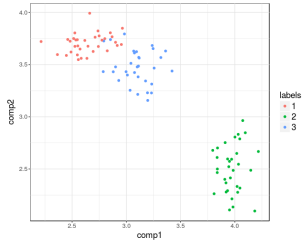


$\rightarrow$ Low-rank representation of $\mathbf{X}$

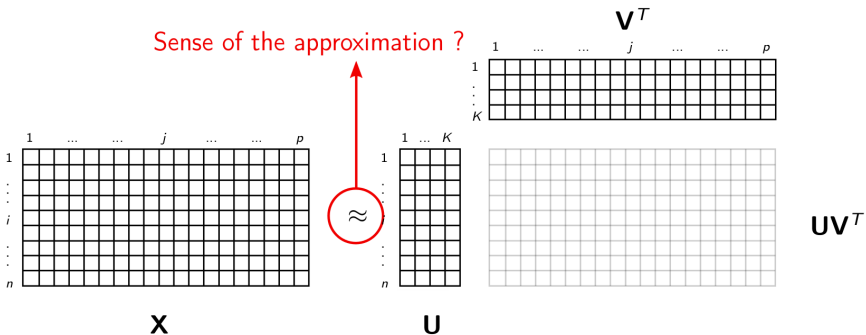
Matrix factorization: $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$



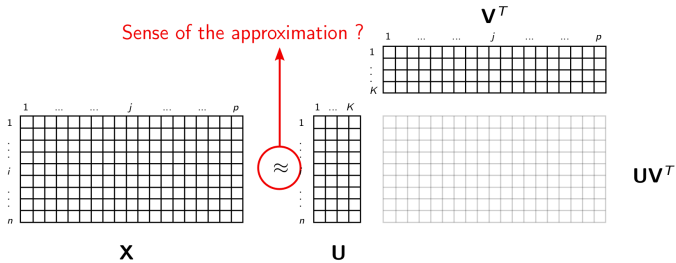
Data visualization:
scatter plot $(u_{i1}, u_{i2})_{i=1:n}$



Approximation $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$?



Approximation $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$?



Principal Component Analysis:

- Find a linear projection of $\mathbf{X}$ with maximum variance
- SVD algorithm: $\underset{\mathbf{U}\in\mathbb{R}^{n\times K},\mathbf{V}\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{U}\mathbf{V}^T\right\|_F^2$
- **Least squares approximation**

RNA-seq data = Counts

Relation between geometry and underlying model
$\|\cdot\|_2 \leftrightarrow$ Gaussian distribution

- First idea: $X_{ij} \sim \mathcal{P}(\lambda)$
- Highly expressed genes

  $\hookrightarrow$ large λ

  $\hookrightarrow$ Gaussian approximation

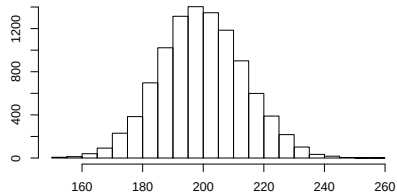


Figure: $\mathcal{P}(200)$ empirical distribution

RNA-seq data = Counts

Relation between geometry and underlying model
$\|\cdot\|_2 \leftrightarrow$ Gaussian distribution

- First idea: $X_{ij} \sim \mathcal{P}(\lambda)$
- Highly expressed genes

  $\hookrightarrow$ large λ

  $\hookrightarrow$ Gaussian approximation

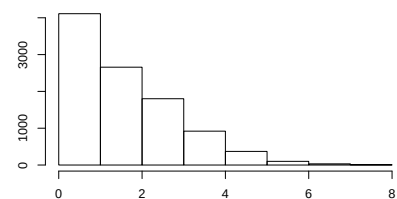


Figure: $\mathcal{P}(2)$ empirical distribution

Need for a probabilistic PCA

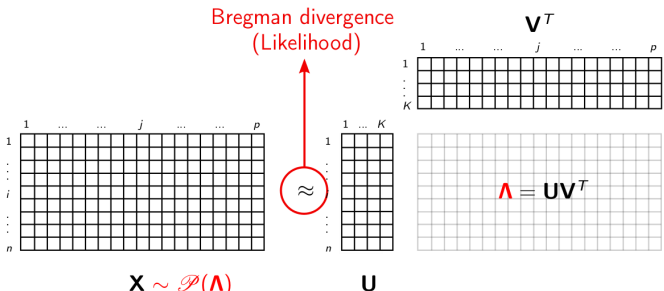
- **Over-dispersion** in RNA-seq data $\rightarrow \mathrm{Var}(X_{ij}) > \mathbb{E}[X_{ij}]$
- Single-cell data: **zero-inflation** $\rightarrow \mathbb{P}(X_{ij} = 0) > e^{-\lambda}$

Embed PCA with a **probabilistic model**

- $X_{ij} \sim$ probability distribution in the exponential family
- Factorization of $\mathbb{E}[\mathbf{X}]$ rather than $\mathbf{X}$
- Replace $\|\cdot\|_2$ approximation by likelihood-based approaches

Generalized PCA[2] and Poisson NMF [4]

- $X_{ij} \sim \mathcal{P}(\lambda_{ij})$ with the Poisson rate matrix $\mathbf{\Lambda} = [\lambda_{ij}]_{n \times p}$
- Decompose $\mathbb{E}[\mathbf{X}] = \mathbf{\Lambda}$ such that $\lambda_{ij} = \sum_k U_{ik} V_{kj}$



Random Intensity Models

- First Strategy : Poisson-Gamma Models :

$$\mathbf{\Lambda} \sim \Gamma(\alpha, \beta), \quad \mathbf{X} \mid \mathbf{\Lambda} \sim \mathcal{P}(\mathbf{\Lambda}), \quad \mathbf{X} \sim \mathcal{NB}$$

- Second Strategy : Poisson Log-Normal Models:

$$\mathbf{\Lambda} \sim \mathcal{N}(0, \Sigma), \quad \mathbf{X} \mid \mathbf{\Lambda} \sim \mathcal{P}(\exp \mathbf{\Lambda})$$

- Challenge : compute the posterior intensity:

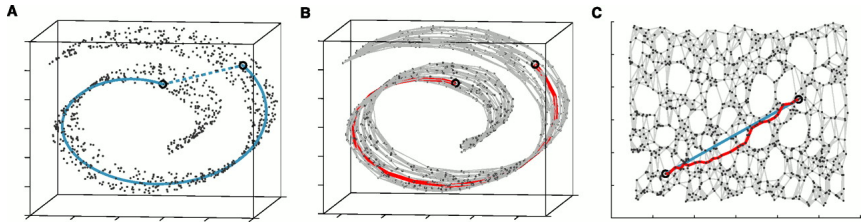
$$\mathbb{E}(\mathbf{\Lambda} \mid \mathbf{X})$$

- Estimate the factors as $\widehat{\mathbf{U}} = \mathbb{E}[\mathbf{U} \mid \mathbf{X}]$ and $\widehat{\mathbf{V}} = \mathbb{E}[\mathbf{V} \mid \mathbf{X}]$
- **Variational inference:** approximation of the posteriors

Outline

Beyond Linear projections

- Linear methods are powerful for planar structures
- High dimensional datasets are characterized by multiscale properties (local / global structures)
- May not be the most powerful for manifolds
- Non Linear projection methods aim at preserving local characteristics of distances



Stochastic Neighbor Embedding [8]

- $(X_1, \ldots, X_n)$ are the points in the high-dimensional space $\mathbb{R}^p$,
- Consider a similarity between points:

$$p_{i|j} = \frac{\exp(-\|X_i - X_j\|^2/2\sigma_i^2)}{\sum_{\ell \neq i} \exp(-\|X_\ell - X_j\|^2/2\sigma_\ell^2)}$$

- Hyper-parameter σ_i locally smooths the data, to be tuned

tSNE and Student / Cauchy kernels

- Consider $(Z_1, \dots, Z_n)$ are points in the low-dimensional space $\mathbb{R}^2$
- Consider a similarity between points in the new representation:

$$q_{i|j} = \frac{\exp(-\|Z_i - Z_j\|^2)}{\sum_{\ell \neq i} \exp(-\|Z_\ell - Z_j\|^2)}$$

- Robustify this kernel by using Student(1) kernels (ie Cauchy)

$$q_{i|j} = \frac{(1 + \|Z_i - Z_j\|^2)^{-1}}{\sum_{\ell \neq i} (1 + \|Z_i - Z_\ell\|^2)^{-1}}$$

KL optimization by Gradient descent

- The Kullback-Leibler divergence can be used as a measure of dissimilarity between distributions:

$$KL(p, q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

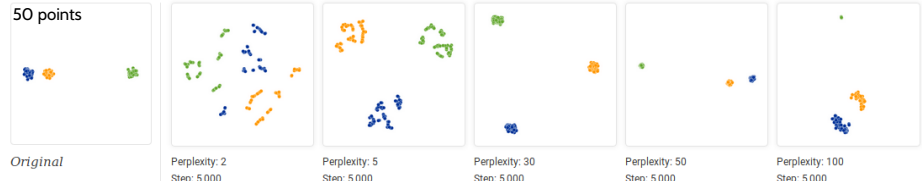
- Minimize the KL between p and q to find $Z \in \mathbb{R}^2$ such that:

$$C(Z) = \sum_{ij} KL(p_{ij}, q_{ij})$$

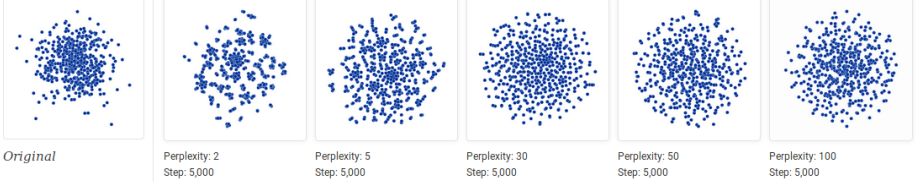
$$\left[\frac{\partial C(Z)}{\partial Z}\right]_i = \sum_j (p_{ij} - q_{ij})(Z_i - Z_j)$$

- Gradient descent with momentum to speed up and improve convergence
- Random initialization

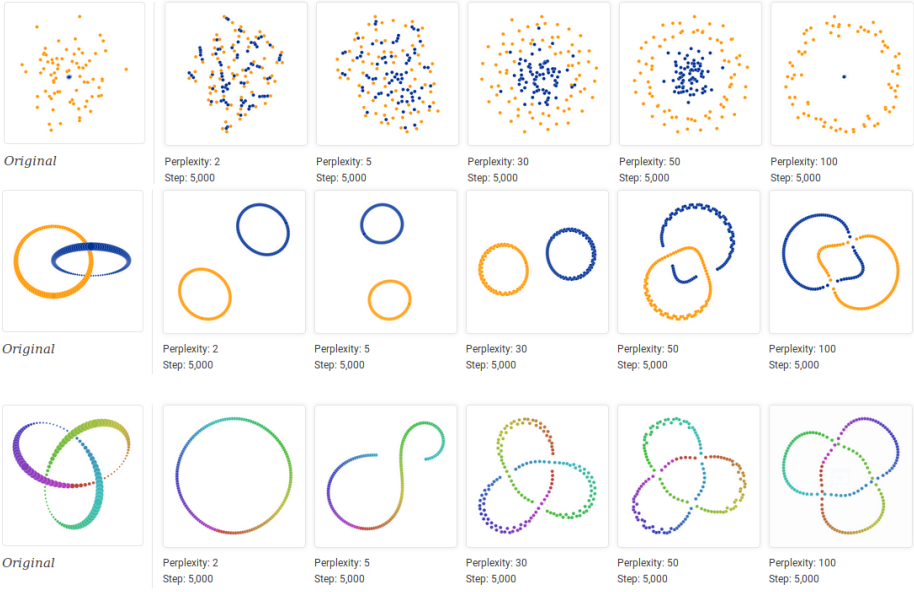
tSNE does not account for between-cluster distance



What about random noise ?



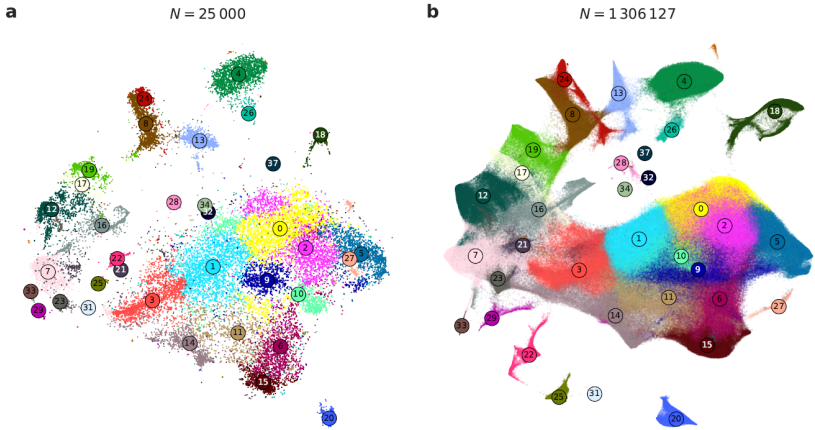
Catching Complex Geometries



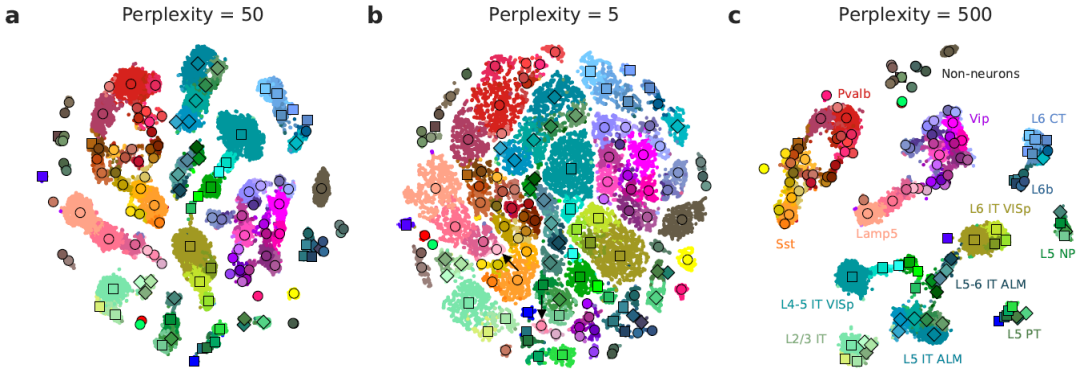
Properties of t-SNE

- Good at preserving local distances (intra-cluster variance)
- Not so good for global representation (inter-cluster variance)
- Good at creating clusters of points that are close, but bad at positioning clusters wrt each other
- Does not handle well high dimensional data (preliminary PCA and feature selection)
- Sensistive to the calibration of the hyperparameter (smoothing)
- Reproducibility of results due to stochastic optimization

tSNE on single cell Gene Expression data [3]



Influence of parameter tuning



Comparisons

- The field is very active and comparisons are performed extensively
- Tuning is a challenge [5] especially for non-linear methods
- Linear methods are robust !
- How to compare dimension reduction methods ?
- Confusion between dimension reduction and clustering ?

Research Challenges

- → What are the statistical / probabilistic foundations of Stochastic Neighbor Embedding ?
- → Can we define a common statistical framework for seemingly unrelated dimension reduction methods ?
- → How to combine non-linear dimension reduction and clustering ?

References

[1] J. N. Campbell, E. Z. Macosko, H. Fenselau, T. H. Pers, A. Lyubetskaya, D. Tenen, M. Goldman, A. M. Verstegen, J. M. Resch, S. A. McCarroll, E. D. Rosen, B. B. Lowell, and L. T. Tsai. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.*, 20(3):484–496, Mar 2017.

[2] Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, pages 617–624, 2001.

[3] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *bioRxiv*, 2018.

[4] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[5] F. Raimundo, C. Vallot, and J. P. Vert. Tuning parameters of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol*, 21(1):212, 08 2020.

[6] A. Regev, S. A. Teichmann, E. S. Lander, and I. et al. Amit. The Human Cell Atlas. *Elife*, 6, 12 2017.

[7] H. van Assel, T. Espinasse, J. Chiquet, and F. Picard. A probabilistic graph coupling view of dimension reduction. *Arxiv*, (2201.13053):1–15, 2022.

[8] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine*