



Subtypes estimation in single-cell multi-omics data via S-multi-SNE

Theodoulos Rodosthenous [✉], Vahid Shahrezaei [✉], Marina Evangelou [✉]

Department of Mathematics, Imperial College London, London, SW7 2AZ, UK

ARTICLE INFO

Keywords:

S-multi-SNE
Single-cell multi-omics
Cell subtype estimation

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) has rapidly become an essential method in modern biology, as it allows global characterisation of transcriptomics in individual cells from different tissues, or organisms. scRNA-seq data are collected from a single individual organism and they are composed as mRNA count matrices with rows representing the cells and columns the genes. New technologies further allow multi-omics data to be obtained on the same set of single cells.

In this paper, multi-SNE and S-multi-SNE, two multi-view dimensionality reduction approaches, are adapted for visualization and classification of cellular subtypes, two important challenges in the analysis of scRNA-seq data. S-multi-SNE has been adapted for cell subtype estimation of scRNA-seq data. In a series of experiments we illustrate that it is possible to identify cell subtypes by leveraging information from individuals from the same species, from different species, and when data are generated by different sequencing technologies. In the conducted analyses, we show that S-multi-SNE consistently outperformed several other machine learning techniques and, in most comparisons, surpassed existing reference-based single-cell classification algorithms.

We further illustrate how two single-cell multi-omics datasets, scATAC-seq and scRNA-seq, can be integrated together through S-multi-SNE for improved cell subtype identification.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has become an essential method in modern biology, as it allows global characterization of transcriptomics in many individual cells (Andrews et al., 2021). scRNA-seq data are collected from a single individual organism, tissue, or cell culture for characterizing global cellular heterogeneity and for identifying the different cell types present in the sample. They are composed as mRNA count matrices with rows representing the cells and columns the genes. The analysis of scRNA-seq data entails several statistical challenges due to the different characteristics of the data that include high dimensions, sparsity, technical noise, and batch effects (Lähnemann et al., 2020).

There are several grand challenges in the statistical analysis of scRNA-seq data one of which is mapping cells to cell subtypes (Lähnemann et al., 2020). Arendt et al. (2016) define the cell subtypes as ‘a set of cells in an organism that change in evolution together, partially independent of other cells, and are evolutionarily more closely related to each other than to other cells’. In agreement with other researchers, Pliner et al. (2019) named four key challenges for the annotation of cells:

- (i) this process is expensive, time consuming and prone to errors, as it requires extensive literature review and multiple manual tests,
- (ii) additional data and parameter adjustments require manual re-evaluation of previous annotations,
- (iii) annotations are not easily transferred between datasets generated by independent groups, and
- (iv) annotations are typically ad hoc.

While we acknowledge the standard practice of annotation lies in linear latent space, current evidence from the literature show an increasingly support towards the incorporation of non-linear methods in annotation pipelines. Bi et al. (2024), demonstrate that semi-supervised pipelines combining batch correction, non-linear dimensionality reduction via UMAP, and clustering significantly improve both annotation of known cell types and discovery of novel types. Similarly, the SnapATAC2 pipeline (Zhang et al., 2024) utilises spectral embedding, a non-linear dimensionality reduction method belonging to the same manifold learning category as t-SNE for single-cell multi-omics analysis, highlighting the field’s growing confidence in the biological fidelity preserved by such methods as part of end-to-end analytical workflows.

* Corresponding author.

E-mail addresses: theodoulosr@gmail.com (T. Rodosthenous), v.shahrezaei@imperial.ac.uk (V. Shahrezaei), m.evangelou@imperial.ac.uk (M. Evangelou).

<https://doi.org/10.1016/j.jtbi.2026.112442>

Received 27 February 2025; Received in revised form 27 February 2026; Accepted 15 March 2026

Available online 25 March 2026

0022-5193/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Recent advances in large-scale single-cell transcriptomic profiling have enabled the systematic characterisation and subsequent annotation of various cell-types. CellTypist is a tool that facilitates in automated and reproducible cell type annotation at scale (Domínguez Conde et al., 2022).

A range of different approaches have been adapted for classification of single cells to their respective cell subtypes. This can be simply achieved by clustering the cells and then using known cell subtype-specific biomarker genes to annotate them manually - following the above definition of cell subtypes by Arendt et al. (2016). In recent years, the focus has shifted from cluster-based discovery and manual annotation new subtypes, to the automatic estimation of cell subtypes based on a set of predefined marker genes or publicly available well-annotated reference datasets, sometimes known as cell atlases (Zhao et al., 2020; Lähnemann et al., 2020; Pasquini et al., 2021).

Numerous reference-based classification methods have been developed that identify characteristic features (e.g. marker genes) for each subtype, from which the subtype estimation of query cells is achieved. Zhao et al. (2020) provide a comprehensive evaluation of several such methods, in which they concluded that out of the compared classification tools, none perfectly and completely solve all problems. SciBet (Li et al., 2020), SingleR (Aran et al., 2019), scMCA (Han et al., 2018) and scMap (Kiselev et al., 2018) use similarity measures to determine the most relevant reference cells for each query cell. CHETAH (de Kanter et al., 2019) and scClassify (Lin et al., 2020) are tree-based solutions that utilise the hierarchical structure of the cells, while Seurat (Satija et al., 2015) and SingleCellNet (Tan and Cahan, 2019) use random forests to estimate the cell subtypes. The classification tasks of scPred (Alquicira-Hernandez et al., 2019), CaSTLe (Lieberman et al., 2018), CellFishing (Sato et al., 2019) and scID (Boufeua et al., 2020) are based on SVM, XG-Boost, kNN and LDA, respectively. More recently, Wei and Zhang (2021) and Chen et al. (2021) proposed semi-supervised solutions for cell subtype estimation. Wei and Zhang (2021) combines the construction of a graph Laplacian matrix with sparse LogReg, while scSemiCluster (Chen et al., 2021) is a solution based on deep discriminative and deep generative clustering. Several new methods leveraging large language models are being developed to create foundational models for scRNA-seq, and these approaches also show promise for addressing cell annotation challenges (Chen et al., 2023; Hao et al., 2024; Cui et al., 2024).

In parallel development, new technologies allow the acquisition of multi-omics on the same set of cells (Baysoy et al., 2023). For example, the Assay of Transposase Accessible Chromatin with high-throughput sequencing (scATAC-seq) measures the chromatin accessibility in single cells (Yu et al., 2020); it is often measured along with scRNA-seq data. There are specific challenges associated with analysis of multi-omics single-cell data and this is an active area of research (Heumos et al., 2023). In this manuscript we are taking advantage of single-cell multi-omics data for addressing key biological questions.

To address the challenges of cell type classification and of the analysis of single-cell multi-omics data, we explored the implementation of novel multi-view manifold learning methods. We recently proposed several methods for multi-view data visualization via manifold learning (Rodosthenous et al., 2024). Multi-view data consist of multiple data types generated for the same samples (e.g., multi-omics data). One of these methods multi-SNE, based on the well known t-distributed Stochastic Neighbour Embeddings (t-SNE) (van der Maaten and Hinton, 2008), was shown to work most robustly. In the work of Rodosthenous et al. (2024) the application of Multi-SNE was illustrated for the visualization of multi-omics single-cell data. S-multi-SNE is an extension of multi-SNE that performs semi-supervised classification through non-linear dimensionality reduction (Rodosthenous et al., 2021a). S-multi-SNE treats one of the data-views as response, containing the labelling information. In contrast to many classification solutions, the training and testing sets are combined together, separated by the missing (Not Available - NA) values found in the data-view that represents the response. In this paper, we have adapted S-multi-SNE for cell subtype classification

of scRNA-seq data using reference datasets, similar to the other methods proposed in the literature. The main idea of the adapted version is the inclusion of the reference panels as the training part in one of the data-views, whereas any unlabelled samples are treated as the testing part. The adapted version of S-multi-SNE is able to provide estimated classes for the unlabelled samples and low-dimensional embedding for all samples, which can be used to visualise labelled and unlabelled samples together, a property inherited from multi-SNE.

Our S-multi-SNE framework fundamentally differs from standard t-SNE applications by capturing complementary labeling information which provide the benefit of mitigating information loss typically associated with single-view non-linear reductions. Through its semi-supervised nature, the proposed solution ensures that dimensionality reduction is guided by known cell type information, preserving class-relevant structure in the low-dimensional space.

The structure of the manuscript is as follows: Section 2 presents existing machine learning classification approaches, and single-cell classification approaches that are the reference-based solutions CHETAH, scPred and SingleCellNet. Single-cell pre-processing and batch correction steps are also presented. In the same section, the two multi-view approaches: multi-SNE and S-multi-SNE and their adaptation for single-cell type classification are presented. Section 3 presents the publicly available data analysed as part of this work. Section 4 presents a series of experiments for single-cell type classification using different training and test sets. The application of both multi-SNE and S-multi-SNE on scRNA-seq data is illustrated as well as a comparison of S-multi-SNE with several standard machine learning and three reference-based solutions CHETAH, scPred and SingleCellNet is presented. Section 5 presents the analysis of single-cell multi-omics data. Through the conducted experiments, we illustrate the promise of multi-SNE and S-multi-SNE for single-cell subtype annotation and data visualization. We further illustrate how is possible to obtain accurate cell type classifications by leveraging information from individuals from the same species, from different species, and when data are generated by different sequencing technologies, as well as taking advantage of single-cell multi-omics datasets.

The code is available at https://github.com/theorod93/scSubtypes_estimation.

2. Methods

2.1. S-multi-SNE

Multi-SNE proposed by Rodosthenous et al. (2024) is an extension of t-SNE tailored for multi-view data visualization and dimensionality reduction. Multi-SNE aims to generate a unified low-dimensional embedding by simultaneously optimizing the information from multiple data views. It is primarily designed to provide a single visualization of multi-view data and to facilitate downstream clustering. In this paper, multi-SNE is utilised to visualise the multi-view data of interest.

Multi-SNE produces a single, unified low-dimensional embedding of samples by integrating information across all M data-views simultaneously. For each data-view m , the algorithm first computes a pairwise probability distribution P^m over samples using a Gaussian kernel. A corresponding low-dimensional distribution Q is defined over the embedding space using the Student's t-distribution with one degree of freedom, consistent with standard t-SNE. The cost function minimises the weighted sum of KL-divergences between each view-specific distribution P^m and the shared embedding distribution Q :

$$C_{\text{multi-SNE}} = \sum_m w^{(m)} KL(P^m || Q), \quad (1)$$

where the weights $w^{(m)}$ represent the weighting factor for each data-view, with default being $w^{(m)} = \frac{1}{M}$.

S-multi-SNE is a semi-supervised extension of multi-SNE that incorporates class label information as an additional data-view. The labelling

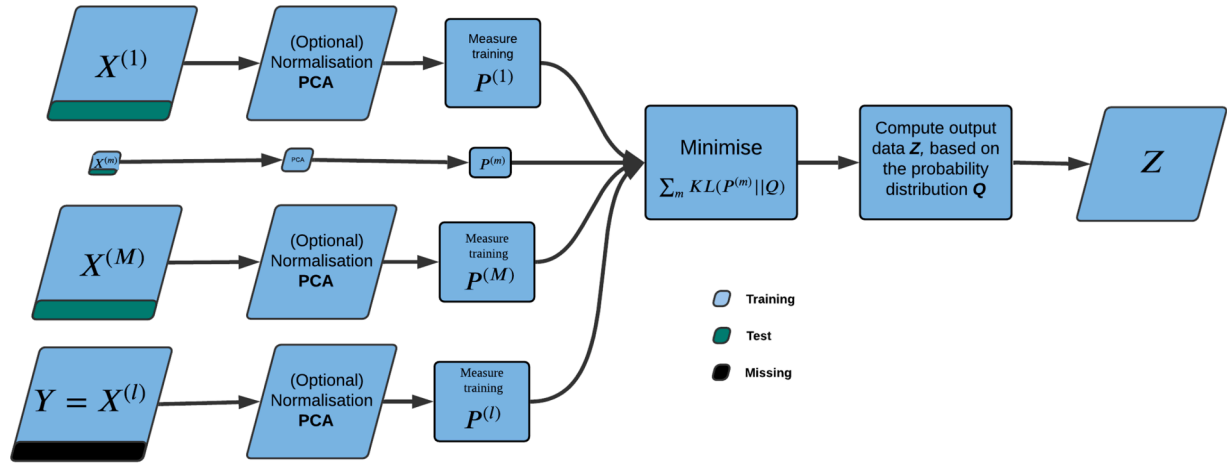


Fig. 1. Schematic overview of the S-multi-SNE algorithm. S-multi-SNE allows for "missing" values to be integrated along with data-views that do not have any missing information.

information is encoded as a binary matrix of size $N \times C$, where C is the number of classes and each entry indicates class membership. Crucially, this label view is only available for the training samples; test samples have missing label entries (Fig. 1).

The algorithm handles this asymmetry through an indicator matrix $\mathfrak{I}^l$, which masks out pairwise contributions involving test samples in the labelling view. The modified cost function becomes:

$$C_{S\text{-}multi\text{-}SNE} = \sum_m^{M-1} w^{(m)} KL(P^{(m)} || Q) + \mathfrak{I}^{(l)} w^{(l)} KL(P^{(l)} || Q) \quad (2)$$

This ensures that labelled training samples are drawn together in the embedding according to their class membership, while test samples are simultaneously projected into the same space by virtue of their other data-view similarities.

Although t-SNE, and by definition its extensions, does not preserve exact distances, it is fundamentally designed to preserve local neighbourhood structure. While the global layout may not accurately reflect the original data structure, the local organisation of points identification remains well-preserved.

S-multi-SNE (Rodosthenous et al., 2021a) is an extension of multi-SNE for semi-supervised classification and visualization of multi-view data. S-multi-SNE integrates label information from training samples into the dimensionality reduction process, enabling both effective visualization and classification of unlabelled test samples. S-multi-SNE is designed for scenarios with multi-view data and partially labelled samples. Since, all individuals (training and test) are analysed simultaneously, S-multi-SNE has the ability of projecting the test single-cell samples in the same set of axes as the training cells.

For further details on the underlying mathematics for multi-SNE and S-multi-SNE we refer the readers to the original papers (Rodosthenous et al., 2021b, 2024).

In the next subsection we present how S-multi-SNE has been adapted for cell type classification of single-cell data.

2.2. Semi-supervised cell-subtype allocation using S-multi-SNE

Consider two single-cell datasets, e.g. scRNA-seq or multi-omics datasets, named A and B . Commonly the dataset of the two sequenced individuals, A and B , (of the same or different species) share the same features (e.g. genes in scRNA-seq or orthologous genes in the case of analysing data from different species). For classification tasks (including cell subtype estimation), one dataset represents the reference dataset with well-curated labels being available as an additional data-view, and is used as training set (e.g. A), and the other data set will be annotated

using supervised learning, and will be considered as the test set (e.g. B). To estimate the cells subtype through S-multi-SNE, the two batches of data (individuals) are corrected and vertically integrated by using existing and reliable tools, specifically developed for single-cell data (see next subsection). The integrated data are then taken as a data-view.

The labelling data-view for reference dataset (e.g. A), which describes the subtypes of the single cells, is denoted by $X^L \in \mathbb{R}^{N \times C}$, where N is the total number of cells, and C is the total number of classes. This labelling data-view is constructed as a binary matrix, for which the cell (i, j) takes the value 1 if sample i belongs to class j , and zero otherwise. In other words, each column of the labelling data-view corresponds to each cell subtype. For the test dataset the labels on this data-view will be set as NA as missing. S-multi-SNE then can be used to embed both datasets jointly. The output of S-multi-SNE is then a low-dimensional embeddings of the two samples A and B , including the ones in the test set with missing labels. These projections can then be treated as input features into any general-purpose classifier to predict the classes of the unlabelled samples. Here we have used the kNN algorithm by Fix and Hodges (1951). We note that this subtype estimation process does not use any gene markers, nor their relationships with cell subtypes.

The following steps describe the methodology used in performing semi-supervised cell subtype estimation via S-multi-SNE:

1. Pre-processing the data and ensuring that both datasets describe the same genes.
2. Data integration and batch correction via Seurat v4.
3. Applying S-multi-SNE to retrieve low-dimensional embeddings of all samples (training and test).
4. Subtype estimation of the testing set (with missing labels) via kNN on the obtained low-dimensional embeddings based on the labels of the training set (known labels).

In step 2, several different integration and batch correction methods specifically developed for single-cell data are explored, with the batch correction algorithms applied presented next.

In step 4, a range of k values was investigated. The results presented in this manuscript are based on the best-performing k value within the range $S_k = \{2, 5, 8, 10, 20\}$. Although in most solutions the optimal value was found to be $k = 5$, the solution was found not to be sensitive to the choice of the k value with almost identical classification performance for all values of $k \in S_k$.

Detailed descriptions of the general and single-cell specific classification algorithms, as well as the batch correction algorithms used throughout this manuscript, are provided in the Supplementary Material.

3. Datasets

Although, single-cell data have only recently been available in abundance, the increasing number of publicly accessible data encourages researchers to face and solve the challenges presented in this manuscript and beyond. As part of this work, datasets from both human and mouse species were analysed. In addition, human datasets from different tissues were accessed including lung, pancreas and kidney as well from different blood cells. Table 1 summarises the data analysed, and provides information about the tissue, species, technology used for generating the data, the number of available individuals, and provides references for additional information for each dataset.

Single-cell (multi-omics) data is heterogeneous and high-dimensional with missing values (often called drop-outs in scRNA-seq data) (Lähnemann et al., 2020). Batch correction in single-cell data poses a particular challenge that many researchers have tackled. In addition to these characteristics, single-cell data are also described as sparse with numerous low-quality samples (cells). Thus it is important to clean the data before analysing them. This section ends with a description of pre-processing steps taken for cleaning the datasets and obtaining high-quality datasets for analysis.

Mouse single cells from lung¹

This dataset consists of three mouse individuals with data taken from lung tissue. The data are taken from the Mouse Cell Atlas (Han et al., 2018). Separate scRNA-seq counts matrix are available, one for each individual, denoted by M_i^l , $\forall i = 1, 2, 3$; we refer to it as *mouse-lung* data. The accession number for the raw data files GSE108097.

Human single cells from lung²

This dataset consists of two human individuals with data taken from lung tissue. The data are taken from the Human Cell Landscape (Han et al., 2020). The scRNA-seq counts matrices of this dataset are denoted by H_i^l , $\forall i = 1, 2$; we refer to this dataset as *human-lung* data. The accession number for the raw data files GSE134355.

Human single cells from pancreas³

This dataset consists of eight human individuals with data taken from pancreas tissue sequenced through different technologies (Lab, 2019). The scRNA-seq counts matrices of this dataset are denoted by H_i^p , $\forall i = 1, 2$ and we refer to it as *human-pancreas* data. Butler et al. (2018) used this dataset to introduce the data integration tool Seurat v2. The accession number for the raw data files are GSE81076, GSE85241 (for technology CelSeq), GSE86469 (Fluidigm C1), E-MTAB-5061 (SMART-Seq2), and GSE84133 (inDrop).

A549⁴

This dataset consists of a single mouse individual with data taken from kidney tissue (Cao et al., 2018). Each cell received treatment for 1, 2 or 3 h. Thus the annotation of this dataset is not on the subtypes of the cells, but on the treatment time. It consists of scRNA-seq and scATAC-seq data on the same cells and it is denoted by M^{A549} ; we refer to this dataset as *mouse-A549* data. The accession number for the raw data files GSE117089.

PBMC - 10xGenomics⁵

This dataset consists of paired Peripheral Blood Mononuclear Cell PBMC profiles from a single human individual, generated by the 10x Genomics Multiome ATAC + RNA kit (Hao et al., 2021); it is denoted

by H^{PBMC} and we refer to it as *human-PBMC* data. This dataset was accessed through the R package SeuratData.

CBMC - Seurat³

This dataset consists of paired Cord Blood Mononuclear Cells (CBMC) profiles from a single human individual, generated by CITE-seq (Stoeckius et al., 2017). It consists of scRNA-seq data and surface proteins measured by Antibody-Derived Tags (ADTs) on the same cells and it is denoted by H^{CBMC} ; we refer to this dataset as *human-CBMC* data.

4. Cell subtype estimation using single-cell data

As discussed in the Introduction, one of the major challenges of the statistical analysis of scRNA-seq data is the mapping of cells to cell subtypes. In this section we showcase that cell subtypes in an individual organism can be estimated using another data from another individual from the species or from a different one as a reference dataset.

Through the analysis of single-cell RNA-seq data we aimed to assess whether the annotations are transferable in four practical situations:

- Training and test sets are taken from *different individuals of the same species*.
- Training and test sets are taken from *different individuals of different species*.
- Training and test sets are taken from *different sequencing technologies*.
- Training and test sets are taken from the *same individual*. A small percentage (5%, 10%, 20%) of the cells was considered annotated and taken as training.

In addition to the above, we explored the impact of *batch correction methods* on cell type annotation. In the experiments conducted, the semi-supervised approach S-multi-SNE approach was applied alongside other standard classification machine learning approaches and cell type classification approaches designed for single-cell data (described in Section 2). The training of the methods was performed on existing data that contain the true subtypes of the cells. The training and test sets were taken from separate studies, and we have used the sets of genes that are expressed in all cells in the studies. In the case of studies of different species, such as mouse and human, we limit ourselves to the set of orthologous genes.

The notation $M_i^l \rightarrow M_j^l$ corresponds to the implementation of the methods with M_i^l and M_j^l being the training and test sets, respectively. The same notation is used in all experiments conducted and presented in the manuscript. The subscripts correspond to the individuals available from each species as described in Section 3. The name of the study, as expressed in the first column of Table 1 is used to refer to the dataset considered for each analysis. For multi-omics analyses, explicit mention to the assay is made when necessary. For example, H_1^{PBMC} refers to the 1st individual in the H^{PBMC} study.

4.1. Different individuals of the same species

The first case investigated situations where the training and test sets are from individuals of the *same* species and with the same sequencing technology used. For this analysis, the mouse-lung (M^l) and human-lung (H^l) datasets were used (see Table 1 for further information for the datasets). In the mouse-lung dataset there are in total 31 subtypes of cells, whereas in the human-lung dataset there are in total 16 cell subtypes.

S-multi-SNE classified the cells of the test set with F_1 score at least 0.85 and 0.90 on human and mouse, respectively (Table 2). Visually, by incorporating the labelling information as an additional data-view in S-multi-SNE, the training data points of the same subtype are enforced to be clustered together (Fig. 2). The data points of the test set are found to be close to their corresponding cell subtype clusters.

Since S-multi-SNE projects the test single-cell samples in the same set of axes as the training cells data points with known subtypes are

¹ <https://figshare.com/s/865e694ad06d5857db4b>

² https://figshare.com/articles/dataset/HCL_DGE_Data/7235471?file=23043329

³ <https://github.com/satijalab/seurat-data>

⁴ <https://github.com/sqjin/scAI/tree/master/data>

⁵ <https://atlas.fredhutch.org/nygc/multimodal-pbmc/>

Table 1

Single-cell datasets description. Notation and characteristics of the datasets explored in this paper.

	Study	Tissue	Species	Technologies	Individuals	Assays	Reference	No. Modalities
	M^l	Lung	Mouse	Microwell-seq	3	scRNA-seq	(Han et al., 2018)	1
	H^l	Lung	Human	Microwell-seq	2	scRNA-seq	(Han et al., 2020)	1
	H^p	Pancreas	Human	Smart-seq	1	scRNA-seq	(Lab, 2019)	1
Multi-omics	M^{A549}	Kidney	Mouse	Cel-seq	2	scRNA-seq	(Cao et al., 2018)	2
				Indrop	4			
				Fluidigm	1			
				sci-CAR	1			
	H^{PBMC}	PBMC	Human	10X Genomics	1	scATAC-seq scRNA-seq	(Hao et al., 2021)	2
	H^{CBMC}	CBMC	Human	ATAC + RNA CITE-seq	1	scATAC-seq scRNA-seq ADT	(Stoeckius et al., 2017)	2

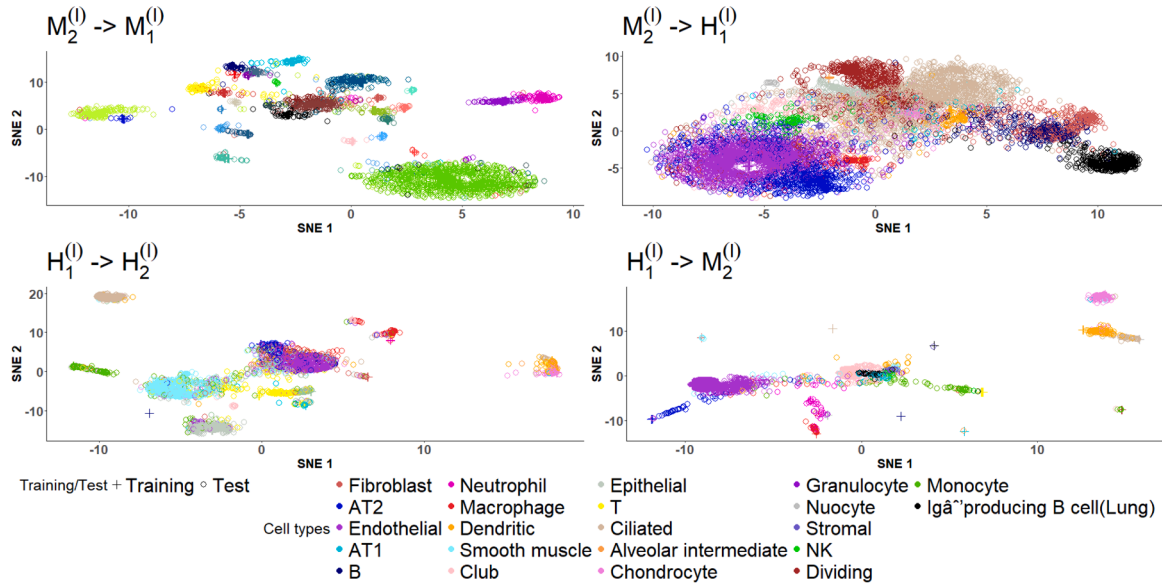


Fig. 2. Data visualisation of single-cell data, via S-multi-SNE. The training and testing individuals originate from the same or different species. The visualisations of the following four models are presented in this figure: $M_2^l \rightarrow M_1^l$, $M_2^l \rightarrow H_1^l$, $H_1^l \rightarrow H_2^l$, and $H_1^l \rightarrow M_2^l$, where $X \rightarrow Y$ denotes the model with X and Y representing the training and test datasets, respectively.

visualised along with points with unknown labels on the same set of axes (Fig. 2). Consequently, the subtypes of the test cells can be predicted by implementing kNN, on the obtained low-dimensional embeddings (Fig. 3).

The performance of S-multi-SNE was slightly improved when two individuals were used as training and one as testing, especially in predicting the subtypes of sample M_3^l (Table 2). A similar observation was observed when two individuals were taken as testing and one as training. Since, S-multi-SNE was implemented once to predict the cell subtypes of multiple individuals, this procedure was overall faster than predicting the cell subtypes from individuals one by one.

This two-step classification process has shown promising results. Cells of the same subtype are visualised near each other, regardless if they belonged in the training set (with known labels) or in the test set (with missing labels).

4.2. Different species

In the first set of experiments, the subtypes of the test samples were estimated by training S-multi-SNE on a separate individual taken from the same species and tissue. This section addresses whether the annotations are transferable between species. The question of interest is: *Can we estimate the subtypes of cells from an individual taken from a different*

Table 2

S-multi-SNE performance across different individuals of the same species. Accuracy, precision, recall and F_1 score of S-multi-SNE applied on various combinations of mouse-lung (M^l) and human-lung (H^l) datasets.

S-multi-SNE across samples					
Training	Test	Accuracy	Precision	Recall	F_1 score
M_1^l	M_2^l	0.88	0.97	0.90	0.94
M_1^l	M_3^l	0.84	0.96	0.87	0.91
M_2^l	M_1^l	0.91	0.95	0.96	0.95
M_2^l	M_3^l	0.87	0.95	0.91	0.93
M_3^l	M_1^l	0.88	0.95	0.93	0.94
M_3^l	M_2^l	0.89	0.96	0.92	0.94
M_1^l & M_2^l	M_3^l	0.90	0.97	0.92	0.94
M_1^l & M_3^l	M_2^l	0.89	0.97	0.92	0.94
M_2^l & M_3^l	M_1^l	0.92	0.96	0.96	0.96
M_1^l	M_2^l & M_3^l	0.89	0.97	0.92	0.95
M_2^l	M_1^l & M_3^l	0.89	0.95	0.93	0.94
M_3^l	M_1^l & M_2^l	0.88	0.96	0.91	0.94
H_1^l	H_2^l	0.78	0.86	0.89	0.87
H_2^l	H_1^l	0.76	0.86	0.86	0.86

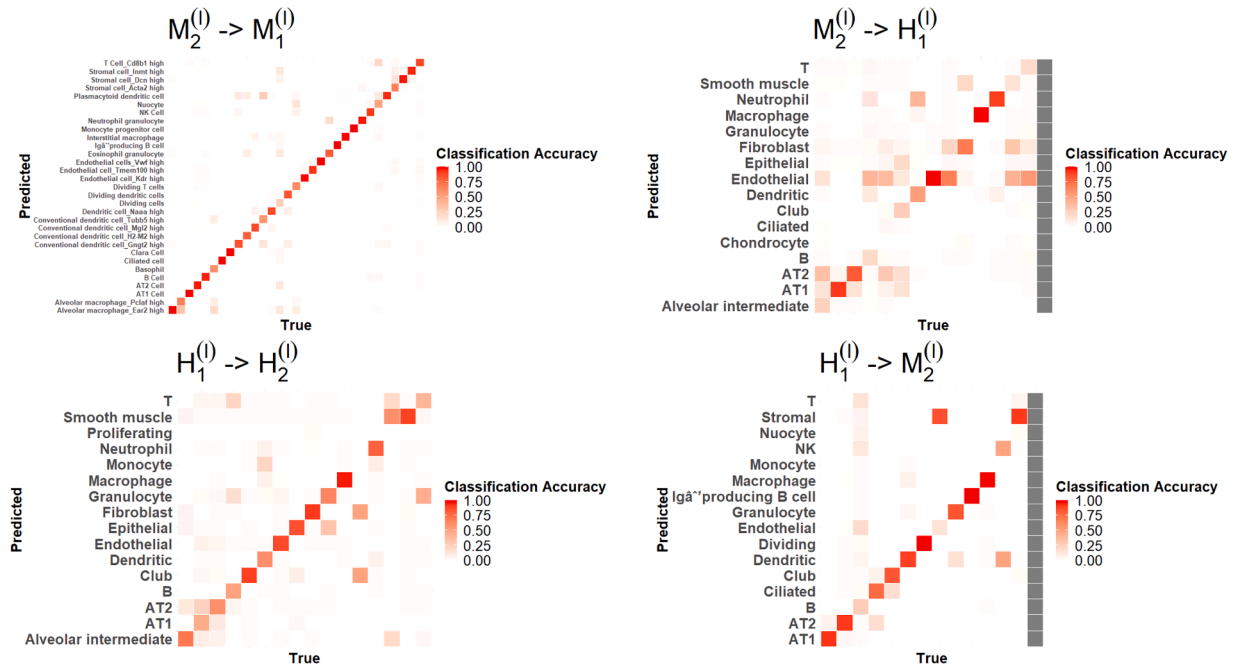


Fig. 3. Heatmaps of cell subtype estimation, via S-multi-SNE. The training and testing individuals originate from the same or different species. From left to right, the columns in each heatmap, represent the same cell subtypes, in the same order as the rows, starting from the bottom to the top, with the only exception being $H_1^l \rightarrow M_2^l$, for which some of the celltypes are not represented cross-species. The grey columns represent the cell subtypes missing from the training data and therefore cannot be predicted in the test individuals. The classification accuracy of the following four models is presented in this figure: $M_2^l \rightarrow M_1^l$, $M_2^l \rightarrow H_1^l$, $H_1^l \rightarrow H_2^l$, and $H_1^l \rightarrow M_2^l$, where $X \rightarrow Y$ denotes the model with X and Y representing the training and test datasets, respectively.

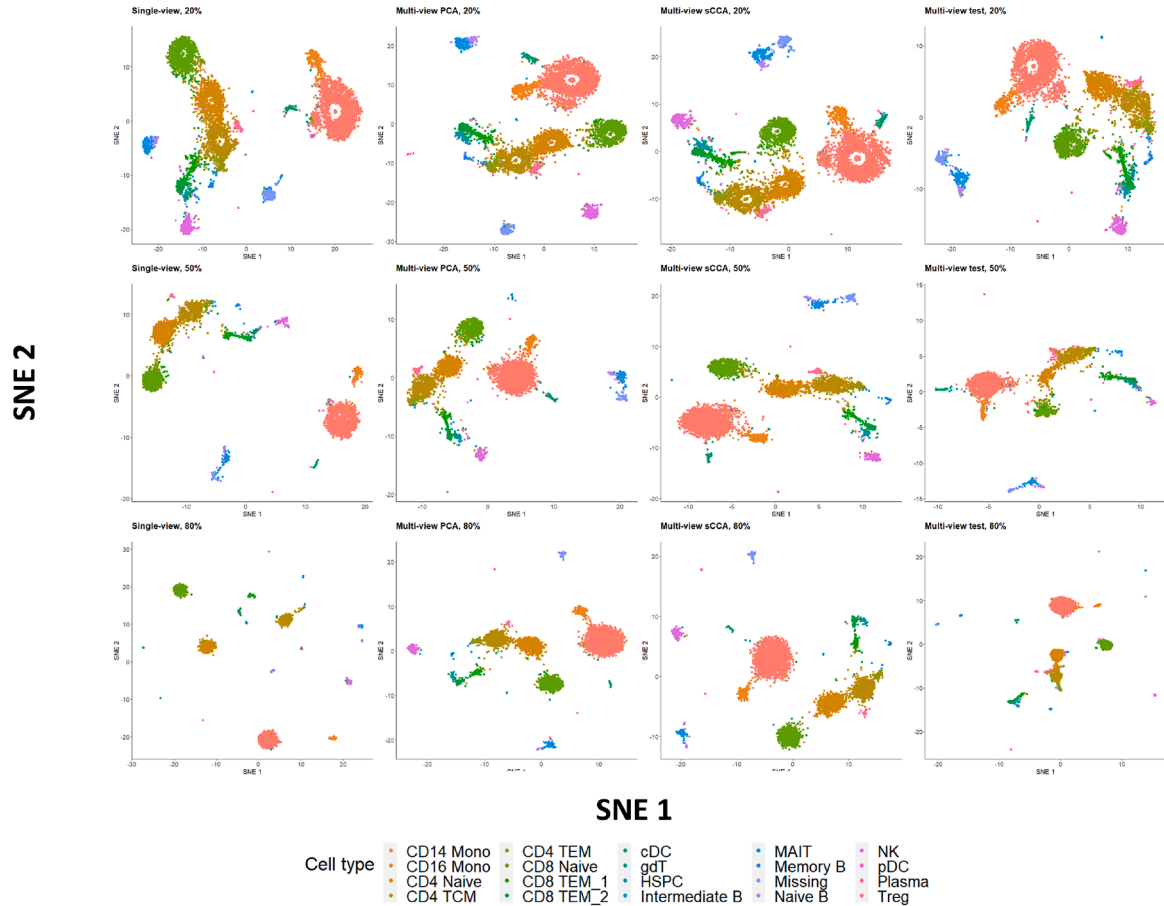


Fig. 4. Visualisations on human-pBMC. S-multi-SNE visualisations on the four scenarios as highlighted in Table 8 with different sample sizes for training and testing sets.

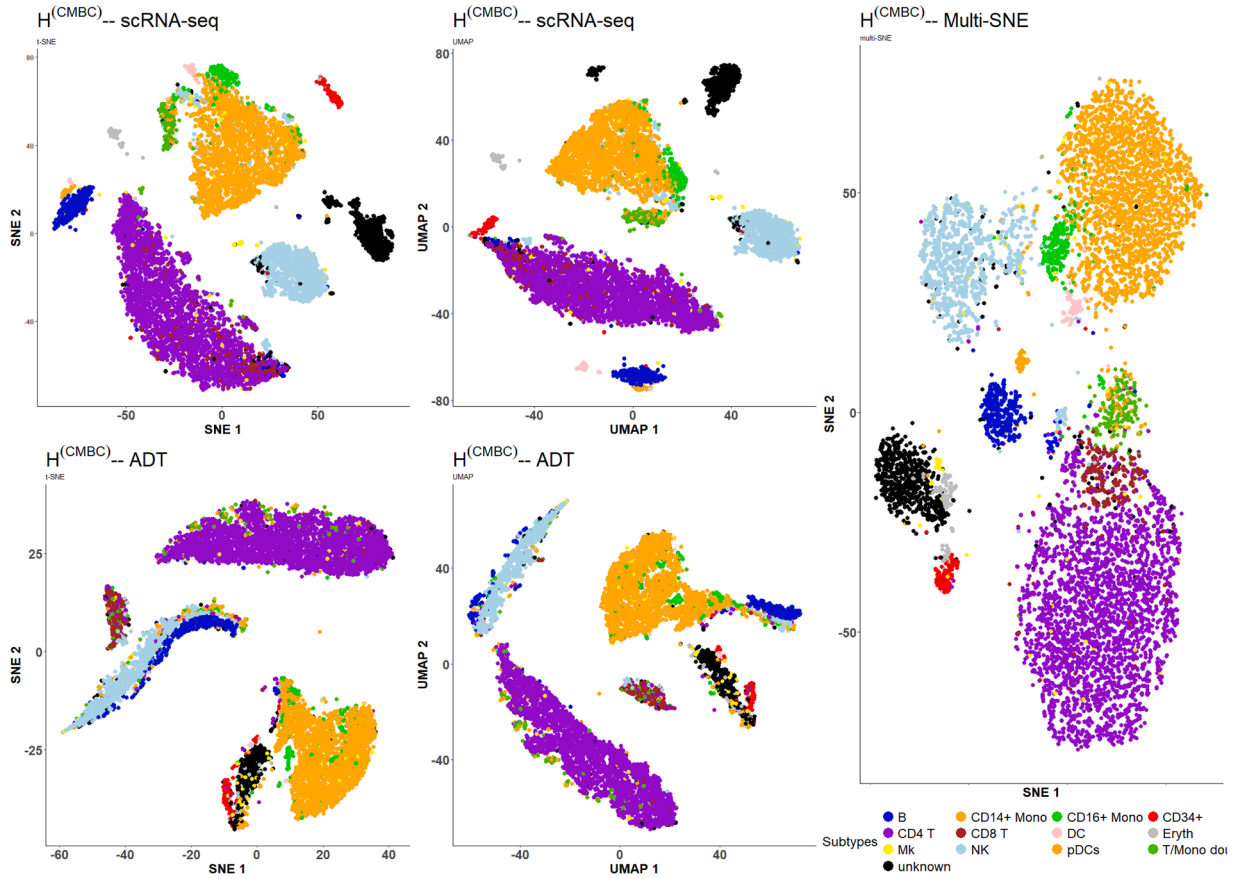


Fig. 5. Visualisations on human-CBMC. scRNA-seq, Antibody-Derived Tags (ADT) on both t-SNE and UMAP projects are complemented by a multi-view projection through multi-SNE on the $H^{(CBMC)}$ dataset. To facilitate quantitative comparison, Normalised Mutual Information (NMI) values are reported for each projection: UMAP on ADT 0.488, UMAP on RNA 0.5, t-SNE on ADT 0.456, t-SNE on RNA 0.543, multi-SNE 0.651.

species than the training species? In particular we addressed the question whether training S-multi-SNE on human-lung data, can be used to estimate the subtypes of mouse-lung, and vice versa?

Additional challenges to the ones discussed previously arise in this case. Firstly, there is an increased heterogeneity, since the cells are not only taken from separate individuals, but they are also sequenced on individuals from different species. Secondly, not all cell subtypes exist in the individuals of both species. It is impossible to classify a sample (cell) to its corresponding class, if that class is not represented in the training data. A mapping between the subtypes of the different species was used in order to align them in an agreed manner.

In this experiment, we implemented the following models: (a) $M_i^l \rightarrow H_j^l$ and (b) $H_j^l \rightarrow M_i^l$, where $i, j = 1, 2$. S-multi-SNE performed better in classifying the cells of mouse individuals, than human individuals (Table 3, column: Original). The main reason for this observation is due to the number of cells that belong to a subtype unseen by the training set; it is larger in human than in mouse samples. This is a consequence related to the second challenge discussed above.

Fig. 3 depicts the classification accuracy of S-multi-SNE between individuals of the same, or different species. Although most of the cell subtypes are shared between individuals, subtypes that are absent from the training set-as noted earlier-cannot be accurately predicted in the test set.

To mitigate the challenge of having new and untrained subtypes in the test set, two solutions were considered:

- randomly select a percentage (g%) of the test samples to be part of the training set
- select a percentage (g%) of the test samples that belong to the untrained subtypes to be part of the training set

Table 3

Predicting individuals across different species. The F_1 score of S-multi-SNE applied on data taken from the same tissue (lung), but from different species are presented, mouse-lung and human-lung. The subscripts of the two datasets present the individuals used as part of the training and test sets. Partial annotation of 5%, 10%, and 20% of the test samples was imposed. The last column, named Original, presents the datasets in their original form without taking any percentages as described in solutions (i) and (ii).

S-multi-SNE across species								
	Training	Test	Random test cells			New test subtypes		
			5%	10%	20%	5%	10%	20%
M_1^l	H_1^l	H_1^l	0.77	0.77	0.79	0.73	0.74	0.75
M_1^l	H_2^l	H_2^l	0.77	0.78	0.78	0.70	0.71	0.76
M_1^l	H_1^l	H_1^l	0.79	0.79	0.82	0.77	0.77	0.77
M_2^l	H_2^l	H_2^l	0.80	0.82	0.82	0.75	0.75	0.80
H_1^l	M_1^l	M_1^l	0.90	0.90	0.90	0.84	0.83	0.83
H_1^l	M_2^l	M_2^l	0.93	0.95	0.95	0.80	0.80	0.87
H_2^l	M_1^l	M_1^l	0.91	0.93	0.95	0.83	0.83	0.84
H_2^l	M_2^l	M_2^l	0.90	0.90	0.90	0.80	0.81	0.82

These two solutions were chosen because of their different applicability suitability. Solution (ii) ensures that the training set contains samples from all subtypes of the test set. Since we would not know the true subtype of the cells before annotation, this solution is not practically applicable in a real-life scenario. However, taking a random subset from the test set (as in solution (i)) and manually annotating it, can be practically applied.

An increase in the F_1 score of the methods was observed when either solution was implemented, even when adding a 5% of the test samples to

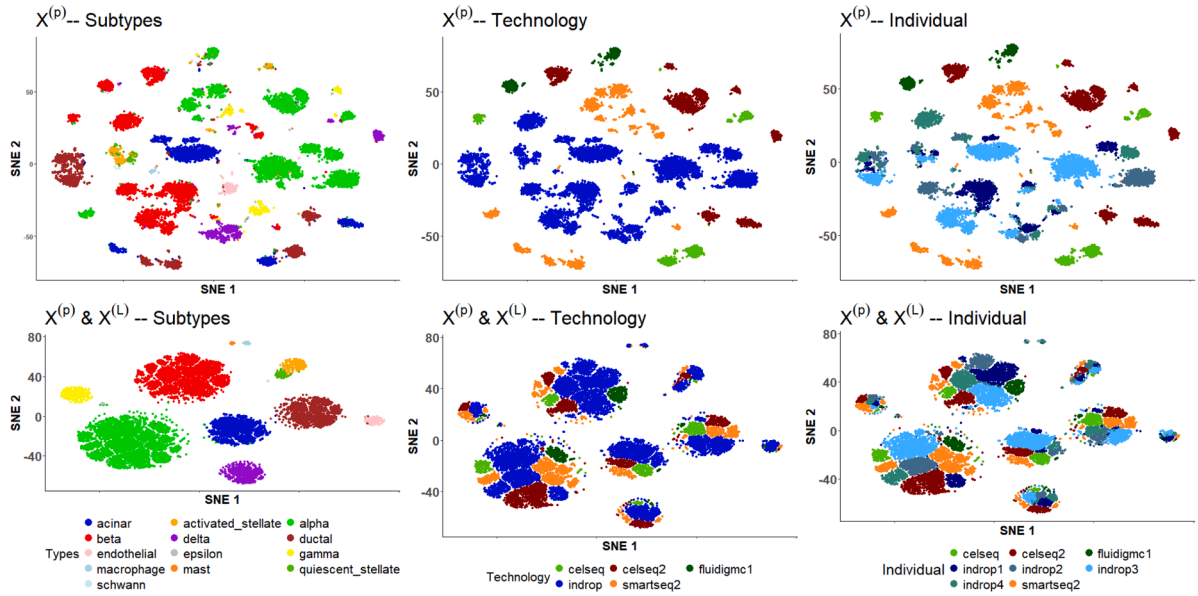


Fig. 6. Visualisations of human-pancreas data. t-SNE and multi-SNE projections with complete labelling information on $H^{(p)}$.

Table 4

Predicting individuals across different species using several supervised methods. The F_1 score of CHETAH, scPred, SingleCellNet, O2PLS and S-multi-SNE applied on data taken from the same tissue (lung), but the individuals come from different species. The S-multi-SNE column corresponds to the Original column of Table 3.

Cell subtype classification performance						
Training	Test	CHETAH	scPred	SingleCellNet	O2PLS	S-multi-SNE
M_1^I	H_1^I	0.51	0.58	0.42	0.72	0.75
M_1^I	H_2^I	0.60	0.50	0.40	0.55	0.76
M_2^I	H_1^I	0.50	0.51	0.42	0.80	0.79
M_2^I	H_2^I	0.55	0.51	0.38	0.71	0.80
H_1^I	M_1^I	0.77	0.87	0.79	0.79	0.83
H_1^I	M_2^I	0.79	0.87	0.76	0.80	0.87
H_2^I	M_1^I	0.73	0.87	0.77	0.82	0.83
H_2^I	M_2^I	0.74	0.83	0.74	0.72	0.83

the training set (Table 3). In particular, solution (i) produced overall better results than solution (ii). In $H^I \rightarrow M^I$ models, the F_1 score improved from 0.83 – 0.87 to 0.90 – 0.95, while in $M^I \rightarrow H^I$ models the increase in performance was less notable (from 0.75 – 0.80 to 0.77 – 0.82).

Standard machine learning approaches and recently proposed reference-based classifiers described in Supplementary material were implemented on the same data and compared with S-multi-SNE. S-multi-SNE outperformed all standard classification methods considered, where amongst them SVM, NN, and MultiReg performed the best (Table S1 in Supplementary material). The standard machine learning methods were applied to the original features of the data. An alternative strategy would have been to pre-process the features e.g. using principal components analysis (PCA). Such a strategy might have improved the performance of the methods but as the F_1 score of these methods was almost half of the S-multi-SNE's score, we believe that the additional pre-processing would have not contributed enough to recover the performance difference between the standard machine learning methods (SVM, MultiReg, RF, NN, kNN, LDA).

Overall, S-multi-SNE was found to be competitive with existing classifiers tailored for cell subtype estimation (Table 4). In particular, S-multi-SNE outperformed CHETAH, scPred and SingleCellNet in estimating human cells (training in mouse cells), while it performed equally well with scPred in estimating mouse cells (training in human cells). One of the approaches that performed comparatively well to S-multi-SNE is

Two-way Orthogonal Partial Least Squares (O2PLS) (Trygg and Wold, 2003), which is a generalisation of PLS that decomposes the variation of two input data-views into three parts, joint, orthogonal and noise. Due to the inner relations of the two data-views, it is possible to predict the values in one of the matrices, given the other. Similarly to S-multi-SNE, O2PLS incorporates the labelling information in its analysis as one of the data-views. Although the performances of the two algorithms were close, S-multi-SNE outperformed O2PLS in all runs, except $M_2^I \rightarrow H_1^I$, while O2PLS was superior to other machine learning classifiers.

4.3. Different sequencing technologies

S-multi-SNE alongside other methods have shown promising results in estimating the subtypes of cells taken from separate individuals, even in cases where the samples are from separate species (Section 4.2). In this section, we explore situations where individuals belong to the same species, but different technologies were used to sequence the cells of each individual. The human-pancreas (H^p) dataset was used for this experiment (see Table 1 for further details for the dataset).

For simplicity, all models explored in this section, $H_j^p \rightarrow H_i^p, i, j = \{1, \dots, 8\}$, contained a single individual as training and a single, separate individual for testing.

Regardless of whether the training and test individuals were sequenced with the same technology or not, S-multi-SNE classified the subtypes of cells with an average F_1 score of 0.95 (Table 5). No distinction was observed between the performance of the models and the sequencing technologies between the training and test individuals (Table 5).

Since all models performed well, a consensus classification assessment was measured. With H_j^p as the test individual, the classifications from all training individuals, $H_{i \neq j}^p$, were taken into consideration. The consensus solution is trivial when all models are in agreement. When the models are in disagreement though, taking the consensus reflects the selecting the class estimated by most models (with most votes). Considering the consensus of the models resulted in an increased average F_1 score of 0.97 (Table 5).

4.4. Same individuals

Suppose that cells from a single individual are sequenced and a small percentage of them is annotated. The research question addressed in

Table 5

S-multi-SNE performance across different individuals of the same species, sequenced by different technology. F_1 score of S-multi-SNE applied on the data taken from the human-pancreas (H^p) dataset.

S-multi-SNE across technologies								
Training	Test SMART-Seq2	CelSeq 1	CelSeq 2	Fluidigm C1	inDrop 1	inDrop 2	inDrop 3	inDrop 4
SMART-Seq2	–	0.97	0.98	0.98	0.94	0.96	0.98	0.97
CelSeq 1	0.96	–	0.93	0.91	0.90	0.96	0.95	0.96
CelSeq 2	0.97	0.97	–	0.97	0.95	0.96	0.97	0.98
Fluidigm C1	0.97	0.96	0.96	–	0.88	0.95	0.95	0.96
inDrop 1	0.87	0.95	0.92	0.96	–	0.92	0.95	0.96
inDrop 2	0.97	0.97	0.95	0.97	0.96	–	0.96	0.98
inDrop 3	0.97	0.95	0.98	0.97	0.96	0.96	–	0.96
inDrop 4	0.95	0.97	0.95	0.97	0.95	0.96	0.96	–
Consensus	0.98	0.98	0.98	0.98	0.95	0.97	0.98	0.98

Table 6

F_1 score of S-multi-SNE applied on the same individual with training on 5%, 10%, 20% samples. The mean (and standard deviation) F_1 score obtained by applying S-multi-SNE, CHETAH, and SingleCellNet on data taken from the same tissue (lung) of both mouse and human individuals. 5%, 10% and 20% of the samples were randomly chosen as training with the remaining acting as test sets.

Cell subtype estimation, F_1 score						
Training %	Algorithm	M'_1	M'_2	M'_3	H'_1	H'_2
5%	S-multi-SNE	0.72 (0.03)	0.59 (0.01)	0.63 (0.03)	0.38 (0.01)	0.26 (0.03)
	CHETAH	0.72 (0.05)	0.58 (0.04)	0.61 (0.04)	0.17 (0.04)	0.17 (0.05)
	SingleCellNet	0.81 (0.03)	0.52 (0.02)	0.80 (0.02)	0.39 (0.03)	0.22 (0.04)
10%	S-multi-SNE	0.85 (0.03)	0.64 (0.01)	0.65 (0.02)	0.42 (0.01)	0.34 (0.06)
	CHETAH	0.77 (0.04)	0.63 (0.03)	0.68 (0.03)	0.17 (0.03)	0.24 (0.06)
	SingleCellNet	0.84 (0.03)	0.75 (0.02)	0.85 (0.03)	0.39 (0.05)	0.26 (0.05)
20%	S-multi-SNE	0.87 (0.07)	0.74 (0.01)	0.73 (0.02)	0.49 (0.01)	0.41 (0.05)
	CHETAH	0.80 (0.04)	0.70 (0.04)	0.72 (0.05)	0.23 (0.05)	0.32 (0.06)
	SingleCellNet	0.89 (0.04)	0.82 (0.02)	0.89 (0.06)	0.39 (0.04)	0.35 (0.06)

this section is: *Can we estimate the subtypes of the remaining cells?*. For creating as realistic experiments as possible, 5%, 10% and 20% of the cells were randomly chosen as training and the remaining acted as testing. For these experiments, the human-lung (H^l) and mouse-lung (M^l) datasets were used. The random training/testing splits were repeated 100 times.

In this experiment, S-multi-SNE, CHETAH and SingleCellNet were applied to evaluate and compare their performance on estimating cell subtypes using subsets of the cell populations of individuals. It is worth noting that some cell types are represented by very few cells in the dataset (less than 5), which limits the statistical power of the classification for those populations and prevents the use of methods such as scPred that impose a minimum cell count per subtype.

On M'_i for $i = 1, 2, 3$, S-multi-SNE performed equally or better than CHETAH (Table 6). With two exceptions, SingleCellNet was the superior classification algorithm on mouse individuals. On H'_1 , as the number of training samples increases, S-multi-SNE obtains a steeper improvement in its performance than CHETAH and SingleCellNet (Table 6). On both human individuals, S-multi-SNE outperformed the other two classification algorithms.

In this experiment, S-multi-SNE had strong performance and in particular on human individuals. For all classification algorithms, the F_1 score improves as the percentage of training cells increases.

4.5. Batch correction

As the experiments in the previous sections showed that S-multi-SNE overall is able to produce accurate cell subtype annotations. Through a set of experiments, we investigated whether the choice of batch correction algorithm has an effect on the performance of S-multi-SNE.

In all experiments so far, batch correction was performed by *Seurat v4*. Here, other standard batch correction algorithms that have been proposed in the literature for single-cell data are explored. In particular, Seurat v3, Seurat v4, LIGER, faster Mutual Nearest Neighbours (MNN), classic MNN and cluster-based MNN were implemented and compared, as described in Table S2 in Supplementary material. A naive integration with no batch correction was used as a benchmark for this comparison, i.e., a simple concatenation of the cells from both individuals was applied.

As in the earlier experiments, the same four steps (described in Section 2.2) were implemented in this section, too. The difference lies in step 2 - Data integration and batch correction via Seurat v4. Seurat v4 is replaced with the batch correction algorithm presented in Table Table S2 in Supplementary material. To quantitatively evaluate and compare these algorithms, kNN was then implemented and the F_1 score was measured.

Between individuals of the same species, all batch correction algorithms performed well with little difference in their F_1 scores (Table 7). Cluster-based MNN was the best performing solution, with an F_1 score at least 0.95 in all combinations. Correction via LIGER produced the lowest classification results, but they were still much better than without any batch correction.

On the other hand, most batch correction algorithms performed poorly in integrating data from two distinct species, mouse and human. In particular all algorithms, except Seurat v4, perform similarly to a naive integration with no batch correction (Table 7). Thus, they appear to fail in correcting the batches. Only Seurat v4 was able to correct the batches taken from individuals of different species and consequently produce strong classification results. These results support our choice of implementing the Seurat v4 as the default batch correction approach for the adapted version of S-multi-SNE for cell subtype classification.

Table 7

Exploring the effect the batch correction approaches. The F_1 score of S-multi-SNE applied on single-cell data without batch correction and after correcting the batches via Seurat v4, LIGER, faster MNN, classic MNN and cluster-based MNN is reported. The best performed batch correction algorithm for each training/test combination is presented in **bold**.

Batch correction comparison						
Training	Test	No correction	Seurat v4	LIGER	faster MNN	Seurat v3
						Cluster-based MNN
M_1^1	M_2^1	0.76	0.94	0.87	0.96	0.91
M_1^1	M_3^1	0.75	0.91	0.86	0.93	0.89
M_2^1	M_1^1	0.84	0.95	0.92	0.95	0.91
M_2^1	M_3^1	0.78	0.93	0.86	0.92	0.87
M_3^1	M_1^1	0.83	0.94	0.89	0.93	0.91
M_3^1	M_2^1	0.77	0.94	0.84	0.94	0.90
M_1^1	H_1^1	0.48	0.75	0.49	0.54	0.52
M_1^1	H_2^1	0.50	0.76	0.41	0.47	0.45
M_2^1	H_1^1	0.47	0.79	0.46	0.55	0.50
M_2^1	H_2^1	0.43	0.80	0.41	0.48	0.45
H_1^1	M_1^1	0.44	0.83	0.57	0.54	0.59
H_1^1	M_2^1	0.44	0.87	0.54	0.43	0.52
H_2^1	M_1^1	0.41	0.83	0.62	0.65	0.65
H_2^1	M_2^1	0.53	0.83	0.54	0.52	0.53

5. Single-cell multi-omics

Although the majority of available single-cell data contain a single data-view (e.g. scRNA-seq), multi-omics data are taking over with more sequencing technologies producing multiple data-views on the same cells.

Using the publicly available multi-omics single-cell data (Section 3), we investigated the use of single-cell multi-omics data for cell identification. Both multi-SNE and S-multi-SNE are applied to datasets that contains both scRNA-seq and scATAC-seq on the same set of single cells. The two methods were explored for addressing the key question whether multi-omics datasets can be beneficial in cell subtype identification and visualisation.

5.1. Cell subtype classification using single-omics and multi-omics data

In this section, we investigate whether incorporating both scRNA-seq and scATAC-seq in S-multi-SNE improves the classification performance. In particular, if the test samples are multi-view, scRNA-seq and scATAC-seq, and the training samples are single-view, scRNA-seq, would the inclusion of additional data-views, scATAC-seq, in the test set improve the performance of S-multi-SNE? The purpose of this analysis is to explore whether cell subtype estimation on single-cell multi-omics data can be achieved by training the model on single-view data. The human-PBMC dataset was used in this experiment. To evaluate S-multi-SNE, the data points were split to training and test sets with 20%, 50% and 80% training rates.

Four scenarios were tested and compared in this experiment as described in Table 8. The first scenario only considers scRNA-seq data and it acts as a benchmark. The second and third scenarios perform S-multi-SNE on both scRNA-seq and scATAC-seq, but with different pre-training (PCA and sCCA (Rodosthenous et al., 2020)) approaches. PCA is commonly used as a pre-training algorithm in similar tasks, as it can reduce the dimensions and reduce the sparsity effect of the data. Sparse Canonical Correlation Analysis (sCCA) solutions are extensions of CCA applicable in high-dimensional data ($p \gg n$). CCA can be considered as an extension of (PCA) applied on two datasets rather than one dataset, with a goal to produce sparse latent variables that achieve maximal correlation between the data-views, instead of achieving maximal variance as in PCA.

The fourth scenario expresses a situation where the training set does not contain scATAC-seq, but the test set does. The motivation behind

the last scenario is based on the fact that single-cell multi-omics data are frequently being sequenced, but at the same time the literature lacks of well-annotated multi-omics data that can be used as training.

The bar is set high as the benchmark model (scenario 1) performs well with F_1 score of 0.93 even when the model was trained on only 20% of the single cells. Even so, multi-view approaches performed better than the single-view benchmark, improving the F_1 score to 0.96 on 20% training rate (Table 9). On all three training rates, both models had an identical classification performance regardless of the pre-training algorithm. In the fourth scenario, and especially with training rates of 50% and 80%, the performance of the model was superior to the benchmark model, but slightly inferior to the other models. The inclusion of scATAC-seq in the test set appears to influence positively the S-multi-SNE output.

Qualitatively, the same conclusions are obtained; the single cells that belong in small clusters/subtypes are separated more clearly in the second and third scenarios, while the fourth scenario projects the training and test cells closer together (Fig. 4).

Further exploration of matrix factorisation-based integration methods, such as MOFA (Argelaguet et al., 2018), as alternative pre-training steps for S-multi-SNE represents a promising direction for future work. The benchmarking study by Cantini et al. (2021) showcases that MOFA and CCA-based approaches are among the better-performing methods for multi-omics integration.

5.2. Cell subtype visualisation using single-omics and multi-omics data

We further explored whether multi-omics data can improve the visualisations obtained from single-omics data. To achieve this, we applied multi-SNE on the single-cell multi-omics data of the human-CBMC $H^{(CBMC)}$ dataset and compared the obtained projects qualitatively with the projections of t-SNE applied to the two omics datasets separately.

Both scRNA-seq and Antibody-Derived Tags (ADT) data-views produced good visualisations via t-SNE (Fig. 5). In the scRNA-seq projection, $CD4^+ T$ and $CD8^+ T$ cells are mixed together. In the ADT projection the $CD4^+ T$ and $CD8^+ T$ cells are well separated, but the $T/Mono$ doublets and $CD16^+ Mono$ cells are scattered within the $CD4^+ T$ and $CD8^+ T$ clusters. Concluding, the two t-SNE projections are not aligned and do not agree on the separation of smaller clusters.

On the other hand, multi-SNE distinguishes those three cell subtypes ($T/Mono$ doublets, $CD4^+ T$ and $CD8^+ T$) in clearly separate clusters (Fig. 5). This observation suggests that multi-SNE was able to capture important information from both data-views and provide a clearer separation between the subtypes of the cells. Similar observations can also be noticed between other subtypes. For example, the separation between $CD14^+ Mono$ and $CD16^+ Mono$ is clearer in the multi-SNE projection than either in the scRNA-seq, or ADT projections.

The effect of including complete (no missing data) label data-view on the embeddings produced by multi-SNE was also explored. Multi-SNE was used to visualise the human-pancreas ($H^{(P)}$) dataset.

The aim here is to solely project the cells separated by their corresponding subtype. Thus, the use of labelling information is permitted, in this experiment no classification or further use of the label data-view is required. In other words, the only goal of this experiment is to visualise the data by incorporating the complete label data-view in the algorithm. Two data-views were considered: (a) all cells (from every technology) integrated and batch corrected via Seurat v4, denoted by $X^{(P)}$ and (b) annotation (subtypes) of all cells, denoted by $X^{(L)}$.

Implementing t-SNE at the entire dataset (i.e. on $X^{(P)}$), separates the cells by both subtype, and individual. However, no concise structure is observed (Fig. 6). Incorporating the subtype information of each cell as an additional data-view, multi-SNE provides some structure to the overall visualisation. Multi-SNE ‘constrains’ the cells of the same subtype to be clustered together. The 13 cell subtypes are clearly separated to their respective clusters. Simultaneously, the 5 technologies and 8 individuals

Table 8
S-multi-SNE comparison scenarios.

	Scenario	Training	Test	Notes
1	Single-view test	scRNA-seq	scRNA-seq	Benchmark
2	Multi-view PCA	scRNA-seq scATAC-seq	scRNA-seq scATAC-seq	Pre-training via PCA Typical application of S-multi-SNE on two data-views
3	Multi-view sCCA	scRNA-seq scATAC-seq	scRNA-seq scATAC-seq	Pre-training via sCCA Different pre-training approach
4	Multi-view test	scRNA-seq	scRNA-seq scATAC-seq	Investigate the scenario for which the scATAC-seq data are missing from training

Table 9

Cell type classification using single-view and multi-view data of human-PBMC dataset. The average and standard deviation of accuracy, precision, recall and F_1 score of S-multi-SNE are reported. S-multi-SNE is applied on the H^{PBMC} dataset using different proportions of training data and on different scenarios of the data as reported in Table 8.

S-multi-SNE on single-view and multi-view data					
Training %	Scenarios	Accuracy	Precision	Recall	F_1 score
20%	Single-view	0.91 (0.004)	0.92 (0.003)	0.93 (0.005)	0.93 (0.003)
	Multi-view PCA	0.93 (0.003)	0.98 (0.003)	0.95 (0.004)	0.96 (0.002)
	Multi-view sCCA	0.93 (0.004)	0.98 (0.004)	0.95 (0.004)	0.96 (0.003)
	Multi-view test	0.93 (0.004)	0.98 (0.002)	0.95 (0.004)	0.96 (0.002)
50%	Single-view	0.93 (0.002)	0.94 (0.002)	0.95 (0.002)	0.94 (0.002)
	Multi-view PCA	0.95 (0.002)	0.98 (0.002)	0.97 (0.002)	0.98 (0.001)
	Multi-view sCCA	0.95 (0.003)	0.98 (0.005)	0.97 (0.002)	0.98 (0.003)
	Multi-view test	0.94 (0.003)	0.97 (0.003)	0.97 (0.003)	0.97 (0.002)
80%	Single-view	0.93 (0.006)	0.96 (0.005)	0.96 (0.004)	0.96 (0.003)
	Multi-view PCA	0.96 (0.004)	0.97 (0.003)	0.98 (0.003)	0.98 (0.002)
	Multi-view sCCA	0.96 (0.004)	0.98 (0.005)	0.98 (0.002)	0.98 (0.003)
	Multi-view test	0.94 (0.005)	0.98 (0.005)	0.97 (0.003)	0.97 (0.003)

can still be viewed separately, all within a ‘bigger’ cluster, representing the cell subtypes.

This observation suggests that multi-SNE can provide a more structured visualisation by incorporating the labelling information ($X^{(L)}$) of the samples (in this case cell subtypes). This process though should only be used for visualisation tasks; no further analysis on the embeddings would be sensible, since the labelling information of all samples was incorporated. On the other hand, suppose that a second response is also available (denoted by $X_2^{(L)}$). Then it would not be unreasonable to use these embeddings for further analysis in regards to investigating $X_2^{(L)}$; in fact, the use of $X^{(L)}$ might prove to be helpful. Such implementation could be potentially useful for hierarchical clustering, a potential future direction of exploration for both multi-SNE and S-multi-SNE.

6. Discussion

The analysis of high-dimensional and sparse single-cell data poses a lot of challenges and the rapid progress of single-cell protocols and sequencing technologies demands for the development of algorithms to address these challenges. In this manuscript, we explored a new approach for cell subtype annotation in single cell data. The implemented approach is an adaptation of the recently proposed multi-view manifold learning approach S-multi-SNE that is an extension of multi-SNE. It is important to note that, unlike standard t-SNE where classification can be performed post-hoc on an unsupervised non-linear embedding, S-multi-SNE incorporates label information directly into the dimensionality reduction objective, producing a semi-supervised latent space that is explicitly guided to preserve class-relevant structure.

Even though both multi-SNE and S-multi-SNE are general machine learning solutions for visualisation and semi-supervised classification, and are not specifically tailored for single-cell data, in the conducted work we illustrated that they can have very good performance for the

analysis of single-cell data. Using a range of datasets and scenarios we have illustrated that both multi-SNE and S-multi-SNE perform well on single-view and multi-view single-cell data, with or without the labelling information.

Towards addressing the challenges described in the Introduction, S-multi-SNE provides an automated and cheaper solution without the need on relying on a reference dataset with well-defined cellular subtypes such as a cell atlas that requires an extensive literature review to be created. Further, it does not require any adjustments, or re-evaluation, and it was demonstrated that it allows transferable annotations between individuals of the same species, different species, or individuals sequenced from different technologies. A general limitation of S-multi-SNE lies in its transductive properties where a new run is required for every test individual. However, this restriction is not necessarily a drawback for this particular task, since usually the estimation is performed on each individual separately. A general solution would require adjustments, such as parameter tuning. Similarly to all reference-based classification solutions, S-multi-SNE suffers from the limitation of only estimating cell subtypes that are covered by the reference (training) dataset.

In our work, we have focused on situations where some information is available and have demonstrated that when this minimal information is accessible, annotation accuracy can be improved dramatically. This finding aligns with recent comprehensive surveys indicating that supervised methods consistently outperform unsupervised approaches when reference data is available, achieving significantly higher accuracy rates across diverse experimental settings. Our semi-supervised S-multi-SNE framework provides a practical compromise that addresses several key limitations in current cell type annotation workflows. Specifically, the approach reduces annotation burden by requiring only 10% manually labelled cells while maintaining discovery capability for novel cell types through the unsupervised component, and enables knowledge transfer from well-annotated reference datasets to new experimental contexts. Moreover, our results demonstrate that the multi-view approach can effectively harness existing annotations from one species (such as mouse) to annotate cell types in another species (such as human), facilitating cross-species knowledge transfer that can be instrumental in scientific discovery.

Although the analysis of single-cell data is progressing fast, it is still at its early stages. In addition to the challenges of this new type of data, several complications and drawbacks arise, while estimating the cell subtypes of an unannotated dataset. For example, laboratories may use a different hierarchy in annotating the cells (e.g. T-cells being split into sub-categories such as CD4 and CD8). Moreover, laboratories use different approaches and software packages to obtain cell annotations. The publicly available annotations are heavily dependent on the laboratory producing the data and the researchers sequencing and clustering the cells. Additional complications emerge in cross-species cell subtype estimation. The missing cell subtypes from one of the species (query or reference) is the most conspicuous issue. It is impossible to predict a class of a categorical response without training a model on that class. As mentioned in Section 4, to overcome these complications, a mapping between the cell subtypes (both within same species and cross-species) was constructed. The performance of our models does depend enormously

on the mapping between cell subtypes and the batch correction method used to integrate the scRNA-seq data of separate individuals. The superior performance of S-multi-SNE in cross-species mapping compared to the other tools we have tested is worth emphasizing. This kind of annotation is useful, when new datasets on less studied species become available and the cellular sub-types can be learned from evolutionary close species.

Recent developments of single-cell technologies are quickly making different kind of single-cell multi-omics data a possibility and specific challenges arise in analysing these kinds of data (Baysoy et al., 2023; Heumos et al., 2023). We demonstrated here through several applications to multi-omics data that multi-SNE and S-multi-SNE are capable of visualizing and providing accurate cell sub-type annotations of such data as well. Through the implementation of two alternative pre-training steps for multi-SNE, we showed that pre-training with sCCA, instead of PCA that is the standard approach of t-SNE, is a promising alternative that can provide both improved visualisation and classification tasks. These findings suggest that the correlated low-dimensional canonical vectors appear to contain more important information than the low-dimensional principal components. Nowadays, specific methods for analysis of multi-omics data are fast being developed including methods based on deep generative models (Schuster et al., 2024; Cui et al., 2024; Hao et al., 2024). While, pure statistical and machine learning approaches will have their role to play, we believe to fully take into account technical noise and stochastic dynamical aspects of genetic networks underlying single-cell multi-omics data one requires combining mechanistic models and machine-learning approaches and this will be an important area of future developments (Gorin et al., 2023; Volteras et al., 2024).

In summary, S-multi-SNE and multi-SNE can be utilised to tackle two important (analytical) steps in single-cell data analysis, classification (e.g. cell subtype estimation) and visualisation. Multi-SNE offers a unified single visualisation based on multiple data-views available (Rodosthenous et al., 2024). It was shown that such projections offer a consensus view of the data that reflect their true structure clearer with fewer clustering bundles between the different cell subtypes.

CRediT authorship contribution statement

Theodoulos Rodosthenous: Formal analysis, Methodology, Software, Visualization, Writing - original draft; **Vahid Shahrezaei:** Conceptualization, Supervision, Writing - review & editing; **Marina Evangelou:** Conceptualization, Supervision, Writing - review & editing.

Declaration of competing interest

None

Acknowledgments

We would like to acknowledge discussions with Dr Nathan Skene and his PhD student Brian Schilder who specialise in computational biology. In particular, we benefited from consulting with them with regards to the mapping of cellular sub-types across species used in this paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.jtbi.2026.112442](https://doi.org/10.1016/j.jtbi.2026.112442)

References

Alquicira-Hernandez, J., Sathe, A., Ji, H.P., Nguyen, Q., Powell, J.E., 2019. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.* 20 (1), 1–17.
 Andrews, T.S., Kiselev, V.Y., McCarthy, D., Hemberg, M., 2021. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat. Protoc.* 16 (1), 1–9.

Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al., 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* 20 (2), 163–172.
 Arendt, D., Musser, J.M., Baker, C. V.H., Bergman, A., Cepko, C., Erwin, D.H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M.D., et al., 2016. The origin and evolution of cell types. *Nat. Rev. Genet.* 17 (12), 744–757.
 Argelaguet, R., Velten, B., Arnold, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O., 2018. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* 14 (6), MSB178124.
 Baysoy, A., Bai, Z., Satija, R., Fan, R., et al., 2023. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* 24 (10), 695–713. <https://doi.org/10.1038/s41580-023-00615-w>
 Bi, C., Bai, K., Li, X., Zhang, X., 2024. HiCat: a semi-supervised approach for cell type annotation. *arXiv:2412.06805*.
 Boufeka, K., Seth, S., Batada, N.N., 2020. scID uses discriminant analysis to identify transcriptionally equivalent cell types across single-cell RNA-seq data with batch effect. *IScience* 23 (3), 100914.
 Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R., 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36 (5), 411–420.
 Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., Baudot, A., 2021. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* 12 (1), 124.
 Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al., 2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 361 (6409), 1380–1385.
 Chen, J., Xu, H., Tao, W., Chen, Z., Zhao, Y., Han, J.-D.J., et al., 2023. Transformer for one stop interpretable cell type annotation. *Nat. Commun.* 14 (1), 223. <https://doi.org/10.1038/s41467-023-35923-4>
 Chen, L., He, Q., Zhai, Y., Deng, M., 2021. Single-cell RNA-seq data semi-supervised clustering and annotation via structural regularized domain adaptation. *Bioinformatics* 37 (6), 775–784.
 Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B., et al., 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* 21 (8), 1470–1480. <https://doi.org/10.1038/s41592-024-02201-0>
 Dominguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T., Howlett, S.K., Suchanek, O., Polanski, K., King, H.W., et al., 2022. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 376 (6594), eabl5197.
 Fix, E., Hodges, J., 1951. An important contribution to non-parametric discriminant analysis and density estimation. *Int. Stat. Rev.* 3 (57), 233–238.
 Gorin, G., Vastola, J.J., Pachter, L., et al., 2023. Studying stochastic systems biology of the cell with single-cell genomics data. *Cell Syst.* 14 (10), 822–843.e22. <https://linkinghub.elsevier.com/retrieve/pii/S2405471223002442>. <https://doi.org/10.1016/j.cels.2023.08.004>
 Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al., 2018. Mapping the mouse cell atlas by microwell-seq. *Cell* 172 (5), 1091–1107.
 Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al., 2020. Construction of a human cell landscape at single-cell level. *Nature* 581 (7808), 303–309.
 Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, III, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al., 2021. Integrated analysis of multi-modal single-cell data. *Cell* 182, 293–304.
 Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., Satija, R., et al., 2024. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* 42 (2), 293–304. <https://doi.org/10.1038/s41587-023-01767-y>
 Heumos, L., Schaar, A.C., Lance, C., Litnitskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., Single-cell Best Practices Consortium, Aliee, H., Ansari, M., Badia-i Mompel, P., Büttner, M., Dann, E., Dimitrov, D., Dony, L., Frishberg, A., He, D., Hediye-zadeh, S., Hetzel, L., Ibarra, I.L., Jones, M.G., Lotfollahi, M., Martens, L.D., Müller, C.L., Nitzan, M., Ostner, J., Palla, G., Patro, R., Piran, Z., Ramírez-Suástegui, C., Saez-Rodriguez, J., Sarkar, H., Schubert, B., Sikkema, L., Srivastava, A., Tanevski, J., Virshup, I., Weiler, P., Schiller, H.B., Theis, F.J., et al., 2023. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* 24 (8), 550–572. <https://doi.org/10.1038/s41576-023-00586-w>
 de Kanter, J.K., Lijnzaad, P., Candelli, T., Margaritis, T., Holstege, F. C.P., 2019. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.* 47 (16), e95.
 Kiselev, V.Y., Yiu, A., Hemberg, M., 2018. scmap: projection of single-cell RNA-seq data across data sets. *Nat. methods* 15 (5), 359–362.
 Lab, S., 2019. *panc8.SeuratData: Eight Pancreas Datasets Across Five Technologies*. R package version 3.0.2.
 Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al., 2020. Eleven grand challenges in single-cell data science. *Genome Biol.* 21 (1), 1–35.
 Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., Ren, X., Zhang, Z., 2020. SciBet as a portable and fast single cell type identifier. *Nat. Commun.* 11 (1), 1–8.
 Lieberman, Y., Rokach, L., Shay, T., 2018. CaSTLe-classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One* 13 (10), e0205499.
 Lin, Y., Cao, Y., Kim, H.J., Salim, A., Speed, T.P., Lin, D.M., Yang, P., Yang, J. Y.H., 2020. scClassify: sample size estimation and multi-scale classification of cells using single and multiple reference. *Mol. Syst. Biol.* 16 (6), e9389.

- van der Maaten, L., Hinton, G., 2008. Visualising data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.
- Pasquini, G., Rojo Arias, J.E., Schäfer, P., Busskamp, V., et al., 2021. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* 19, 961–969. <https://linkinghub.elsevier.com/retrieve/pii/S2001037021000192>. <https://doi.org/10.1016/j.csbj.2021.01.015>
- Pliner, H.A., Shendure, J., Trapnell, C., 2019. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* 16 (10), 983–986.
- Rodosthenous, T., Shahrezaei, V., Evangelou, M., 2020. Integrating multi-omics data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. *Bioinformatics* 36 (17), 4616–4625.
- Rodosthenous, T., Shahrezaei, V., Evangelou, M., 2021a. Semi-supervised classification and visualization of multi-view data. *JSM 2021 - Sect. Stat. Learn. Data Sci.*, 1066–1080.
- Rodosthenous, T., Shahrezaei, V., Evangelou, M., 2021b. S-multi-SNE: semi-supervised classification and visualisation of multi-view data. *arXiv:2111.03519*.
- Rodosthenous, T., Shahrezaei, V., Evangelou, M., et al., 2024. Multi-view data visualisation via manifold learning. *PeerJ Comput. Sci.* 10, e1993. <https://peerj.com/articles/cs-1993>. <https://doi.org/10.7717/peerj-cs.1993>
- Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A., 2015. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33 (5), 495–502.
- Sato, K., Tsuyuzaki, K., Shimizu, K., Nikaido, I., 2019. CellFishing: an ultrafast and scalable cell search method for single-cell RNA sequencing. *Genome Biol.* 20 (1), 1–23.
- Schuster, V., Dann, E., Krogh, A., Teichmann, S.A., 2024. multiDGD: a versatile deep generative model for multi-omics data. *Nat. Commun.* 15 (1), 10031. <https://www.nature.com/articles/s41467-024-53340-z>. <https://doi.org/10.1038/s41467-024-53340-z>
- Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., Smibert, P., 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14 (9), 865–868.
- Tan, Y., Cahan, P., 2019. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. *Cell Syst.* 9 (2), 207–213.
- Trygg, J., Wold, S., 2003. O2-PLS, a two-block (x–y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.* 17 (1), 53–64.
- Volteras, D., Shahrezaei, V., Thomas, P., 2024. Global transcription regulation revealed from dynamical correlations in time-resolved single-cell RNA sequencing. *Cell Syst.* 15 (8), 694–708.e12. <https://linkinghub.elsevier.com/retrieve/pii/S2405471224002011>. <https://doi.org/10.1016/j.cels.2024.07.002>
- Wei, Z., Zhang, S., 2021. CALLR: a semi-supervised cell-type annotation method for single-cell RNA sequencing data. *Bioinformatics* 37 (Supplement_1), i51–i58.
- Yu, W., Uzun, Y., Zhu, Q., Chen, C., Tan, K., 2020. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *Genome Biol.* 21 (1), 1–17.
- Zhang, K., Zemke, N.R., Armand, E.J., Ren, B., 2024. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nature Methods* 21 (2), 217–227.
- Zhao, X., Wu, S., Fang, N., Sun, X., Fan, J., 2020. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Briefings Bioinf.* 21 (5), 1581–1595.