PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'ENS PSL

Development of methods for paired single-cell multimodal integration

Soutenue par
Geert-Jan Huizing
Le 27 septembre 2024

École doctorale n°515
Complexité du Vivant

Spécialité
Bioinformatique et biologie des systèmes

Composition du jury :

Andrea Rau INRAE	*Présidente du jury*
Ahmed Mahfouz Leiden University Medical Center	*Rapporteur*
Franck Picard CNRS, ENS de Lyon	*Rapporteur*
Laurence Calzone Institut Curie	*Examinatrice*
Laura Cantini CNRS, Institut Pasteur	*Directrice*
Gabriel Peyré CNRS, ENS PSL	*Co-directeur*

1794 ENS | PSL

Remerciements

La rédaction de cette thèse marque la fin d'un long parcours que je n'aurais jamais pu accomplir seul. Je tiens donc à exprimer ma profonde gratitude envers toutes les personnes qui m'ont soutenu, accompagné et inspiré tout au long de cette aventure.

Je commence par remercier chaleureusement mes directeurs de thèse, Laura et Gabriel. Vous m'avez ouvert les portes de la recherche, mais plus encore, vous m'avez appris à me prendre au sérieux, à réfléchir avec rigueur et à créer avec ambition. Votre mentorat a été inestimable, et j'ai eu la chance de bénéficier de vos conseils avisés, que ce soit dans mes recherches, dans l'enseignement ou la présentation de mes travaux. Merci pour votre soutien et vos encouragements constants.

Je souhaite ensuite adresser un immense merci à l'ensemble de mon jury. Vos remarques, votre bienveillance et votre curiosité ont fait de ma soutenance un moment précieux et mémorable. À mes rapporteurs, Franck et Ahmed, qui ont eu la lourde tâche de lire et d'évaluer mon manuscrit durant l'été, merci pour votre implication et pour m'avoir permis d'approfondir ma réflexion. À Andrea, merci pour ton rôle essentiel de présidente du jury et pour tes questions judicieuses lors de la soutenance.

Je tiens également à remercier mes amis et compagnons de route de l'ENS et de l'Institut Pasteur. Je suis content d'avoir pu partager avec vous conférences, échanges scientifiques, sorties sportives, soirées pizza, et autres visites au musée. Durant les moments de doute, j'ai pu compter sur votre oreille attentive et votre empathie, et j'espère vous avoir rendu la pareille lorsque vous en aviez besoin. Un merci tout particulier à Jules et Rémi, présents tout au long de cette aventure.

Je remercie profondément mes parents, qui m'ont toujours encouragé, avec amour et compassion, à repousser mes limites. Merci de m'avoir montré la voie par votre exemple de persévérance et de dévouement. Merci de tout cœur à mon frère Pascal et à ma sœur Margot, pour votre soutien et votre gentillesse. Pascal, je te souhaite tout le meilleur pour ta thèse, que tu commences au moment où je termine la mienne.

À mes amis de Provence, merci pour cette amitié indéfectible qui a traversé les années.

À mes amis de Paris, vous avez été une source inestimable de soutien, de positivité et d'inspiration. C'est grâce à vous que j'ai osé me lancer dans l'aventure de la thèse, et c'est grâce à vous que je suis allé jusqu'au bout. Je suis extrêmement chanceux de vous avoir à mes côtés et vous remercie du fond du cœur pour votre encouragement et votre bienveillance.

À Ania, je ne saurais exprimer toute ma reconnaissance pour ton soutien inconditionnel. Ton amour et ta patience ont été une bouée essentielle pendant cette thèse.

Yasmine, je suis infiniment reconnaissant pour tous les instants passés ensemble. J'aurais tant aimé partager ce moment, et bien d'autres, avec toi.

Summary

Characterizing cellular diversity (i.e., cell types and cell states) and the mechanisms driving transitions between cell states is a long-standing goal of cell biology, made increasingly attainable by advances in high-throughput single-cell omics sequencing techniques. These technologies deliver quantitative molecular profiling at unprecedented resolution, facilitating the unbiased discovery of previously unknown cell types across tissues and elucidating differentiation trajectories between cell states in health and disease. Recently, spatial transcriptomics and paired single-cell multi-omics sequencing have enabled the joint profiling of complementary modalities, yielding insights into spatial patterns and regulatory mechanisms. However, analyzing paired multimodal data is challenging due to high dimensionality, noise, and the vast differences between modalities.

A promising strategy to address the challenges of paired single-cell multimodal data integration lies in Optimal Transport (OT), a geometrically meaningful way to compare probability distributions. Recent breakthroughs have reframed this centuries-old theory as a powerful tool for machine learning, which has proven particularly effective in analyzing single-cell omics data. Successful applications include modeling population dynamics and aligning data across different modalities. However, most existing applications focus on *unpaired* modalities rather than the joint analysis of *paired* modalities.

This thesis proposes novel OT-based methods for paired single-cell multimodal integration and is organized along two main research directions. The first direction focuses on *joint dimensionality reduction* of paired single-cell multi-omics data. Unlike existing methods, this thesis views the cell as a discrete distribution of molecular features (such as genes, genomic regions, or proteins). OT then provides a natural way to compare cells by leveraging the unique geometry underlying the features in each omics layer while mitigating the noise inherent to single-cell data. This new notion of similarity between cells is used as a machine learning objective to embed cells into an interpretable low-dimensional representation.

The second direction focuses on *trajectory inference* using spatial transcriptomics. This thesis proposes a causal model of cellular differentiation that formalizes Waddington's epigenetic landscape as a neural network trained on gene expression data profiled through space and time. The proposed method leverages a quadratic extension of OT to impose rotational and translational invariance for spatial coordinates. In contrast, existing methods cannot leverage spatial information since they rely on the original formulation of OT, which is sensitive to such transformations.

This thesis's methodological contributions are evaluated thoroughly in diverse biological settings such as the immune system and development, demonstrating that the developed methods provide valuable biological insights across modalities.

Resumé

Caractériser la diversité cellulaire (c'est-à-dire les types et les états cellulaires) et les mécanismes qui conduisent aux transitions entre les états cellulaires est un objectif de longue date de la biologie cellulaire, rendu de plus en plus atteignable par les avancées des techniques de séquençage omique à haut débit sur cellule unique. Ces technologies permettent un profilage moléculaire quantitatif à une résolution sans précédent, facilitant la découverte non biaisée de types de cellules auparavant inconnus à travers les tissus et l'élucidation des trajectoires de différenciation entre les états cellulaires en santé et en maladie. Récemment, la transcriptomique spatiale et le séquençage multi-omique pairé sur cellule unique ont permis le profilage conjoint de modalités complémentaires, offrant des renseignements sur les motifs spatiaux et les mécanismes de régulation. Cependant, l'analyse des données multimodales pairées est difficile en raison de la haute dimensionnalité, du bruit et des vastes différences entre les modalités.

Une stratégie prometteuse pour relever les défis de l'intégration des données multimodales pairées sur cellule unique réside dans le Transport Optimal (OT), une méthode géometriquement pertinente pour comparer des distributions de probabilité. Des avancées récentes ont fait de cette théorie séculaire en un outil puissant pour l'apprentissage automatique, particulièrement efficace dans l'analyse des données omiques sur cellule unique. Des applications fructueuses incluent la modélisation de la dynamique des populations et l'alignement des données à travers différentes modalités. Cependant, la plupart des applications existantes se concentrent sur de modalités *non pairées* plutôt que sur l'analyse conjointe de modalités *pairées*.

Cette thèse propose de nouvelles méthodes basées sur l'OT pour l'intégration multimodale pairée sur cellule unique et s'organise autour de deux axes de recherche principaux. Le premier axe se concentre sur la *réduction dimensionnelle conjointe* des données multi-omiques pairées sur cellule unique. Contrairement aux méthodes existantes, cette thèse considère la cellule comme une distribution discrète de caractéristiques moléculaires (telles que des gènes, des régions génomiques ou des protéines). L'OT fournit alors un moyen naturel de comparer les cellules en tirant parti de la géométrie unique des caractéristiques de chaque couche omique tout en atténuant le bruit inhérent aux données sur cellule unique. Cette nouvelle notion de similarité entre les cellules est utilisée comme une fonction d'objectif en apprentissage automatique pour intégrer les cellules dans une représentation de faible dimension interprétable.

Le deuxième axe se concentre sur l'*inférence de trajectoire* utilisant la transcriptomique spatiale. Cette thèse propose un modèle causal de différenciation cellulaire qui formalise le paysage épigénétique de Waddington par un réseau de neurones entraîné sur des données d'expression profilées dans l'espace et le temps. La méthode proposée exploite une extension quadratique de l'OT pour imposer une invariance rotationnelle et translationnelle aux coordonnées spatiales. Par opposition, les méthodes existantes ne peuvent pas tirer parti de l'information spatiale puisqu'elles reposent sur la formulation originale de l'OT, sensible ces isométries.

Les contributions méthodologiques de cette thèse sont évaluées de manière approfondie dans divers contextes biologiques tels que le système immunitaire et le développement, démontrant que les méthodes développées fournissent des reseignements biologiques précieux à travers les modalités.

Contents

1 Introduction **6**
- 1.1 High-throughput single-cell sequencing analysis . . . 7
- 1.2 Paired multimodal single-cell sequencing integration . . . 13
- 1.3 Optimal Transport . . . 16
- 1.4 Contributions of this thesis . . . 28

2 Optimal Transport improves cell-cell similarity inference in single-cell omics data **32**
- 2.1 Introduction . . . 33
- 2.2 Materials and methods . . . 34
- 2.3 Results . . . 40
- 2.4 Discussion . . . 46

3 Unsupervised ground metric learning using Wasserstein Singular Vectors **49**
- 3.1 Introduction . . . 50
- 3.2 Unsupervised Wasserstein metric learning . . . 53
- 3.3 Large scale stochastic power iterations . . . 58
- 3.4 Parallelization with entropic regularisation . . . 59
- 3.5 Metric learning for single-cell genomics . . . 61
- 3.6 Discussion . . . 63

4 Paired single-cell multi-omics data integration with Mowgli **65**
- 4.1 Introduction . . . 66
- 4.2 Results . . . 67
- 4.3 Discussion . . . 77
- 4.4 Methods . . . 78

5 Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein **86**
- 5.1 Introduction . . . 87
- 5.2 Results . . . 89
- 5.3 Discussion . . . 97
- 5.4 Methods . . . 99

6 Discussion **108**

Bibliography **112**

Introduction (en français) **140**

Appendices **167**

Notations

$\mathbb{R}$: Real numbers.
$\mathbb{N}$: Nonnegative integers.
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$: Metric spaces.
$\mathcal{M}(\mathcal{X}), \mathcal{P}(\mathcal{X})$: Measures on $\mathcal{X}$, probability measures on $\mathcal{X}$ (see page 17).

$\mathbf{A}, \mathbf{B}, \mathbf{C}$: Matrices.
$\mathbf{a}, \mathbf{b}, \mathbf{c}$: Vectors.
a, b, c: Scalars.
ρ, μ, ν: Measures.
θ, φ, ξ: Parameters of a neural network.

Σ_n: Probability n-simplex (see page 17).
$\mathcal{U}(\mathbf{a}, \mathbf{b})$: Discrete couplings (see page 18).

$f_{\#}\mu$: Pushforward measure (see page 100).
$\delta_{\mathbf{x}}$: Dirac measure (see page 29).
$\frac{\mathrm{d}\mu}{\mathrm{d}x}$: Density with respect to Lebesgue (see page 23).

$\langle \mathbf{x}, \mathbf{y} \rangle \stackrel{\text{def.}}{=} \sum_i x_i y_i$: Scalar product in $\mathbb{R}^d$.
$\|\mathbf{x}\|_p \stackrel{\text{def.}}{=} \left(\sum_i x_i^p\right)^{1/p}$: p-norm in $\mathbb{R}^d$.

$\mathbf{Id}_n, \mathbf{1}_n$: Identity matrix in $\mathbb{R}^{n \times n}$, ones vector in $\mathbb{R}^n$.
softmax: Soft argmax (see page 80).

$\mathcal{W}_c(\mu, \nu), \mathrm{W}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$: Wasserstein distance (see page 18).
$\mathcal{W}_c^\varepsilon(\mu, \nu), \mathrm{W}_{\mathbf{C}}^\varepsilon(\mathbf{a}, \mathbf{b})$: Entropy-regularized Wasserstein distance (see page 19).
$\overline{\mathcal{W}}_c^\varepsilon(\mu, \nu), \overline{\mathrm{W}}_{\mathbf{C}}^\varepsilon(\mathbf{a}, \mathbf{b})$: Sinkhorn divergence (see page 20).

$\mathcal{GW}(\mu, \nu), \mathcal{GW}^\varepsilon(\mu, \nu)$: (entropy-regularized) Gromov-Wasserstein distance (see page 21).
$\mathcal{FGW}(\mu, \nu), \mathcal{FGW}_\alpha^\varepsilon(\mu, \nu)$: (entropy-regularized) Fused Gromov-Wasserstein distance (see page 22).

$\nabla_{\mathrm{W}}\mathcal{F}(\mu)$: Wasserstein gradient (see page 23).
$\delta\mathcal{F}(\mu)$: First variation (see page 23).
$\mathrm{div}, \nabla, \Delta$: Divergence, gradient, Laplacian.

Chapter 1

Introduction

In the mid-17th century, Robert Hooke made a groundbreaking discovery while observing a piece of cork through his microscope. He observed structures that he named "cells" and, as a result, marked the beginning of cellular biology. [Hooke et al., 1665] Cells have since been identified as life's fundamental structural and functional units, and biologists have endeavored to map the diverse cell types that comprise multicellular organisms. Additionally, they have sought to understand the transient cell states that occur during development, disease progression, and tissue regeneration [Trapnell, 2015].

Recent efforts to cartography the landscape of cell identity, like the Human Cell Atlas project [Regev et al., 2017], have moved beyond microscopy to quantify cell states at the molecular level. Specifically, high-throughput single-cell RNA sequencing (scRNA-seq) technologies have enabled the characterization of cell identity through gene expression. Given the success of these methods, the cell has even been compared to a "bag of RNA." [Quake, 2021].

However, cell identity is influenced by a complex interplay of various molecular layers, including prote*omics*, transcript*omics*, and epigen*omics*. Accordingly, two critical new technological frontiers have emerged to understand the cell more comprehensively. The first is high-throughput multiomics single-cell sequencing, which simultaneously measures multiple omics within the same cell. The second is spatial transcriptomics, which quantifies gene expression while maintaining the spatial context of the cells. Together, these advancements fall under the umbrella of paired single-cell multimodal sequencing. [Vandereyken et al., 2023]

The joint analysis of multiple modalities in paired single-cell multimodal data, also called *paired single-cell multimodal integration*, is challenging and requires tailored computational methods. Indeed, single-cell data is sparse, noisy, and high-dimensional, and complementary modalities often have vastly different properties [Vandereyken et al., 2023]. This thesis focuses on developing computational methods addressing these challenges using the mathematical framework of Optimal Transport (OT) [Villani et al., 2009, Santambrogio, 2015]. OT is a promising way to compare and study high-dimensional objects in a geometrically meaningful way, and recent computational advances have made OT an essential tool in modern machine learning applications [Peyré et al., 2019]. In particular, OT has garnered significant interest in single-cell analysis [Hashimoto et al., 2016, Schiebinger et al., 2019, Demetci et al., 2022b], motivating the contributions in this thesis.

Contents

1.1 High-throughput single-cell sequencing analysis . . . 7
1.1.1 Single-cell sequencing technologies . . . 7
1.1.2 Computational analysis of single-cell sequencing data . . . 9
1.1.3 Biological impact of single-cell analysis . . . 12
1.2 Paired multimodal single-cell sequencing integration . . . 13
1.2.1 Paired multimodal single-cell sequencing technologies . . . 13
1.2.2 Computational paired single-cell multimodal integration . . . 14
1.2.3 Biological impact of paired multimodal single-cell integration . . . 16
1.3 Optimal Transport . . . 16
1.3.1 Mathematical background on Optimal Transport . . . 17
1.3.2 Computational background on Optimal Transport . . . 24
1.3.3 Applications of Optimal Transport in single-cell biology . . . 27
1.4 Contributions of this thesis . . . 28

1.1 High-throughput single-cell sequencing analysis

High-throughput single-cell omics sequencing refers to a set of technologies capable of quantifying the molecular identity of individual cells. These technologies analyze various molecular layers, including epigenomics and transcriptomics, within each cell. They have enabled the precise characterization of cell types and of differentiation trajectories, thereby enhancing our understanding of cellular diversity and dynamics in complex biological systems [Trapnell, 2015].

In recent years, numerous such single-cell sequencing technologies have been developed, along with computational methods to analyze the ever-growing volume of data produced in single-cell experiments. In this section I provide an overview of the field of single-cell sequencing, including the technologies that have been developed (Section 1.1.1), the computational methods that have been developed for the analysis of single-cell data (Section 1.1.2), and the applications of single-cell sequencing (Section 1.1.3).

1.1.1 Single-cell sequencing technologies

Single-cell sequencing technologies rely on *high-throughput sequencing* platforms such as Illumina, which enables the parallel sequencing of vast amounts of DNA strands [Goodwin et al., 2016]. Protocols such as scRNA-seq and scATAC-seq utilize this technology to sequence DNA libraries built from the molecular information contained in individual cells, enabling precise molecular quantification at single-cell resolution [Nawy, 2014, Preissl et al., 2023].

Single-cell sequencing protocols contrast with *bulk* sequencing protocols such as RNA-seq and ATAC-seq, which construct libraries from a pool of cells [Wang et al., 2009, Buenrostro et al., 2013]. Since they average signal from heterogeneous groups of cells, bulk assays suffer from Simpson's paradox, which is a statistical phenomenon where trends that appear in different groups of data disappear or reverse when these groups are combined [Trapnell, 2015].

In this section, I will cover in more detail the three single-cell modalities that are relevant to the

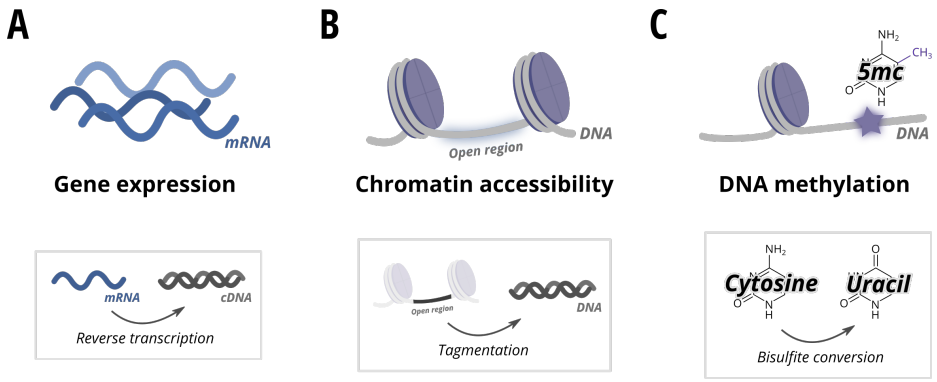
rest of the thesis, i.e. single-cell transcriptomics, single-cell chromatin accessibility, and single-cell methylation.

Single-cell transcriptomics *Single-cell RNA sequencing* (scRNA-seq) is a high-throughput technology that allows the measurement of gene expression at the single-cell resolution [Ziegenhain et al., 2017, Svensson et al., 2017, Mereu et al., 2020] for exponentially growing numbers of cells per experiment [Svensson et al., 2018, Kharchenko, 2021]. scRNA-seq protocols start with dissociating the tissue into a suspension of single cells, which are then physically isolated from each other using plate-based [Islam et al., 2011, Ramsköld et al., 2012, Picelli et al., 2013] or droplet-based [Macosko et al., 2015, Klein et al., 2015] technologies. Recent protocols introduce a molecular barcode unique to each cell, which attaches to the mRNAs and allows pooling the cells before sequencing [Klein et al., 2015]. In the case of combinatorial barcoding [Cao et al., 2017, Rosenberg et al., 2018], cells do not even need to be physically isolated. As in bulk RNA-seq, the captured mRNAs are reverse transcribed to cDNA molecules by reverse transcriptase [Svensson et al., 2018]. The cDNAs are then amplified because of the small amount of genetic material in a single cell: ~ 0.1 picograms of mRNA, compared to the microgram required by RNA-seq [Huang et al., 2018, Wang and Navin, 2015]. To keep track of the absolute amount of mRNAs before amplification, recent protocols tag each mRNA with a unique molecular identifier (UMI) [Ziegenhain et al., 2017]. The cDNA is then sequenced, reading either the 3' end of the transcript (e.g. [Macosko et al., 2015]) or the full-length transcript [Picelli et al., 2014]. The sequencing outputs reads which undergo a processing pipeline like CellRanger, which outputs a *count matrix* $\mathbf{A} \in \mathbb{N}^{m \times n}$ where m is the number of different mRNA species (There are around $m = 20,000$ protein-coding genes in the human genome) and n is the number of unique barcodes.

Single-cell chromatin accessibility The *single-cell assay for transposase-accessible chromatin sequencing* (scATAC-seq) is a high-throughput technology that allows the measurement of chromatin accessibility at single-cell resolution by measuring Tn5 transposase hypersensitivity sites [Preissl et al., 2023]. This information is crucial to understand the regulatory landscape of the cell, as chromatin architecture shapes the transcriptional state of the cell [Klemm et al., 2019]. scATAC-seq protocols build upon bulk ATAC-seq [Buenrostro et al., 2013] and leverage amplification to achieve single-cell resolution [De Rop et al., 2023]. As in single-cell transcriptomics, technological advances such as droplet-based isolation and combinatorial indexing [Lareau et al., 2019, Cusanovich et al., 2015] have allowed scaling to large numbers of cells. Representing scATAC-seq as a count matrix is less straightforward than with scRNA-seq, since there are no universal features to map the reads to [Heumos et al., 2023]. Rather than manipulating the reads directly, it is common to divide the genome into uniformly sized *bins* (for instance of 5kbp) or variable-length *peaks* and count Tn5 transposition events within these features [Fang et al., 2021].

Single-cell methylation *Single-cell DNA methylation sequencing* is a high-throughput technology that allows the measurement of DNA methylation at single-cell resolution. The most common form of DNA methylation is the addition of a methyl group to the 5th carbon of a cytosine nucleotide (denoted 5mC). This epigenetic modification usually happens when the cytosine is followed by a guanine (a CpG dinucleotide) and is often associated with repression of gene expression. It thus offers precious information on regulatory mechanisms underlying cell identity [Schwartzman and Tanay, 2015, Preissl et al., 2023]. Single-cell DNA methylation is typically measured using whole-genome bisulfite sequencing techniques, which convert unmethylated cytosines to uracil while leaving methylated cytosines unchanged [Preissl et al., 2023]. Examples for such technologies include scBS-seq [Smallwood et al., 2014], snmC-seq [Luo et al., 2017], and snmC-seq2 [Luo et al., 2018]. Similarly to scATAC-seq, the sequencing information is usually aggregated into bins, and one counts the

Figure 1.1: Overview of single-cell sequencing modalities.



(A) Single-cell RNA sequencing (scRNA-seq) technologies use reverse transcription to measure gene expression; **(B)** Single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq) technologies use tagmentation to measure chromatin accessibility; **(C)** Single-cell 5mC sequencing technologies use bisulfite conversion to measure DNA methylation.

number of methylated CpG cites within each bin to produce a cell-by-bin matrix [Luo et al., 2017].

Sparsity in single-cell data In scRNA-seq, the count matrix **A** contains a large proportion of zero entries, which may be due to the absence of transcription or to technical limitations of the scRNA-seq technology [Lähnemann et al., 2020]. The typical scRNA-seq dataset contains around 90% of zeros, and even greater sparsity has been observed in recent datasets with large numbers of cells, making sparsity an important theme in scRNA-seq analysis [Bouland et al., 2023]. Similarly, the typical scATAC-seq dataset contains hundreds of thousands of peaks, but around 90% of the count matrix's entries are zero [Baek and Lee, 2020]. A reason for this sparsity is the limited amount of genetic material available (two copies of the chromosome in diploid organisms), which also affects single-cell methylation sequencing [Lähnemann et al., 2020].

1.1.2 Computational analysis of single-cell sequencing data

As single-cell sequencing technologies become more widely applied, many methods have been developed to analyse the ever-growing volume of data from single-cell experiments. For instance, [Zappia and Theis, 2021] numbered over 1,000 tools for scRNA-seq analysis, covering a diversity of tasks. Fortunately, standards have emerged to ensure the intercompatibility of these methods, such as Seurat and Bioconductor in R [Butler et al., 2018, Amezquita et al., 2020] or Scanpy and AnnData in Python [Wolf et al., 2018, Virshup et al., 2021].

Below, I outline the main computational tasks in single-cell analysis discussed in this thesis. A more comprehensive account of methods in single-cell analysis may be found in [Luecken and Theis, 2019, Heumos et al., 2023].

Preprocessing

Single-cell analysis pipelines usually subject the count matrices $\mathbf{A} \in \mathbb{N}^{m \times n}$ (m features and n cells) described above to preprocessing steps such as *quality control* and *normalization* to account for the limitations of the sequencing protocols [Luecken and Theis, 2019, Heumos et al., 2023].

Quality control consists in filtering out uninformative cells such as doublets or dead cells. In scRNA-seq, quality control is typically based on *(i)* the total number of transcripts captured in a cell (the count depth) *(ii)* the number of genes per barcode and *(iii)* the fraction of counts from mitochondrial genes. [Luecken and Theis, 2019]. In scATAC-seq, quality control is typically based on *(i)* total fragments per cell *(ii)* log total fragments and *(iii)* and transcription start site enrichment score [Heumos et al., 2023]. In single-cell methylation sequencing, quality control is typically performed at the sequence level [Iqbal and Zhou, 2023].

Then, normalization steps aim to reduce technical noise and to facilitate downstream analysis. For instance, in scRNA-seq the capture of mRNAs is a stochastic process and differences in count depth are thus poorly informative with regard to cellular heterogeneity [Luecken and Theis, 2019]. Many normalization strategies have been proposed, including Pearson residuals [Lause et al., 2021], but a straightforward yet widely used approach is simply to divide each cell's counts by its count depth, followed by a nonlinear transformation such as $x \mapsto \log(1 + x)$ [Ahlmann-Eltze and Huber, 2023]. In ATAC-seq, peaks are often binarized [Fang et al., 2021] to encode open or closed chromatin, though recent work has shown a benefit to modeling fragment counts [Martens et al., 2024].

The shifted log transform preserves the sparsity of the data but reduces its skewness, making normalized values closer to a normal distribution [Luecken and Theis, 2019]. Principal Component Analysis (PCA), which is often the next step in single-cell analysis, is well suited to normally distributed data. Indeed, [Tipping and Bishop, 1999] show that PCA arises from the maximum likelihood estimation of the parameters of a Gaussian latent variable model. However, [Townes et al., 2019] argue that the sparse and skewed count data is better modeled with Poisson of Negative-Binomial distributions, and thus propose using a generalization of probabilistic PCA [Collins et al., 2001]. The generalized PCA can be approximated with the usual PCA on Pearson residuals (which are dense and approximately normally distributed) [Lause et al., 2021, Townes et al., 2019].

Dimensionality reduction

Single-cell data is typically very high-dimensional in nature. Consequently, many dimensionality reduction methods have been developed to facilitate visualization and downstream applications, such as clustering and trajectory inference [Heumos et al., 2023]. Prior to dimensionality reduction, features are often subset to the ones that vary most across the dataset [Luecken and Theis, 2019, Heumos et al., 2023].

Visualization of single-cell data relies for the overwhelming majority on t-SNE [Van der Maaten and Hinton, 2008] or UMAP [McInnes et al., 2018]. While these methods are efficient in visualizing discrete clusters of cells [Kobak and Berens, 2019], other methods, such as PHATE [Moon et al., 2019] and Isomap [Tenenbaum et al., 2000], are often preferred for visualizing continuous trajectories like development.

Downstream applications like clustering and trajectory inference also greatly benefit from dimensionality reduction. The most widely used approach is PCA, though many other tools have been

developed [Heumos et al., 2023]. Of particular interest to this thesis are Nonnegative Matrix Factorization (NMF) approaches, which offer intuitive part-based representations of data [Lee and Seung, 1999]. Neural network-based embeddings have gained attention for their expressive power in modeling single-cell data [Gayoso et al., 2022]. However, despite recent efforts to improve their interpretability [Ainsworth et al., 2018, Svensson et al., 2020] neural networks often lack the biological interpretability of linear models like NMF.

Unsupervised clustering

A common goal in single-cell data analysis is the identification of cell types and cell states within a sample [Luecken and Theis, 2019]. This is typically achieved through unsupervised clustering, where cells with similar profiles are grouped into clusters, which are then annotated based on known markers [Kiselev et al., 2019, Xiong et al., 2019]. Consequently, the quality of clustering is critical for accurate biological discovery.

In the early days of single-cell analysis, the methods used in bulk assays were applied in the single-cell setting. For instance, MARS-seq [Jaitin et al., 2014] used hierarchical clustering to detect immune cell populations in the spleen. However, due to increasing dataset sizes and improved performances, most modern single-cell analysis pipelines rely on grpah clustering (or community detection) algorithms like Leiden [Traag et al., 2019] or Louvain [Blondel et al., 2008].

Unsupervised clustering's success in revealing cellular heterogeneity in high-dimensional single-cell data appears to defy the curse of dimensionality, which suggests that in high dimensions, data points become uniformly distributed [Beyer et al., 1999]. But despite the apparent high dimensionality, single-cell data often organize into low-dimensional manifolds that can be captured by dimensionality reduction [Luecken and Theis, 2019, Moon et al., 2018].

A general challenge with the annotation of clusters is also the double-dipping phenomenon: when using statistical tests for differential expression of marker genes, one should account for the fact that the two populations are not independent since they come from clustering. [Lähnemann et al., 2020, Gao et al., 2024, Hivert et al., 2024]

Trajectory inference

Trajectory inference differs from clustering in that it seeks a continuous characterization of cell identity along cell fates, rather than a discrete one. [Saelens et al., 2019, Trapnell, 2015]

Pseudotime-based methods. Monocle [Trapnell et al., 2014] initiated this active field of research by introducing the concept of pseudotime, which orders cells along a biological process, such as cell differentiation or response to a stimulus. Since then, many other pseudotime-based trajectory inference methods have been developed, including Slingshot [Street et al., 2018], diffusion pseudotime [Haghverdi et al., 2016], and Palantir [Setty et al., 2019].

Velocity-based methods. A fundamentally different tool for trajectory inference is RNA velocity, which predicts the future state of a cell based on the ratio of spliced and unspliced mRNA [La Manno et al., 2018, Bergen et al., 2020]. However, RNA velocity relies on simple kinetic models that can misinterpret cell dynamics, for instance in the case of transient boosts in transcription [Bergen et al., 2021].

OT-based methods. Of particular interest in this thesis is the use of Optimal Transport (OT)

for trajectory inference in datasets where several time points have been profiled. Waddington-OT [Schiebinger et al., 2019] uses OT to compute transition probabilities between pairs of cells at adjacent time points, then uses these probabilities to infer cell trajectories. Another OT-based approach was pioneered by [Hashimoto et al., 2016], who used OT to learn a causal model of cell differentiation through a neural network trained to compute the differentiation potential of cells.

A convenient wrapper around different methods for trajectory inference is CellRank [Lange et al., 2022, Weiler et al., 2023], which allows the combination of several methods to improve the accuracy of inferred trajectories. Cellrank also enables the discovery of potential drivers genes of a trajectory, by regressing their expression against the pseudotime variable, using techniques such as splines or generalized additive models [Lange et al., 2022, Weiler et al., 2023].

1.1.3 Biological impact of single-cell analysis

The analysis of single-cell sequencing data has been used to study a wide range of biological questions, including *(i)* the characterization of cellular diversity and function, and *(ii)* the understanding of dynamic processes such as development and disease.

Characterizing cellular diversity and function Since its early days, single-cell omics sequencing techniques have been used to probe cellular heterogeneity across tissues, in health and disease. For instance, the MARS-seq sequencing technique was used to identify transcriptionally distinct clusters of immune cells in the spleen [Jaitin et al., 2014], the bone marrow [Paul et al., 2015], and the intestine [Gury-BenAri et al., 2016]. The scBS-seq protocol was used to explore 5mC heterogeneity in embryonic stem cells [Smallwood et al., 2014]. For instance, 10X Genomics's Single Cell 3' platform for snRNA-seq has been used to probe the transcriptomic signatures of specific cell-types in the brain, in the context of autism spectrum disorder [Velmeshev et al., 2019] and Alzheimer's disease [Leng et al., 2021]. The combinatorial indexing-based sci-ATAC-seq protocol was used to define cell-type specific regulatory signatures in pancreatic cells in the context of type 2 diabetes [Rai et al., 2020]. Today, much effort is devoted to assembling large multi-donor atlases from data profiled in different experiments. The flagship initiative is the Human Cell Atlas [Regev et al., 2017], which has already delivered a transcriptomic atlas of the lung [Sikkema et al., 2023], a chromatin accessibility atlas of the brain [Li et al., 2023b], and a DNA methylation atlas of the brain [Tian et al., 2023]. These large annotated datasets can serve as a reference to annotate smaller single-cell datasets [Lotfollahi et al., 2024].

Understanding the dynamic nature of cell identity Besides characterizing cellular heterogeneity, the second focus of single-cell sequencing research has been to understand the dynamics of cell identity in contexts such as development, ageing and disease. From the first single-cell experiments, researchers described continuous profiles between well characterized cell types, and sought to understand the mechanisms underlying these dynamics. The analysis of transcriptional profiles of Th17 immune cells obtained with SMART-seq revealed a continuum of cell states along differentiation and and genes involved in experimental autoimmune encephalomyelitis [Gaublomme et al., 2015]. A turning point in the study of single-cell dynamics was the formalisation of *pseudotime* and trajectory inference [Trapnell et al., 2014], which paved the way for large-scale studies of dynamic processes. For instance, Diffusion Pseudotime [Haghverdi et al., 2016] and Waddington OT [Schiebinger et al., 2019] were used to analyze trajectories in over 100,000 cells from nine time points during mouse development [Pijuan-Sala et al., 2019].

1.2 Paired multimodal single-cell sequencing integration

Understanding cell types, cell states and the evolution of cells in development, ageing, and disease require a holistic approach based on the simultaneous measurement of multiple molecular layers in the same cell [Vandereyken et al., 2023]. This is facilitated by paired multimodal single-cell sequencing, an umbrella term referring to refer to *(i)* the joint sequencing of multiple omics and *(ii)* spatially resolved omics sequencing [Heumos et al., 2023].

In this section I cover the multimodal protocols utilized in this thesis (Section 1.2.1), the integrative analysis of multimodal data (Section 1.2.2), an applications of paired multimodal single-cell sequencing and integration (Section 1.2.3).

1.2.1 Paired multimodal single-cell sequencing technologies

Simultaneous single-cell multiomics profiling

Paired single-cell multiomics technologies such as CITE-seq [Stoeckius et al., 2017] or the commercial 10x Multiome platform extend single-omics technologies by enabling the simultaneous readout of several molecular layers from the same cells. They generally complement transcriptomics with proteomic or epigenetic information to give a more complete characterization of the cell [Vandereyken et al., 2023].

One of the most widespread single-cell multiomics approach is the joint profiling of gene expression and surface protein abundance, using protocols like REAP-seq [Han et al., 2014] or CITE-seq [Stoeckius et al., 2017]. These technologies build upon droplet-based scRNA-seq through the addition of antibody-conjugated oligonucleotides. They enable measuring up to hundreds of proteins, but the necessity of creating a specific panel of antibodies limits the opportunities for unbiased discovery at the proteomic level. The joint analysis of the high-plex RNA layer and the low-plex protein layer is thus essential [Vandereyken et al., 2023].

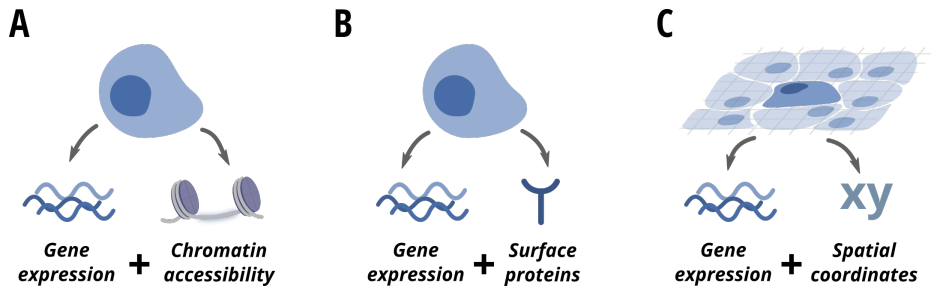
The other main type of single-cell multiomics protocol is the joint profiling of gene expression and chromatin accessibility by combining reverse transcription and tagmentation. A range of technologies have been developed, like 10x Multiome, sci-CAR [Cao et al., 2018] or scCAT-seq [Liu et al., 2019], which differ in terms of isolation and barcoding strategies, and in separation strategies for mRNA and DNA fragments [Vandereyken et al., 2023]. Recent technologies like TEA-seq enable the joint profiling of surface proteins on top of gene expression and chromatin accessibility [Swanson et al., 2021].

Spatially resolved single-cell omics

Single-cell sequencing typically requires the dissociation of the tissue (see Section 1.1.1). However, dissociation removes the cells of their context, obscuring spatial patterns and cell-cell interactions [Trapnell, 2015, Vandereyken et al., 2023, Larsson et al., 2021]. *Spatial transcriptomics* technologies enable to profile cells without dissociation, bridging the gap between the structural and functional aspects of the cell [Palla et al., 2022]. For this reason, spatial transcriptomics was named method of the year 2020 by Nature Methods [Marx, 2021].

Technologies for spatially resolved transcriptomics can be separated into *optics-based* and *spot-based* categories [Palla et al., 2022].

Figure 1.2: Paired multimodal single-cell sequencing integration.



(A) Technologies like 10X Multiome enable the joint profiling of gene expression and chromatin accessibility; **(B)** Technologies like CITE-seq [Stoeckius et al., 2017] enable the joint profiling of gene expression and a panel of surface proteins; **(C)** Technologies like Stereo-seq [Chen et al., 2022a] enable the profiling of a cell's gene expression along with its spatial coordinates.

- Optics-based techniques such as MERFISH [Xia et al., 2019] use fluorescent probes for a predefined set of genes, and detect each tagged mRNA along with its spatial coordinates at subcellular resolution. However, optics-based techniques do not allow to profile the whole transcriptome, which limits its potential for unbiased discovery [Palla et al., 2022].
- This thesis focuses on spot-based methods like Stereo-seq [Chen et al., 2022a], which enable full transcriptome profiling using spatial barcoding [Palla et al., 2022]. Spot-based methods output a count matrix $\mathbf{A} \in \mathbb{N}^{m \times n}$ where m is the number of genes and n is the number of spatial barcodes. Additionally, each of the n barcodes corresponds to a spatial position $(x, y) \in \mathbb{R}^2$.

Early spot-based methods such as [Ståhl et al., 2016] (which was later acquired by 10X Genomics) had a spatial resolution greater than the typical cell size, i.e. a number n of barcodes smaller than the number cells. Subcellular resolution has since been achieved by technologies such as Slide-seq [Rodriques et al., 2019], HDST [Vickovic et al., 2019] and Stereo-seq [Chen et al., 2022a], which offer a number n of barcodes greater than the number cells by using grids of DNA-barcoded beads.

This high resolution comes at the cost of a low count depth per bead, which is typically addressed by pooling barcodes into cell-sized bins or based on the cell segmentation of histology images [Chen et al., 2022a]. One then obtains a count matrix $\mathbf{A} \in \mathbb{N}^{m \times n}$ where m is the number of genes and n is the number of cells, as in Section 1.1.1.

1.2.2 Computational paired single-cell multimodal integration

The growing popularity of single-cell multimodal profiling requires tailored methods for the analysis of the resulting data, different from those developed for single-cell unimodal omics analysis. The joint analysis of several paired modalities is also known as *single-cell multimodal integration*. A wealth of methods have been proposed to fill this gap. As in single-cell unimodal analysis, standards have emerged to harmonize the application of these methods, such as Giotto [Dries et al., 2021] and SpatialExperiment [Righelli et al., 2022] in R or Squidpy [Palla et al., 2021] and SpatialData [Marconato et al., 2024] in Python.

Early steps of the analysis pipeline such as quality control and normalization are typically applied separately for each modality [Heumos et al., 2023]. In consequence, in the following I focus on dimensionality reduction and trajectory inference methods tailored for paired single-cell multimodal data.

Joint dimensionality reduction

Joint dimensionality reduction techniques enable to embed cells in a low-dimensional space informed by several modalities. State-of-the-art joint dimensionality reduction methods for single-cell multi-omics integration can be broadly classified as *linear models* such as matrix factorization [Cantini et al., 2021] and *nonlinear models* such as neural networks [Gayoso et al., 2022] or graph-based methods [Hao et al., 2021].

On the one hand, linear methods such as MOFA+ [Argelaguet et al., 2020] and MEFISTO [Velten et al., 2022] benefit from their interpretability since their low-dimensional embeddings can be directly related to the original data through their loadings, which represent aggregated features such as "metagenes" or "metapeaks" [Stein-O'Brien et al., 2018]. However, they are limited by their linearity, which may not capture the complex relationships between the different modalities [Stanojevic et al., 2022].

On the other hand, deep learning methods such as Cobolt [Gong et al., 2021], MultiVI [Ashuach et al., 2023], totalVI [Gayoso et al., 2021], GLUE [Cao and Gao, 2022] or graph-based models such as Weighted Nearest Neighbours [Hao et al., 2021] are better equipped to deal with the nonlinearities of single-cell multimodal data. But unlike linear models they do not benefit from a natural feature-level interpretability.

Striking a balance between the interpretability of linear methods and the expressivity of nonlinear methods is thus crucial to get the most out of single-cell multimodal data, and is a goal of Chapter 4.

Trajectory inference

As in single-cell unimodal analysis, trajectory inference methods based on pseudotime [Welch et al., 2017, Hou and Ji, 2022, Pham et al., 2023], velocity [Li et al., 2023a] and OT [Shen et al., 2023, Klein et al., 2023] are used to assign cells to continuous cell fate trajectories [Weiler et al., 2023].

Pseudotime-based methods. Pseudotime methods based on cell-cell similarity like Diffusion Pseudotime [Haghverdi et al., 2016] are typically based on the gene expression layer only, though recent work has sought to leverage joint spatial information [Pham et al., 2023, Hou and Ji, 2022] or epigenomic information [Welch et al., 2017].

Velocity-based methods. Likewise, RNA velocity's original model only accounts for gene expression information [La Manno et al., 2018, Bergen et al., 2020] but has recently been extended to include epigenetic information [Li et al., 2023a] and protein information [Gorin et al., 2020].

OT-based methods. Moscot [Klein et al., 2023] and spaTrack [Shen et al., 2023] have extended Waddington OT [Schiebinger et al., 2019] to spatial transcriptomics using a quadratic extension of OT and a custom spatially informed cost for linear OT respectively. While OT-based methods could in principle be applied to the joint dimensionality reduction of multiomics data, such experiments have not been reported yet.

Inferring cell fate trajectories from spatial transcriptomics is critical to discover the mechanisms underlying spatial organization in dynamic biological processes such as development [Palla et al., 2022]. Chapter 5 proposes a novel method based on Wasserstein gradient flows and quadratic OT (see Section 1.3.1), which enables to learn a causal model of differentiation from slices of spatial transcriptomics profiled at several time points.

1.2.3 Biological impact of paired multimodal single-cell integration

Paired multimodal single-cell technologies offer a precious glimpse into the regulatory and spatial contexts of cell identity. Below, I outline some successes of the integrative analysis of multimodal data in extending the insights gathered from unimodal omics.

Characterizing cellular diversity and function Cells have traditionally been annotated based on surface proteins using flow cytometry. For instance, in immunology, CD4 is a marker for Helper T cells. The joint profiling of transcriptomics and surface proteins has thus offered the possibility to bridge the gap between traditional surface protein-based cell type annotation and the unbiased discovery of transcriptomic signatures. For instance, [Hao et al., 2021] leverage the joint profiling of immune markers and gene expression to identify a subpopulation of CD8+ memory T cells in peripheral blood with characteristics of tissue-resident memory cells, pointing towards a previously unknown intermediary population. The same study [Hao et al., 2021] found that the joint profiling of transcriptomics and chromatin accessibility was more pertinent in separating naive CD8+ and naive CD4+ T cells than the joint profiling of transcriptomics and surface proteins, pointing to the importance of epigenetic profiles in uncovering cellular heterogeneity. Finally, spatially-resolved transcriptomics has discovered spatial patterns of gene expression that would be obscured by dissociation in single-cell RNA-seq. For instance, [Chen et al., 2020] study the cellular niche surrounding amyloid plaque in Alzheimer's disease, identifying transcriptional responses in the early and late stages of the disease.

Understanding the dynamic nature of cell identity The joint profiling of chromatin accessibility and gene expression can uncover regulatory dynamics underlying differentiation. For example, SHARE-seq was used to identify cell fates in hair follicle differentiation [Ma et al., 2020b]. This study observed that changes in chromatin accessibility preceded changes in gene expression during differentiation and pointed out transcription factors regulation differentiation within the different lineages. [Genshaft et al., 2016] used joint profiled RNA and proteins to quantify response to treatment at several time points in a breast cancer cell line, highlighting genes that exhibit different transcriptomic and proteomic patterns. An exciting perspective for the study of the dynamics of cellular identity is the use of spatial transcriptomics to build comprehensive spatiotemporal atlases, such as [Chen et al., 2022a, Liu et al., 2022] in mouse and zebrafish development and [Wei et al., 2022] in axolotl regeneration. Chapter 5 studies these datasets in more detail.

1.3 Optimal Transport

In the late 18th century, Gaspard Monge formulated the problem of transporting piles of earth from one location to another in a way that minimizes the total cost of transportation [Monge, 1781]. Two centuries later, a tractable formulation of Optimal Transport (OT) as a linear program earned Leonid Kantorovich a Nobel prize for economics [Kantorovich, 1942]. Today, developments in the efficient computation of OT have made it a powerful tool in machine learning, statistics, and signal processing [Peyré et al., 2019].

Pioneering works such as [Schiebinger et al., 2019] and [Hashimoto et al., 2016] have started an active field of research at the interface of OT and single-cell biology, which motivating this thesis's OT approach to single-cell multimodal integration. In this section I provide a brief overview of the mathematical background of OT 1.3.1, its computational aspects 1.3.2, and key applications of OT to single-cell biology 1.3.3.

1.3.1 Mathematical background on Optimal Transport

The following paragraphs outline the theory underlying this thesis's contributions, including the original Monge-Kantorovich problem, the quadratic Gromov-Wasserstein problem, and Wasserstein gradient flows. The monographs [Villani, 2003, Santambrogio, 2015] provide a more comprehensive account of the theory of OT.

The shared goal of all formulations is to match a source measure $\mu \in \mathcal{M}(\mathcal{X})$ to a target measure $\nu \in \mathcal{M}(\mathcal{Y})$ while minimizing some notion of cost. Here, $\mathcal{M}(\mathcal{X})$ denotes the space of Radon measures on a metric space $\mathcal{X}$. Typically, one considers probability measures $\mu \in \mathcal{P}(\mathcal{X})$, meaning that μ is positive and that $\int_{\mathcal{X}} \mathrm{d}\mu = 1$, though there exist extensions to positive measures with arbitrary mass [Peyré et al., 2019].

All numerical applications in this thesis involve discrete probability distributions on $\mathcal{X} = \mathbb{R}^d$. In this discrete setting, let us denote $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{a} \in \Sigma_n \overset{\text{def.}}{=} \{\mathbf{u} \in \mathbb{R}^n : \forall i \in [1, n], u_i \geq 0 \text{ and } \sum_i u_i = 1\}$.

Different chapters in this thesis use the mathematical framework of OT to represent two drastically different biological settings.

- In Chapter 2, Chapter 3, and Chapter 4, **the distribution $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ represents a cell**. The vectors $\mathbf{x}_i \in \mathbb{R}^d$ correspond to genes (or other molecular features), embedded in some d-dimensional space. The histogram $\mathbf{a} \in \Sigma_m$, where m is the number of genes, is proportional to the normalized count data.

- In Chapter 5 instead, **the distribution $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ represents a population of cells**. The cells $\mathbf{x}_i \in \mathbb{R}^d$ are embedded in a d-dimensional space, and correspond to weights a_i. The histogram $\mathbf{a} \in \Sigma_n$, with n the number of cells, may be uniform or proportional to an estimated growth rate [Schiebinger et al., 2019].

Most definitions in this section are given in the discrete case but extend naturally to continuous probability distributions, as described in [Peyré et al., 2019]. The algorithms presented in Section 1.3.2 to solve the various OT problems focus however on the discrete setting.

Monge's matching problem

Gaspard Monge's original formulation of OT [Monge, 1781] is a matching problem, where is worker seeks to move piles of sand from one configuration $\mu \in \mathcal{P}(\mathcal{X})$ to another configuration $\nu \in \mathcal{P}(\mathcal{X})$ with the least amount of effort.

In a discrete setting where μ and ν have the same number of points and uniform masses, i.e.

$$\mu = \frac{1}{n}\sum_i \delta_{\mathbf{x}_i} \text{ with } \mathbf{x}_i \in \mathcal{X}, \quad \nu = \frac{1}{n}\sum_j \delta_{\mathbf{y}_j} \text{ with } \mathbf{y}_j \in \mathcal{X}, \tag{1.1}$$

Monge's OT amounts to the assignment problem

$$\min_{\sigma \text{ permutation}} \frac{1}{n}\sum_i c(\mathbf{x}_i, \mathbf{y}_{\sigma(i)}). \tag{1.2}$$

In Equation 1.2 the function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is the *ground metric* that encodes the cost of transportation, typically a squared Euclidean distance. The bijective function $\sigma : [1, \dots, n] \to [1, \dots, n]$ *transports* the point $\mathbf{x}_i$ to the point $\mathbf{y}_{\sigma(i)}$.

In a setting like Chapter 5, where μ and ν are two populations of cells, Equation 1.2 compares μ and ν by transporting the cells $\mathbf{x}_i$ to cells $\mathbf{y}_{\sigma(i)}$ in a way that minimizes a discrepancy between their molecular profiles. However, the assignment problem defined in Equation 1.2 cannot compare μ and ν if they have different numbers of cells.

While this assignment problem can be extended to more general discrete probability measures (different numbers of points and non-uniform weights), it does not always admit a solution in that case [Peyré et al., 2019]. A more tractable formulation of the OT problem is Kantorovich's linear program, described in the next section.

Kantorovich's linear program

Kantorovich's formulation of the OT problem is a linear program [Kantorovich, 1942]. It can be seen as a probabilistic relaxation of Monge's deterministic problem in the sense that it allows *mass splitting*. Let us consider more general discrete probability measures μ and ν, possibly with different numbers of points and non-uniform masses:

$$\mu = \sum_i a_i \delta_{\mathbf{x}_i} \text{ with } \mathbf{a} \in \Sigma_n \text{ and } \mathbf{x}_i \in \mathcal{X}, \quad \nu = \sum_j b_j \delta_{\mathbf{y}_j} \text{ with } \mathbf{b} \in \Sigma_m \text{ and } \mathbf{y}_j \in \mathcal{X}. \tag{1.3}$$

The transfer of mass between discrete probability measures μ and ν (as defined in Equation 1.3) is thus encoded by a coupling $\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \{\mathbf{P} \in \mathbb{R}_+^{n \times m} \text{ such that } \sum_j P_{i,j} = a_i \text{ and } \sum_i P_{i,j} = b_j\}$.

Kantorovich's linear program then reads

$$\mathcal{W}_c(\mu, \nu) \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j} c(\mathbf{x}_i, \mathbf{y}_j) P_{i,j}. \tag{1.4}$$

This is a convex optimization problem which can be solved using the toolbox of linear programming, as explained in Section 1.3.2. Contrarily to Monge's formulation, this optimization problem always admits a solution. However, an optimal solution is not necessarily unique [Peyré et al., 2019].

An important feature of Equation 1.4 is that it not only defines a matching $\mathbf{P}$ between μ and ν, but also a distance $\mathbb{W}_p(\mu, \nu)$ between the two probability measures. Indeed, for the cost $c(\mathbf{x}, \mathbf{y}) = d(\mathbf{x}, \mathbf{y})^p$ with d a distance on $\mathcal{X}$ (typically the Euclidean distance), the quantity $\mathbb{W}_p(\mu, \nu) \overset{\text{def.}}{=} \mathcal{W}_c(\mu, \nu)^{1/p}$ is a distance on $\mathcal{P}(\mathcal{X})$ known as the *p-Wasserstein distance*.

Kantorovich's convex minimization problem admits a dual formulation as a concave maximization problem, which reads as follows:

$$\mathcal{W}_c(\mu,\nu) = \max_{\substack{\mathbf{f}\in\mathbb{R}^n,\ \mathbf{g}\in\mathbb{R}^m \\ \text{s.t. } f_i+g_j\leq c(\mathbf{x}_i,\mathbf{x}_j)}} \langle \mathbf{f},\mathbf{a}\rangle + \langle \mathbf{g},\mathbf{b}\rangle \tag{1.5}$$

Where $\langle \mathbf{f},\mathbf{a}\rangle \overset{\text{def.}}{=} \sum_i f_i a_i$ is the usual scalar product in $\mathbb{R}^n$.

Chapter 2, Chapter 3, and Chapter 4 propose seeing a cell as a histogram $\mathbf{a} \in \Sigma_m$ over a fixed set of m molecular features $\mathbf{x}_i$ (e.g. genes). To compare two cells $\mathbf{a}$ and $\mathbf{b}$ given a pairwise *cost matrix* $\mathbf{C} \in \mathbb{R}^{m\times m}$ between the features $\mathbf{x}_i$, i.e. $C_{i,j} = c(\mathbf{x}_i,\mathbf{x}_j)$, a convenient notation is

$$\mathrm{W}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \mathcal{W}_c(\sum_i a_i\delta_{\mathbf{x}_i}, \sum_j b_j\delta_{\mathbf{y}_j}). \tag{1.6}$$

Entropic regularization

Most of the recent success of OT in high-dimensional machine learning problems builds upon the entropic regularization of Equation 1.4, introduced as follows in [Cuturi, 2013].

$$\mathcal{W}_c^\varepsilon(\mu,\nu) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathcal{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} c(\mathbf{x}_i,\mathbf{y}_j)P_{i,j} - \varepsilon \mathrm{E}(\mathbf{P}), \tag{1.7}$$

where $\mathrm{E} : \mathbf{P} \mapsto -\sum_{i,j} P_{i,j}(\log P_{i,j} - 1)$ is the Shannon entropy. E is 1-strongly concave on $\mathcal{U}(\mathbf{a},\mathbf{b})$, which renders the optimization problem ε-strongly convex. Consequently, it has a unique minimizer $\mathbf{P}^\star$, contrarily to Monge and Kantorovich's formulations.

As in Kantorovich's problem, the entropy regularized formulation admits a dual formulation, defined as follows.

$$\mathcal{W}_c^\varepsilon(\mu,\nu) \overset{\text{def.}}{=} \max_{\mathbf{f}\in\mathbb{R}^n,\ \mathbf{g}\in\mathbb{R}^m} \langle \mathbf{f},\mathbf{a}\rangle + \langle \mathbf{g},\mathbf{b}\rangle - \varepsilon\langle e^{\mathbf{f}/\varepsilon}, \mathbf{K}e^{\mathbf{g}/\varepsilon}\rangle, \tag{1.8}$$

where $\mathbf{K} \overset{\text{def.}}{=} [e^{-c(\mathbf{x}_i,\mathbf{x}_j)/\varepsilon}]_{i,j}$ is the Gibbs kernel associated with the cost function c. [Cuturi, 2013] proposes a fast and straightforward algorithm to compute $\mathcal{W}_c^\varepsilon(\mu,\nu)$: Sinkhorn's algorithm, described in more detail in Section 1.3.2, which coincides with a block coordinate gradient ascent on the dual potentials $\mathbf{f},\mathbf{g}$.

Finally, Danskin's envelope theorem proves that $\mathcal{W}_c^\varepsilon(\mu,\nu)$ is differentiable with regard to the histograms $\mathbf{a},\mathbf{b}$ and the positions $\mathbf{x}_i,\mathbf{y}_j$, which is a desirable property for machine learning. As before, given a pairwise cost matrix $\mathbf{C}$ between vectors $\mathbf{x}_i$, we denote $\mathrm{W}_{\mathbf{C}}^\varepsilon(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \mathcal{W}_c^\varepsilon(\mu,\nu)$.

Chapter 4 uses the convex functional $\mathbf{a} \mapsto \mathrm{W}_{\mathbf{C}}^\varepsilon(\mathbf{a},\mathbf{a}_{\text{ref}})$ as an objective to approximate a cell $\mathbf{a}_{\text{ref}} \in \Sigma_m$ with the factorization $\mathbf{a} = \mathbf{H}\mathbf{w} \in \Sigma_m$. Here, $\mathbf{H} \in \mathbb{R}_+^{m\times d}$ is a dictionary whose columns $\mathbf{h}_j \in \Sigma_m$ are histograms over the m genes (or other molecular features), and the embedding $\mathbf{w} \in \Sigma_d$ is a histogram over the d latent factors.

Sinkhorn divergences

While p-Wasserstein distances define a metric between probability measures, the entropic regularization defined in Equation 1.7 is not a metric. Most importantly for practical numerical problems, the quantity $\mathcal{W}_c^\varepsilon(\mu,\mu) > 0$ for $\varepsilon > 0$. To address this bias, [Genevay et al., 2018] proposed using the Sinkhorn divergence, a debiased version of $\mathcal{W}_c^\varepsilon$ defined as

$$\overline{\mathcal{W}_c^\varepsilon}(\mu,\nu) \overset{\text{def.}}{=} \mathcal{W}_c^\varepsilon(\mu,\nu) - \tfrac{1}{2}\left(\mathcal{W}_c^\varepsilon(\mu,\mu) + \mathcal{W}_c^\varepsilon(\nu,\nu)\right). \tag{1.9}$$

[Feydy et al., 2019] give a thorough account of the properties of Sinkhorn divergences, including the fact that they interpolate between Wasserstein distances and the Maximum Mean Discrepancies (MMD).

Analogously to the previous paragraph, we denote $\overline{\mathrm{W}_{\mathbf{C}}^\varepsilon}(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \mathrm{W}_{\mathbf{C}}^\varepsilon(\mathbf{a},\mathbf{b}) - \frac{1}{2}\left(\mathrm{W}_{\mathbf{C}}^\varepsilon(\mathbf{a},\mathbf{a}) + \mathrm{W}_{\mathbf{C}}^\varepsilon(\mathbf{b},\mathbf{b})\right)$. Chapter 2 proposes using $\overline{\mathrm{W}_{\mathbf{C}}^\varepsilon}$ as a notion of similarity between cells $\mathbf{a},\mathbf{b}$ to improve downstream analysis, and Chapter 3 introduces a novel way to learn a pairwise cost $\mathbf{C}$ between features $\mathbf{x}_i,\mathbf{x}_j$.

Ground metric learning

A crucial aspect of OT is the cost metric c, which encodes the relationships between features. While a squared Euclidean distance between features is sufficient in most cases, practical machine learning problems may require a tailored metric.

- *Feature engineering.* A first strategy is to consider an Euclidean ground cost on a well-chosen embedding of the data. For instance, the Word Mover Distance [Kusner et al., 2015] between bag-of-word histograms uses Euclidean distances on precomputed Word2vec embeddings [Mikolov et al., 2013], i.e.

  $$c(\mathbf{x}_i,\mathbf{y}_j) = \|\mathrm{W2V}(\mathbf{x}_i) - \mathrm{W2V}(\mathbf{y}_j)\|_2$$

  where the function W2V assigns an embedding $\mathrm{W2V}(\mathbf{x}_i) \in \mathbb{R}^{d'}$ to a one-hot encoded word $\mathbf{x}_i \in \mathbb{R}^d$. The Gene Mover Distance [Bellazzi et al., 2021] has adapted this idea to single-cell transcriptomics using Gene2vec emebddings [Du et al., 2019]. Concurrently, the work presented in Chapter 2 proposed materializing molecular features (e.g. genes) as the rows of the count matrix, and computing distances between these vectors directly.

- *Supervised ground metric learning.* In a supervised setting, one might take advantage of methods such as [Cuturi and Avis, 2014, Wang and Guibas, 2012] which optimize a function F on the cone of distance-like matrices based on pairwise information $\omega_{k,l}$ between histograms $\mathbf{a}_k,\mathbf{b}_l$. For instance, [Cuturi and Avis, 2014] consider an image classification problem and define $\omega_{k,l} = 1$ if the histograms $\mathbf{a}_k$ and $\mathbf{b}_l$ have the same label, and $\omega_{k,l} = -1$ otherwise. The objective F is then defined as

  $$F(\mathbf{C}) = \sum_{k,l} \omega_{k,l} \mathrm{W}_{\mathbf{C}}(\mathbf{a}_k,\mathbf{b}_l).$$

- *Adversarial optimization.* Metric learning can also be performed through adversarial optimization. Similarly to Generative Adversarial Networks (GANs), [Genevay et al., 2018] train a generative model f_θ to map a uniform distribution μ to a target distribution ν (e.g. images), while learning a metric c_φ able to dicriminate between generated and real data:

  $$\min_\theta \max_\varphi \overline{\mathcal{W}_{c_\varphi}^\varepsilon}(f_{\theta\#}\mu,\nu).$$

  In the discrete case $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$, the pushworward measure $f_{\theta\#}\mu$ is simply the pointwise application of f_θ to the points $\mathbf{x}_i$. More formally, for $B \subset \mathcal{X}$, $(f_{\theta\#}\mu)(B) = \mu(f_\theta^{-1}(B))$ and verifies $\int g \,\mathrm{d}\,(f_{\theta\#}\mu) = \int g \circ f_\theta \,\mathrm{d}\mu$, for g measurable and $g \circ f_\theta$ integrable with respect to μ [Bogachev and Ruas, 2007].

- *Inverse OT.* Finally, another bi-level optimization problem called inverse OT aims to learn a metric c which minimizes a divergence $\mathcal{D}$ between an observed coupling $\hat{\mathbf{P}}$ and the coupling $\mathbf{P}_c$ induced by c (i.e. the coupling minimizing Equation 1.7), with a regularization term $\mathcal{R}(c)$ [Sun et al., 2020], i.e.
$$\min_c \mathcal{D}(\hat{\mathbf{P}}, \mathbf{P}_c) + \mathcal{R}(c).$$
For instance, [Dupuy and Galichon, 2014] consider relationships between people as an observed coupling $\hat{\mathbf{P}}$ and learn a metric c based on features such as the level of education or personality traits. One can choose $\mathcal{D}$ and $\mathcal{R}$ such that this bilevel optimization problem is convex [Carlier et al., 2023]. [Samaran et al., 2024] frame the alignment of unpaired single-cell data as an inverse OT problem, where the cost c is the Euclidean distance on the latent space of autoencoders.

Chapter 3 introduces a novel approach to learn a cost matrix $\mathbf{C}$ that requires *(i)* no feature engineering *(ii)* no supervision *(iii)* no adversarial optimization and *(iv)* no observed matching. The proposed method is applied to a single-cell dataset to learn a cost $\mathbf{C}$ encoding the geometry of the space of genes. Since there is no straighforward notion of cost between genes, learning $\mathbf{C}$ is of biological interest and defines a distance $\mathrm{W}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ between cells $\mathbf{a}, \mathbf{b} \in \Sigma_m$ (which are histograms on the space of genes).

Gromov-Wasserstein

Many numerical problems such as shape registration involve matching points up to an isometry (e.g. rotations or translations). In its original formulation, OT cannot handle matching in a way that is invariant to isometries. The Gromov-Wasserstein distance as introduced in [Mémoli, 2011] is a quadratic extension of Equation 1.4 that allows for such invariance.

In fact, in enables matching probability measures defined on different metric spaces $\mathcal{X}$ and $\mathcal{Y}$ by matching pairwise distances rather than pairs of points. In the discrete case,
$$\mu = \sum_i a_i \delta_{\mathbf{x}_i} \text{ with } \mathbf{a} \in \Sigma_n \text{ and } \mathbf{x}_i \in \mathcal{X}, \quad \nu = \sum_j b_j \delta_{\mathbf{y}_j} \text{ with } \mathbf{b} \in \Sigma_m \text{ and } \mathbf{y}_j \in \mathcal{Y}. \tag{1.10}$$

The Gromov-Wasserstein problem as introduced by [Mémoli, 2011] reads
$$\mathcal{GW}(\mu, \nu) \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j,i',j'} \ell\left(c_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_{i'}), c_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_{j'})\right) P_{i,j} P_{i',j'} \tag{1.11}$$

where typically, $\ell(s, s') \overset{\text{def.}}{=} |s - s'|^p$ and $c_{\mathcal{X}} = d_{\mathcal{X}}^q, c_{\mathcal{Y}} = d_{\mathcal{Y}}^q$ are powers of distance functions defined on $\mathcal{X}$ and $\mathcal{Y}$ respectively. In this case $\mathcal{GW}(\mu, \nu)^{1/p}$ defines a distance between the metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ [Sturm, 2012].

Contrarily to Kantorovich's linear program (Equation 1.4), this problem in inherently nonconvex and thus harder to solve. To introduce regularity to Equation 1.11, [Solomon et al., 2016, Peyré et al., 2016] introduce an entropic regularization term analogous to Equation 1.7:
$$\mathcal{GW}^{\varepsilon}(\mu, \nu) \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j,i',j'} \ell\left(c_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_k), c_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_l)\right) P_{i,j} P_{i',j'} - \varepsilon \mathrm{E}(\mathbf{P}). \tag{1.12}$$

For $\varepsilon > 0$, [Solomon et al., 2016, Peyré et al., 2016] propose an efficient algorithm to compute an approximate local minimum, which is based on an iterative linearization of the problem (see Section 1.3.2).

As proposed in [Demetci et al., 2022b], μ could represent a distribution of cells $\mathbf{x}_i$ defined on gene expression space $\mathcal{X}$ and μ a distribution of cells $\mathbf{y}_j$ defined on chromatin accessibility space $\mathcal{Y}$. Alternatively, μ and ν may represent two tissue slices in spatial transcriptomics, i.e. distributions of cells in the physical space [Zeira et al., 2022]. In this case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$, but since the two slices are not aligned, μ and ν can only be compared up to an isometry.

Fused Gromov-Wasserstein

In many machine learning settings, one may want to match point clouds where some features are comparable and other features are not. This is the case for example in spatial transcriptomics, where one can compare cells directly in the gene expression space, or up to an isometry in the physical space [Klein et al., 2023].

One can formalize this as $\mu \in \mathcal{P}(\mathcal{Z} \times \mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})$. In the discrete case,

$$\begin{aligned} \mu &= \sum_i a_i \delta_{(\mathbf{x}_i, \mathbf{r}_i)} \text{ with } \mathbf{a} \in \Sigma_n, \mathbf{x}_i \in \mathcal{Z}, \text{ and } \mathbf{r}_i \in \mathcal{X} \\ \nu &= \sum_j b_j \delta_{(\mathbf{y}_j, \mathbf{s}_j)} \text{ with } \mathbf{b} \in \Sigma_m, \mathbf{y}_j \in \mathcal{Z}, \text{ and } \mathbf{s}_j \in \mathcal{Y} \end{aligned} \tag{1.13}$$

The Fused Gromov-Wasserstein distance [Vayer et al., 2018b] interpolates between a Wasserstein linear term and a Gromov-Wasserstein quadratic term:

$$\mathcal{FGW}_\alpha^\varepsilon(\mu, \nu) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \alpha \sum_{i,j,i',j'} \ell\left(c_\mathcal{X}(\mathbf{r}_i, \mathbf{r}_k), c_\mathcal{Y}(\mathbf{s}_j, \mathbf{s}_l)\right) P_{i,j} P_{i',j'} + (1-\alpha) \sum_{i,j} c_\mathcal{Z}\left(\mathbf{x}_i, \mathbf{y}_j\right) P_{i,j} - \varepsilon \mathrm{E}(\mathbf{P}) \tag{1.14}$$

where $\alpha \in [0, 1]$ is a trade-off parameter between the two terms. In the spatial transcriptomics case mentioned above, α interpolates between a discrepancy based on gene expression ($\alpha = 0$) and a discrepancy based on morphology ($\alpha = 1$).

Similarly to the Gromov-Wasserstein distance, the Fused Gromov-Wasserstein distance is nonconvex but can be approximated efficiently for $\varepsilon > 0$ (see Section 1.3.2). Chapter 5 thus uses $\mathcal{FGW}_\alpha^\varepsilon$ to compare a ground-truth spatial transcriptomics experiment μ with a predicted distribution $\rho(\theta)$, where θ are the parameters of a neural network.

Wasserstein gradient flows

In biology, studying dynamic processes often involves minimizing some notion of energy over time. For instance, Waddington compared cellular differentiation to a marble rolling down a landscape towards several possible attractor states, which represent differentiated cell types [Allen, 2015]. [Huang et al., 2007] show that such a potential landscape naturally arises from simple gene regulatory networks.

One can formalize this idea of a cell $\mathbf{x} \in \mathbb{R}^d$ minimizing a functional $J : \mathbb{R}^d \to \mathbb{R}$ as an *Euclidean gradient flow*, i.e. the continuous counterpart of gradient descent. An Euclidean gradient flow $t \mapsto \mathbf{x}_t \in \mathbb{R}^d$ for the functional J verifies the following partial differential equation (PDE):

$$\frac{\partial \mathbf{x}_t}{\partial t} = -\nabla J(\mathbf{x}_t). \tag{1.15}$$

A "forward Euler" discretization of this PDE is simply gradient descent: $\mathbf{x}_{t+\tau} = \mathbf{x}_t - \tau \nabla J(\mathbf{x}_t)$. A "backwards Euler" discretization is given by $\mathbf{x}_{t+\tau} = \arg\min_{\mathbf{x}} \tau J(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2$, which is commonly known as proximal gradient descent [Santambrogio, 2017].

To study population dynamics instead of cellular dynamics, one can resort to *Wasserstein gradient flows*, which generalize Euclidean gradient flows to the space $(\mathcal{P}(\mathcal{X}), \mathbb{W}_2)$ of probability measures [Ambrosio et al., 2005, Santambrogio, 2017].

Given a functional $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$, Wasserstein gradient flows $t \mapsto \mu_t \in \mathcal{P}(\mathbb{R}^d)$ verify the PDE

$$\frac{\partial \mu_t}{\partial t} = \operatorname{div}(\mu_t \nabla_{\mathrm{W}} \mathcal{F}(\mu_t)), \tag{1.16}$$

where the Wasserstein gradient $\nabla_{\mathrm{W}}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d$ is a gradient field defined as $\nabla_{\mathrm{W}}\mathcal{F}(\mu) \overset{\text{def.}}{=} \nabla\left(\delta\mathcal{F}(\mu)\right)$, and $\delta\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}$ is the first variation of $\mathcal{F}$ evaluated at μ, which satisfies

$$\mathcal{F}(\mu + \eta\nu) = \mathcal{F}(\mu) + \eta \int_{\mathbb{R}^d} \delta\mathcal{F}(\mu)(\mathbf{x}) \mathrm{d}\nu(\mathbf{x}) + o(\eta). \tag{1.17}$$

As explained in more detail in [Santambrogio, 2017], forward and backward Euler discretization schemes for Equation 1.16 are given by

$$\mu_{t+\tau} = (\mathrm{Id} - \tau\nabla_{\mathrm{W}}\mathcal{F})_{\#}\mu_t \quad \text{and} \quad \mu_{t+\tau} = \arg\min_{\mu} \tau\mathcal{F}(\mu) + \tfrac{1}{2}\mathbb{W}_2^2(\mu, \mu_t). \tag{1.18}$$

In the case of a potential energy $\mathcal{F}(\mu) = \int_{\mathbb{R}^d} J(\mathbf{x})\mathrm{d}\mu(\mathbf{x})$, the first variation $\delta\mathcal{F}(\mu) = J$ does not depend on μ. Equation 1.16 then becomes an advection PDE

$$\frac{\partial \mu_t}{\partial t} = \operatorname{div}(\mu_t \nabla J). \tag{1.19}$$

In the discrete case, the forward scheme $\mu_{t+\tau} = (\mathrm{Id} - \tau\nabla J)_{\#}\mu_t$ corresponds to Euclidean gradient descent (Equation 1.15) applied independently at each point $\mathbf{x}_i$.

Jordan, Kinderlehrer and Otto [Jordan et al., 1998] show that in the case of the Gibbs-Boltzmann negentropy, $\mathcal{F}(\mu) = \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}x}(\mathbf{x})\right) \mathrm{d}\mu(\mathbf{x})$ and $\mathcal{F}(\mu) = +\infty$ if μ does not have a density, Equation 1.16 becomes to the heat equation

$$\frac{\partial \mu_t}{\partial t} = \Delta\mu_t. \tag{1.20}$$

Combining this with an advection term, i.e.

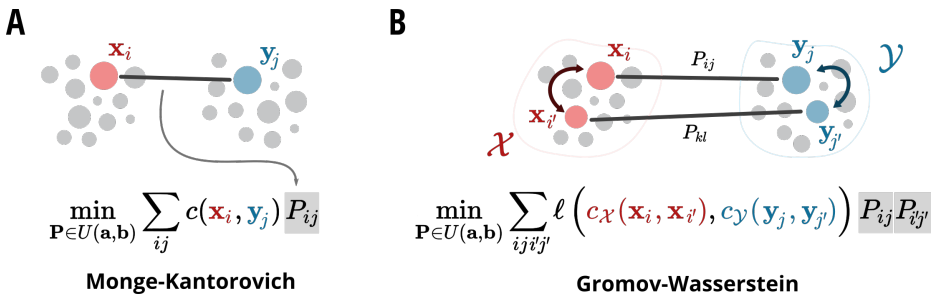
$$\mathcal{F}(\mu) = \int_{\mathbb{R}^d} \left[J(\mathbf{x}) + \beta^{-1} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}x}(\mathbf{x})\right)\right] \mathrm{d}\mu(\mathbf{x}), \tag{1.21}$$

[Jordan et al., 1998] recover the Fokker-Planck equation:

$$\frac{\partial \mu_t}{\partial t} = \operatorname{div}(\mu_t \nabla J) + \beta^{-1}\Delta\mu_t. \tag{1.22}$$

In the context of spatial transcriptomics, Chapter 5 aims to train a neural network J_θ such that observed distributions μ_{t_k} for $k = 0, \dots, K$ can be modeled by a gradient flow for the energy $\mathcal{F}(\mu) = \int_{\mathbb{R}^d} J_\theta(\mathbf{x})\mathrm{d}\mu(\mathbf{x})$, discretized with a forward scheme.

Figure 1.3: Linear and quadratic Optimal Transport.



(A) The Monge-Kantorovich problem outlined in Equation 1.4 seeks to transport points $\mathbf{x}_i \in \mathcal{X}$ to points $\mathbf{y}_j \in \mathcal{X}$ with a coupling $\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})$ in a way that minimizes the overall transportation cost; **(B)** The Gromov-Wasserstein problem outlined in Equation 1.12 seeks to transport points $\mathbf{x}_i \in \mathcal{X}$ to $\mathbf{y}_j \in \mathcal{Y}$ and $\mathbf{x}_k \in \mathcal{X}$ to $\mathbf{y}_l \in \mathcal{Y}$ with a coupling $\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})$ in a way that preserves pairwise distances.

A similar framework has been proposed for gene expression data by [Hashimoto et al., 2016], who use a forward Euler scheme, and [Bunne et al., 2022] who use a backward Euler scheme. However, [Hashimoto et al., 2016, Bunne et al., 2022] use a Sinkhorn divergence to evaluate the predicted expression profiles, while Chapter 5 proposes a debiased Fused Gromov-Wasserstein objective to include spatial coordinates.

1.3.2 Computational background on Optimal Transport

Below, I outline the key principles underlying the efficient computation of OT matchings for discrete probability measures, which enable the applications of OT proposed in this thesis. For a comprehensive account of OT's computational aspects, see [Peyré et al., 2019].

Efficient implementations of the algorithms presented here are implemented in the following Python libraries: POT [Flamary et al., 2021], GeomLoss [Feydy et al., 2019], and OTT [Cuturi et al., 2022]. POT aims to provide a comprehensive toolbox for OT, including solvers for the regularized and unregularized settings. Geomloss focuses on the fast computation of Sinkhorn divergences and their gradients for large-scale machine learning applications. OTT is the most recent addition to the Python landscape of OT solvers. It uses a JAX backend to compute regularized OT problems with a focus on scalability and differentiability.

Linear Optimal Transport

Matching n points to n other points with Monge's formulation is inherently difficult in general, as this problem is non-convex and requires searching through $n!$ possible matchings [Peyré et al., 2019].

The first tractable formulation of OT is thus Kantorovich's linear program, which is a convex optimization problem that can be solved efficiently in $\mathcal{O}(n^3)$ using the Hungarian algorithm when the

marginals are both equal and uniform [Kuhn, 1955], or $\mathcal{O}(n^3 \log(n)^2)$ using the network simplex algorithm [Tarjan, 1997].

The entropy regularized formulation gives rise to an efficient computation of the transport plan using Sinkhorn's algorithm [Sinkhorn, 1964], detailed in Algorithm 1. Sinkhorn's iterations have a complexity of $\mathcal{O}(n^2)$, and correspond to a coordinate gradient ascent on the dual potentials of Equation 1.8. A major advantage of Sinkhorn's algorithm is that is is easily parallelizable and benefits from GPU acceleration [Cuturi, 2013].

Algorithm 1 Sinkhorn's algorithm [Sinkhorn, 1964]

Initialize $\mathbf{f}, \mathbf{g}$
for $k = 1, \dots, n_{\text{iter}}$ **do**
  $\mathbf{f} \leftarrow \varepsilon \log \mathbf{a} - \varepsilon \log(\mathbf{K}\, e^{\mathbf{g}/\varepsilon})$
  $\mathbf{g} \leftarrow \varepsilon \log \mathbf{b} - \varepsilon \log(\mathbf{K}^T e^{\mathbf{f}/\varepsilon})$
end for
Return $\mathbf{P}^\star = \text{diag}(e^{\mathbf{f}/\varepsilon}) \mathbf{K}\, \text{diag}(e^{\mathbf{g}/\varepsilon})$

Using entropy regularized OT as an objective in machine learning settings requires the efficient differentiation of $\mathcal{W}_c^\varepsilon$ with regard to the histograms $\mathbf{a}, \mathbf{b}$ or the positions $\mathbf{x}_i, \mathbf{y}_j$. A naive application of Pytorch or JAX's automatic differentiation would unroll the Sinkhorn iterations, thus becoming costly for large numbers of iterations [Feydy et al., 2019]. However, as described in Proposition 4.6. of [Peyré et al., 2019], one can benefit from Danskin's envelope theorem to get the simple expression

$$\nabla \mathcal{W}_c^\varepsilon(\mu, \nu) = \nabla \mathcal{E}_{\mathbf{f}^\star, \mathbf{g}^\star} \text{ where } \mathcal{E}_{\mathbf{f}^\star, \mathbf{g}^\star} \overset{\text{def.}}{=} \langle \mathbf{f}^\star, \mathbf{a} \rangle + \langle \mathbf{g}^\star, \mathbf{b} \rangle - \varepsilon \langle e^{\mathbf{f}^\star/\varepsilon}, \mathbf{K} e^{\mathbf{g}^\star/\varepsilon} \rangle, \tag{1.23}$$

given $(\mathbf{f}^\star, \mathbf{g}^\star)$ a solution of Equation 1.8 such that the coordinates sum to zero. This efficient computation of the gradient is implemented as the default behavior in GeomLoss [Feydy et al., 2019] and OTT [Cuturi et al., 2022].

Finally, in some instances, minimizing an entropy regularized OT objective does not even require to compute the Sinkhorn iterations. For instance, [Cuturi and Peyré, 2016] provide a closed form expression of the Fenchel-Legendre dual of the function $\mathbf{a} \mapsto \text{W}_{\mathbf{C}}^\varepsilon(\mathbf{a}, \mathbf{b})$ and of its gradient. [Rolet et al., 2016] leverage this expression to provide a fast algorithm for dictionary learning on histograms with an entropy regularized OT objective.

Quadratic Optimal Transport

The non-regularized Gromov-Wasserstein and Fused Gromov-Wasserstein are nonconvex quadratic optimization problems, but a local solution can be achieved using the Frank-Wolfe algorithm [Frank et al., 1956], which requires solving a linear OT problem at each iteration [Vayer et al., 2018b].

The entropic regularization introduced in [Peyré et al., 2016] does not render the problem convex, but it gives rise to an efficient algorithm to compute an approximate local minimum for $\varepsilon > 0$. For a well chosen step size, [Peyré et al., 2016] show that mirror descent for the Kullback-Leibler divergence coincides with an iterative linearization scheme, where each step requires solving Sinkhorn's algorithm. This is detailed in Algorithm 2.

As in the linear case, differentiation through the Algorithm 2 does not require unrolling the iterations. A gradient can be taken for fixed $\mathbf{P}$ and $\tilde{\mathbf{C}}$ [Brogat-Motte et al., 2022, Bunne et al., 2019], which is the default behavior in OTT [Cuturi et al., 2022].

Algorithm 2 Mirror descent for entropy regularized quadratic OT [Peyré et al., 2016]

Initialize $\mathbf{P}$
for $k = 1, \dots, n_{\text{iter}}$ **do**
  $\tilde{\mathbf{C}} \leftarrow [\sum_{i',j'} \ell(c_{\mathcal{X}}(\mathbf{x}_i - \mathbf{x}_{i'}) - c_{\mathcal{Y}}(\mathbf{y}_j - \mathbf{y}_{j'}))P_{i',j'}]_{i,j}$ `# Linearization of the quadratic term`
  $\tilde{\mathbf{C}} \leftarrow \alpha\tilde{\mathbf{C}} + (1-\alpha)[c_{\mathcal{Z}}(\mathbf{x}_i - \mathbf{y}_j)]_{i,j}$ `# Interpolation with the linear term`
  $\mathbf{P} \leftarrow \arg\min_{\tilde{\mathbf{P}}} \sum_{i,j} \tilde{C}_{i,j}\tilde{P}_{i,j} - \varepsilon E(\tilde{\mathbf{P}})$ `# Solved with Sinkhorn's algorithm`
end for
Return $\mathbf{P}$

Wasserstein gradient flows

Numerical applications of Wasserstein gradient flows (presented in Equation 1.16) require the discretization of the flow with a forward or backward scheme (see Equation 1.18).

Let us first consider an Eulerian setting, where one seeks histograms $\mathbf{a}_t \in \Sigma_n$ discretizing a Wasserstein gradient flow for a functional $F : \Sigma_n \to \mathbb{R}$. The backwards scheme reads

$$\mathbf{a}_{t+\tau} = \arg\min_{\mathbf{a} \in \Sigma_n} \tau F(\mathbf{a}) + \tfrac{1}{2}\mathrm{W}_2^2(\mathbf{a}_t, \mathbf{a}). \tag{1.24}$$

Equation 1.24 can be solved efficiently in the one dimensional case [Kinderlehrer and Walkington, 1999]. In higher dimension, there exist finite elements and finite volume schemes [Burger et al., 2009, Carrillo et al., 2015]. In addition, [Peyré, 2015] proposes an approximate method relying on entropic regularization to compute a discretized Wasserstein gradient flow for F convex,

$$\mathbf{a}_{t+\tau} = \arg\min_{\mathbf{a} \in \Sigma_n} \tau F(\mathbf{a}) + \mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}_t, \mathbf{a}), \tag{1.25}$$

which can be solved using a generalized Sinkhorn algorithm.

A Lagrangian setting corresponds to operating on positions $\mathbf{x}_t^{(i)} \in \mathbb{R}^d$ rather than on the histogram $\mathbf{a}_t \in \Sigma_n$. When the functional $\mathcal{F}$ is finite for discrete probability measures, the forward Euler scheme amounts to an Euclidean gradient descent $\mathbf{X}_{t+\tau} = \mathbf{X}_t - \tau\nabla F(\mathbf{X}_t)$ where $F(\mathbf{X}) \overset{\text{def.}}{=} \mathcal{F}(\sum_i a_i \delta_{\mathbf{x}_i})$. The gradient ∇F can be obtained through automatic differentiation.

However, the Gibbs-Boltzmann negentropy is not finite for discrete measures. Discretizing the Fokker-Planck equation (Equation 1.22) is achieved using density estimation [Beirlant et al., 1997] or through the addition of a noise term with a Euler-Maruyama scheme

$$\mathbf{x}_{t+\tau}^{(i)} = \mathbf{x}_t^{(i)} - \tau\nabla J(\mathbf{x}_t^{(i)}) + \sqrt{2\tau\sigma^2}\mathbf{z}_t \tag{1.26}$$

where $\mathbf{z}_t$ is an independent draw from a multivariate normal distribution [Kloeden et al., 1992].

Equation 1.26 is a stochastic discretization scheme, where points $\mathbf{x}_i$ do not interact. Deterministic schemes involving interactions between points are more computationally intensive, but enable to model a wider range of entropy functionals (and thus nonlinear PDEs on the density). The most straightforward approach is to convolve the empirical measure with a regularizing function (i.e. replacing the Diracs with Gaussian mixtures) which is generally called the *blob method* [Carrillo et al., 2019].

Backward solvers also exist for the Langrangian setting [Benamou et al., 2016, Westdickenberg and

Wilkening, 2010]. Recently, [Bunne et al., 2022] proposed to reparametrize the backward step using an Input Convex Neural Network (ICNN) f_ξ:

$$\mu_{t+\tau} = \arg\min_{\xi} \tau\mathcal{F}\left((\nabla f_\xi)_{\#}\mu_t\right) + \tfrac{1}{2}\int \|\mathbf{x} - \nabla f_\xi(\mathbf{x})\|^2 \mathrm{d}\mu_t(\mathbf{x}). \tag{1.27}$$

Chapter 5 is concerned with the problem of learning a functional $J_\theta : \mathbb{R}^d \to \mathbb{R}$ such that observed discrete measures μ_{t_k} can be modeled with a Wasserstein gradient flow for the functional $\mathcal{F} : \mu \mapsto \int J_\theta(\mathbf{x})\mathrm{d}\mu(\mathbf{x})$. Such a setting was originally introduced by [Hashimoto et al., 2016], who consider a forward Euler discretization and rely on automatic differentiation to compute the gradient $\nabla_{\mathbf{x}} J_\theta(\mathbf{x})$. [Bunne et al., 2022] consider the ICNN-based backward Euler discretization scheme described above, which leads to a costly bilevel optimization problem. A recent preprint proposes a less costly approach using an optimality conditions eliminating the need to solve a bilevel optimization problem [Terpin et al., 2024].

1.3.3 Applications of Optimal Transport in single-cell biology

Recent advances in scaling OT for machine learning applications have opened the possibility of analyzing high-dimensional single-cell data across modalities. The paragraphs below showcase some diverse applications of OT in single-cell genomics, highlighting the versatility of this framework to model biological problems and hinting at the new opportunities explored in this thesis.

OT as a cell-cell similarity

Chapter 2, Chapter 3 and Chapter 4 explore the idea of leveraging OT as a distance between single-cell profiles. In this setting, a cell is represented by a discrete probability distribution $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$. The vectors $\mathbf{x}_i \in \mathbb{R}^d$ correspond to genes (or other molecular features), embedded in some d-dimensional space. The histogram $\mathbf{a} \in \Sigma_m$, where m is the number of genes, is proportional to the normalized count data.

Chapter 2 proposes materializing the molecular features $\mathbf{x}_i$ (e.g. genes) as the rows of the count matrix $\mathbf{A}$, and computing distances between these vectors to obtain a pairwise cost matrix $\mathbf{C} \in \mathbb{R}_+^{m\times m}$ between genes. The Sinkhorn divergence $\overline{\mathcal{W}}_c^\varepsilon(\mu,\nu)$ then provides a notion of distance between two cells μ and ν, which is used for clustering tasks.

Concurrently, [Bellazzi et al., 2021] proposed the Gene Mover Distance which materializes genes $\mathbf{x}_i$ as Gene2vec embeddings [Du et al., 2019], but does not extend the approach to other omics. [Tong et al., 2022] repeat an experiment on scRNA-seq data from Chapter 2 with a fast approximation of OT to speed up the computation of pairwise distances between cells.

Chapter 3 introduces a novel metric learning approach to jointly learn a cost matrix $\mathbf{C}$ between features and a cost matrix $\mathbf{D}$ between cells, which recover the cell type structure in a peripheral blood mononuclear cell dataset.

Chapter 4 proposes a joint dimensionality reduction model for paired single-cell multiomics data based on Wasserstein Nonnegative Matrix Factorization (Wasserstein NMF), which uses the entropy regularized OT distance $\mathrm{W}_{\mathbf{C}}^\varepsilon$ as a reconstruction loss.

OT-based alignment across modalities

A notable application of OT for single-cell data, though out of scope for this thesis, is the alignment of *unpaired* modalities. This corresponds to considering two populations of cells $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ and seeking a coupling between individual cells.

For instance, SCOT [Demetci et al., 2022b, Demetci et al., 2022a] uses Gromov-Wasserstein to align single-cell data profiled for different omics. Similarly, NovoSpaRc [Moriel et al., 2021] uses Fused Gromov-Wasserstein to align single-cell transcriptomics to low-plex spatial transcriptomics or arbitrary spatial coordinates.

Finally, [Samaran et al., 2024] frame the alignment of unpaired single-cell data as an inverse OT problem, where a cost function $c : (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y} \mapsto \|f_\theta(\mathbf{x}) - g_\varphi(\mathbf{y})\|^2$ is learned from the data by training neural networks f_θ and g_φ.

Reconstructing trajectory dynamics from time-course data

The application that sparked the single-cell community's interest in OT is the modelization of population dynamics from single-cell snapshots across time. In this setting, one considers discrete distributions $\mu_t = \sum_i a_t^{(i)} \delta_{\mathbf{x}_t^{(i)}}$ where $\mathbf{x}_t^{(i)} \in \mathbb{R}^d$ is the omics profile of cell i at time t. The histogram $\mathbf{a}_t$ represents the weights of the cells, which may be uniform, or proportional to an estimated growth rate [Schiebinger et al., 2019]. Importantly, sequencing imposes lysis of the cells, so no *a priori* matching exists between two adjacent time points μ_t and $\mu_{t+\tau}$.

[Schiebinger et al., 2019] take a static approach to reconstruct cell dynamics from snapshots by computing matchings between adjacent timepoints μ_t and $\mu_{t+\tau}$ using entropy regularized OT. In the context of spatial transcriptomics, [Klein et al., 2023] extend the approach of [Schiebinger et al., 2019] to spatial transcriptomics by considering snapshots $\mu_t = \sum_i a_t^{(i)} \delta_{(\mathbf{x}_t^{(i)}, \mathbf{r}_t^{(i)})}$ where $\mathbf{x}_t^{(i)} \in \mathbb{R}^d$ is still the omics profile of a cell, and $\mathbf{r}_t^{(i)} \in \mathbb{R}^2$ is its position in space. Instead of a linear OT coupling, [Klein et al., 2023] match two adjacent time points using entropy-regularized Fused Gromov-Wasserstein.

Instead, [Hashimoto et al., 2016] take a dynamic approach by learning a potential function J_θ such that the Wasserstein gradient flow for this potential (see Equation 1.19) fits the observations μ_t, using a Sinkhorn divergence objective (see Equation 1.9). Given this context, Chapter 5 fills a critical gap: by providing a dynamic causal model of cells through space and time based on a Fused Gromov-Wasserstein objective, it is to [Hashimoto et al., 2016] what [Klein et al., 2023] is to [Schiebinger et al., 2019].

1.4 Contributions of this thesis

This thesis provides novel contributions to paired single-cell multimodal integration. Chapter 2, Chapter 3, and Chapter 4 focus on clustering and dimensionality reduction for paired single-cell multi-omics, while Chapter 5 focuses on trajectory inference in spatial transcriptomics.

Optimal Transport improves cell-cell similarity inference in single-cell omics data (Chapter 2)

Pairwise similarities between cells are the foundation of most downstream analysis for single-cell omics data, like clustering and visualization. Typical single-cell analysis pipelines use bin-bin distances like the Euclidean distance to compare cells. However, bin-bin distances ignore the interactions between the features. Chapter 2 proposes to compare single-cell omics profiles using OT, which is robust to sparsity and integrates a geometric understanding of the complex structure of the molecular features.

Chapter 2 formalizes the cell as a discrete probability measure $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$. The vectors $\mathbf{x}_i \in \mathbb{R}^m$ correspond to genes (or other molecular features), and the histogram $\mathbf{a} \in \Sigma_m$, where m is the number of genes, is proportional to the normalized count data. We propose choosing $\mathbf{x}_i$ as the rows of the normalized count matrix, and computing a pairwise cost matrix $\mathbf{C}$ between genes. We then propose using the Sinkhorn divergence $\overline{\mathrm{W}}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ as a distance between two histograms $\mathbf{a}, \mathbf{b} \in \Sigma_m$.

We show that the Sinkhorn divergence between cells reflects the cell type structure better than traditionally used distances in single-cell analysis. We benchmark the approach on simulated scRNA-seq data and on real single-cell omics datasets spanning gene expression, chromatin accessibility and DNA methylation. We then integrated our approach in a typical clustering pipeline using the Leiden algorithm, and show an improvement in clustering performance across all datasets.

Unsupervised ground metric learning with Wasserstein Singular Sectors (Chapter 3)

The previous chapter introduces a new way to compare cells, given an ad-hoc cost matrix $\mathbf{C}$ which encodes the geometry of the feature space. However, the absence of straightforward choice of ground cost motivates learning such a metric from the data. Chapter 3 proposes an unsupervised method to learn the ground metric $\mathbf{C}$.

As in the previous chapter, Chapter 3 formalizes cells as histograms $\mathbf{a}_k \in \Sigma_m$ on the space of genes, and seeks a cost $\mathbf{C} \in \mathbb{R}_+^{m \times m}$ between genes. In addition, Chapter 3 also transposes the problem and formalizes the genes as histograms $\mathbf{b}_i \in \Sigma_n$ on the space of cells. We provide an algorithm that computes a pairwise Wasserstein distance matrix $\mathbf{C}$ between genes and $\mathbf{D}$ between cells such that up to rescaling, $C_{i,j} = \mathrm{W}_{\mathbf{D}}(\mathbf{b}_i, \mathbf{b}_j)$ and $D_{k,l} = \mathrm{W}_{\mathbf{C}}(\mathbf{a}_k, \mathbf{a}_l)$. We provide results on existence and uniqueness of these Wasserstein Singular Vectors. In addition, we improve the scalability of the proposed method by introducing entropic regularization and stochastic power iterations.

On a synthetic dataset consisting of translated histograms, we show that the learned metric integrates geometrical properties and that the proposed algorithm converges at a linear rate. On a scRNA-seq dataset of peripheral blood we show that the metric $\mathbf{D}$ between cells better captures the cell type structure of the dataset than other metrics. In addition, the metric $\mathbf{C}$ between genes captures similarities between known markers of immune cell types, making it a biologically interpretable output.

Paired single-cell multi-omics data integration with Mowgli (Chapter 4)

The analysis of paired single-cell multiomics data requires tailored joint dimensionality reduction methods. Methods based on matrix factorization are popular because of their interpretability, but may miss nonlinear relationships accounted for by deep learning methods. It is thus crucial to improve matrix factorization-based methods to strike a balance between interpretability and perfor-

mance. Chapter 4 proposes a joint dimensionality reduction method for paired single-cell multi-omics based on joint Nonnegative Matrix Factorization (joint NMF) with an entropy-regularized OT reconstruction loss.

For each omics layer p (e.g., gene expression), Chapter 4 considers the molecular profile of a cell as a histogram $\mathbf{a}^{(p)} \in \Sigma_{m_p}$ over the space of features (e.g. genes), which is equipped with a cost matrix $\mathbf{C}^{(p)}$. We propose Mowgli, a joint NMF method with an entropy-regularized OT reconstruction loss which seeks to decompose the histogram $\mathbf{a}^{(p)} \in \Sigma_{m_p}$ as a product $\mathbf{H}^{(p)}\mathbf{w}$. The matrix $\mathbf{H}^{(p)} \in \mathbb{R}_+^{m_p \times k}$ is an omics-specific dictionary whose columns $\mathbf{h}_j^{(p)} \in \Sigma_{m_p}$ are histograms on the space of genes, peaks, or proteins. The embedding $\mathbf{w} \in \Sigma_k$ is shared across omics and represents the joint embedding of the cell in a k-dimensional latent space. This embedding is interpretable since each dimension can easily be mapped to a weighted list of genes (or other features).

In challenging controlled settings derived from cell lines data, Mowgli outperforms the state-of-the-art in terms of embedding quality and clustering performance. In more complex and heterogeneous datasets spanning several multiomics technologies, Mowgli's performance is competitive with the state-of-the-art while delivering more interpretable embeddings. Mowgli's highly interpretable embedding allows the in-depth characterization of several precise CD8+ T cell clusters in a TEA-seq dataset, jointly profiled for three omics.

Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein (Chapter 5)

Inferring cell fate trajectories from spatial transcriptomics is critical to discover the mechanisms underlying spatial organization in dynamic biological processes such as development. A promising framework for analyzing sequencing data across time is Wasserstein gradient flow learning. It provides a causal model of differentiation built around a neural network representing the differentiation potential. However, existing gradient flow learning methods are ill-equipped to handle spatial information and cannot analyze spatially resolved transcriptomic data. Chapter 5 proposes a novel method to leverage spatial coordinates that employs a Fused Gromov Wasserstein objective to learn a spatially informed potential.

Chapter 5 shifts the focus from individual cells to populations of cells. A slice of spatial transcriptomics as time t is represented as a discrete probability distribution $\mu_t = \sum_i a_i \delta_{(\mathbf{x}_i, \mathbf{r}_i)}$. The vectors $\mathbf{x}_i \in \mathbb{R}^d$ represent the gene expression profile of the i-th cell, while the vectors $\mathbf{r}_i \in \mathbb{R}^2$ represent the spatial coordinates of the cell. The weights of individual cells are encoded in a histogram $\mathbf{a} \in \Sigma_n$, with n the number of cells, which may be uniform or proportional to an estimated growth rate. We propose STORIES, a model based on Wasserstein gradient flows which learns a highly interpretable differentiation potential J_θ parametrized by a neural network. Indeed, the potential function is analogous to pseudotime as it orders cells according to their stage of differentiation. In addition, the vector $-\nabla J_\theta$ is analogous to velocity as it allows to predict the future transcriptomic state of a cell. Unlike existing methods, STORIES learns J_θ using a Fused Gromov-Wasserstein objective, which allows to compare the spatial coordinates of different slices in a way that is invariant to isometries.

We benchmarked our method on three large Stereo-seq spatiotemporal atlases spanning development and regeneration. In these atlases, we demonstrate that the Fused Gromov-Wasserstein matching is more biologically coherent than the linear matching used in existing potential-based methods. We also performed an in-depth analysis of axolotl neuron regeneration and mouse gliogenesis. In

these biological settings, STORIES recovered expected trajectories and trends of gene expression for known markers. In addition, STORIES uncovered putative driver genes and transcriptional regulators which may motivate further biological investigation.

Software contributions

The methodological developments in this thesis were distributed as open-source Python packages:

- **The `otscomics` package**, developed for Chapter 2, enables to compute pairwise Sinkhorn divergence matrices between cells in single-cell omics data. `otscomics` uses POT [Flamary et al., 2021] for the Sinkhorn algorithm and Pytorch [Paszke et al., 2019] for GPU computations. `otscomics` batches computations to benefit from parallelization in POT's entropic solver. Finally, `otscomics`'s documentation demonstrates how to use the tool with an AnnData [Virshup et al., 2021] object and Scanpy [Wolf et al., 2018] clustering functions.

  `pip install otscomics` https://github.com/cantinilab/ot-scomics

- **The `wsingular` package**, developed for Chapter 3, enables to compute Wasserstein Singular Vectors, Sinkhorn Singular Vectors and their stochastic variants (see Chapter 3). `wsingular` uses POT [Flamary et al., 2021] for its linear solver and its entropic solver, and Pytorch [Paszke et al., 2019] for GPU computations. In addition, `wsingular` provides a convenient function to check the uniqueness of singular vectors using the criterion in Theorem 3.2.5.

  `pip install wsingular` https://github.com/CSDUlm/wsingular

- **The `mowgli` package**, developed for Chapter 4, enables to perform the joint dimensionality reduction on paired single-cell multiomics data using a joint Nonnegative Matrix Factorization (joint NMF) model with a Wasserstein reconstruction loss (see Chapter 4). `mowgli` is implemented in Pytorch [Paszke et al., 2019] for GPU computations and for its L-BFGS optimizer. In addition, `mowgli` provides convenient functions to visualize the results and to perform downstream analysis. Finally, `mowgli` takes MuData [Bredikhin et al., 2022] objects as an input and is seamlessly integrated in the Scverse ecosystem [Virshup et al., 2023].

  `pip install mowgli` https://github.com/cantinilab/mowgli

- **The `stories-jax` package**, developed for Chapter 5, enables to perform trajectory inference on spatial transcriptomics slices profiled at several timepoints. Unlike the previous packages, `stories-jax` relies on JAX [DeepMind et al., 2020] for GPU acceleration and OTT [Cuturi et al., 2022] for fast computation for Fused Gromov-Wasserstein distances. The modular implementation of `stories-jax` allows to choose a forward Euler discretization or an ICNN-based backward Euler discretization, as well as the number of steps in the discretization and the presence of teacher-forcing (see Chapter 5). As before, `stories-jax` is seamlessly integrated in the Scverse ecosystem and takes AnnData objects as an input, making it easy to interface `stories-jax` with standard Python tools for trajectory inference, such as CellRank [Weiler et al., 2023].

  `pip install stories-jax` https://github.com/cantinilab/stories

Chapter 2

Optimal Transport improves cell-cell similarity inference in single-cell omics data

Abstract

High-throughput single-cell molecular profiling is revolutionizing biology and medicine by unveiling the diversity of cell types and states contributing to development and disease. The identification and characterization of cellular heterogeneity are typically achieved through unsupervised clustering, which crucially relies on a similarity metric.

We here propose the use of Optimal Transport (OT) as a cell-cell similarity metric for single-cell omics data. OT defines distances to compare high-dimensional data represented as probability distributions. To speed up computations and cope with the high dimensionality of single-cell data, we consider the entropic regularization of the classical OT distance. We then extensively benchmark OT against state-of-the-art metrics over 13 independent datasets, including simulated, scRNA-seq, scATAC-seq and single-cell DNA methylation data. First, we test the ability of the metrics to detect the similarity between cells belonging to the same groups (e.g. cell types, cell lines of origin). Then, we apply unsupervised clustering and test the quality of the resulting clusters. OT is found to improve cell-cell similarity inference and cell clustering in all simulated and real scRNA-seq data, as well as in scATAC-seq and single-cell DNA methylation data. All our analyses are reproducible through the OT-scOmics Jupyter notebook available at https://github.com/cantinilab/OT-scOmics.

This chapter was published as

Geert-Jan Huizing, Gabriel Peyré, Laura Cantini, *Optimal transport improves cell-cell similarity inference in single-cell omics data*, Bioinformatics, March 2022, https://doi.org/10.1093/bioinformatics/btac084

Contents

2.1 Introduction . . . 33
2.2 Materials and methods . . . 34
2.2.1 scRNA-seq data simulation . . . 34
2.2.2 Single-cell omics data acquisition and preprocessing . . . 35
2.2.3 Baseline metrics . . . 35
2.2.4 Optimal Transport distance . . . 36
2.2.5 Performance evaluation . . . 38
2.3 Results . . . 40
2.3.1 Comparing OT against baseline metrics based on cell-cell similarity detection 41
2.3.2 Comparing OT against baseline metrics based on hierarchical clustering . . . 42
2.3.3 Comparing a typical single-cell clustering workflow against its counterpart based on OT . . . 44
2.3.4 Open-source implementation and distribution . . . 45
2.4 Discussion . . . 46

2.1 Introduction

Allowing the measurement of gene expression in thousands of cells in a single experiment, single-cell RNA sequencing (scRNA-seq) has unveiled the diversity of the cells constituting human tissues [Stegle et al., 2015]. The possibility to assess cellular heterogeneity at a previously inaccessible resolution has profoundly impacted our understanding of development, of the immune system functioning, and of many diseases [Papalexi and Satija, 2018, Potter, 2018, Rajewsky et al., 2020]. While scRNA-seq is now mature, the single-cell technological development has shifted to the measurement of other omics, e.g. DNA methylation, proteome and chromatin accessibility [Lee et al., 2020, Ma et al., 2020a].

A common goal in single-cell data analysis is the identification of the cell types and cell states present in a sample [Luecken and Theis, 2019]. This is typically achieved in a data-driven fashion through unsupervised clustering [Kiselev et al., 2019, Xiong et al., 2019]. Cells with similar transcriptional profiles are assembled into clusters, which are then annotated based on markers [Kiselev et al., 2019]. As a consequence, the quality of such clustering plays a critical role in the derived biological discovery. While numerous clustering algorithms have been proposed, they all rely on a similarity metric for categorizing individual cells. Popular metrics include the Euclidean and Manhattan distances, Cosine similarity and Pearson correlation [Guo et al., 2015b, Kim et al., 2019, Macosko et al., 2015, Satija et al., 2015].

Optimal Transport (OT) emerged in the last decade as a promising mathematical toolkit to analyze and compare high-dimensional data using different variants of the Wasserstein distance [Peyré et al., 2019, Santambrogio, 2015]. Recently, applications of OT to biology have been proposed [Bellazzi et al., 2021, Cao et al., 2022, Demetci et al., 2022b, Schiebinger et al., 2019]. Some works use OT on cells in the context of trajectory inference [Schiebinger et al., 2019] and alignment of unpaired (i.e. independently profiled) scRNA-seq and scATAC-seq data [Cao et al., 2022, Demetci et al., 2022b]. Others apply OT on genes in scRNA-seq data and perform supervised cell classification, such as malignant versus normal cells [Bellazzi et al., 2021]. Of note, classical OT, as applied by [Bellazzi et al., 2021],

requires solving a costly linear optimization problem thus making computations in large single-cell data slow and sometimes computationally intractable. In addition, the use of supervised clustering in biological applications is of limited relevance. Indeed, most of the existing biological studies use single-cell data to discover and characterize the cell populations/types present in a biological sample, a task that is intrinsically unsupervised.

Here, we propose the use of OT as a cell-cell similarity metric for single-cell data. In particular, we use OT with entropic regularization [Cuturi, 2013], expected to control the systematic noise due to the stochasticity of gene expression at single-cell level and to the presence of dropouts. In addition, using OT with entropic regularization we could efficiently analyze datasets with large numbers of cells using a Graphics Processing Unit (GPU). We further extensively benchmark OT against state-of-the-art metrics. We apply the different metrics to single-cell data with known groups (e.g. cell types, cell lines of origin) and we evaluate their ability to detect the similarity between cells belonging to the same group. We then apply different unsupervised clustering algorithms to the computed distance matrices and test the quality of the resulting clusters. Of note, all the tests are performed in three conditions: *(i)* simulated scRNA-seq data, where the effect of the number of cells and of the size and overlap of the clusters can be tested in-depth; *(ii)* real scRNA-seq data, profiled from cell lines and colorectal tissue and *(iii)* other omics data, including scATAC-seq and single-cell DNA methylation.

All the performed analyses are reproducible using the OT-scOmics Jupyter notebook provided on GitHub (https://github.com/cantinilab/OT-scOmics). Users can also employ OT-scOmics to test the various metrics on new single-cell data and to evaluate the performances of other/new metrics.

2.2 Materials and methods

2.2.1 scRNA-seq data simulation

scRNA-seq data with 5000 genes and three underlying clusters have been simulated using the R Bioconductor package Splatter [Zappia et al., 2017]. Splatter simulates scRNA-seq data using the Splat model, built around a Gamma-Poisson distribution. Different parameters can be tuned in the Splatter simulation. Details on the parameters used to run Splatter are available in Supplementary Text A.1.

Using simulated scRNA-seq data, we can assess in-depth the influence of different factors on the quality of the inferred cell-cell similarities. We simulated five scRNA-seq datasets Table 2.1 obtained by varying three main factors:

1. The number of cells constituting the scRNA-seq dataset (`batchCells` parameter in Splatter). Datasets with 500, 1,000 and 10,000 cells are simulated. Of note, we thereby also test whether the different metrics can be computed for large numbers of cells.

2. The overlap of the clusters (`de.prob` parameter in Splatter). Overlapping and well-separated clusters are simulated varying `de.prob` between 0.4 and 0.7, respectively.

3. The equal or unbalanced size of the clusters (`group.prob` parameter in Splatter). We set the probabilities of the clusters either equal (⅓, ⅓, ⅓) or unbalanced (0.75, 0.20, 0.05). The unbalanced case reflects the more realistic scenario of a tissue composed of a mixture of prevalent and rare cell types or states.

2.2.2 Single-cell omics data acquisition and preprocessing

Several publicly available single-cell omics datasets have been employed Table 2.1. Only public datasets providing ground-truth labels for all the profiled cells were considered. The labels are intended to associate each cell to a specific group (e.g. cell type, cell line of origin). Of note, labels are only used to evaluate the quality of our results, as all the performed benchmarking is unsupervised.

For scRNA-seq data four datasets have been considered. First, the scRNA-seq data (called 'Liu scRNA') present in the scCAT-seq joint profiling of [Liu et al., 2019] containing 206 cells profiled from three cancer cell lines (HCT116, HeLa-S3, K562). Next, a bigger dataset composed of 561 cells profiled from seven cell lines (A549, GM12878, H1437, HCT116, IMR90, H1, K562) from [Li et al., 2017] (called 'Li cell lines'). Finally, two colorectal cancer (CRC) datasets, corresponding to primary CRC tumors and matched normal mucosa have been also taken into account. The first (called 'Li Tumor') contains 364 cells clustered into seven cell types: B cells, endothelial cells, epithelial cells, fibroblasts, macrophages, mast cells and T cells. The second (called 'Li NM') is composed of 266 cells clustered according to the same seven cell types.

Other single-cell omics are also included in our analysis: methylation and scATAC-seq data. For scATAC-seq data we considered: *(i)* the dataset included into the scCAT-seq joint profiling of [Liu et al., 2019] (called 'Liu scATAC-seq') composed of 206 cells extracted from three cancer cell lines (HCT116, HeLa-S3, K562); *(ii)* the leukemia scATAC-seq data from [Corces et al., 2016] (called 'Leukemia scATAC'), containing 391 cells and composed of monocytes and lymphoid-primed multipotent progenitors (LMPP) isolated from a healthy human donor, together with leukemia stem cells (SU070_LSC, SU353_LSC) and blast cells (SU070_Leuk, SU353_Blast), isolated from two patients with acute myeloid leukemia. To represent single-cell DNA methylation, we considered two neuronal snmC-seq datasets from [Luo et al., 2017]. The first (called 'scMethylation mouse') is composed of 3377 cells extracted from mouse frontal cortical neurons clustered into 16 neuronal subtypes. The second (called 'scMethylation human') is composed of 2740 cells, extracted from human frontal cortical neurons and clustered into 21 neuronal subtypes.

A summary of the considered datasets is available in Table 2.1. The downloaded datasets had already undergone standard preliminary preprocessing and, following standard practices [Luecken and Theis, 2019], we log-transformed the scRNA-seq counts and selected the 10,000 most varying features. Alternative preprocessing strategies for scRNA-seq [Hafemeister and Satija, 2019, Lun et al., 2016, Yip et al., 2017] have also been tested using the code provided in [Chen et al., 2021], with no impact on the results (Supplementary Figure A.1, Supplementary Table A.1, Supplementary Table A.2).

2.2.3 Baseline metrics

We consider a normalized single-cell omics dataset as a matrix $\mathbf{A} \in \mathbb{R}_{+}^{m \times n}$ whose n columns correspond to cells, and whose m rows correspond to features (e.g. peaks, genes). Given two cells indexed by i and j as columns of $\mathbf{A}$, different metrics are classically used to infer the similarity between their omic profiles $\mathbf{u} = \mathbf{A}[\cdot, i] = (u_1 \dots u_m)$ and $\mathbf{v} = \mathbf{A}[\cdot, j] = (v_1 \dots v_m)$. We here focus on four state-of-the-art metrics, henceforth called baseline metrics, and defined as

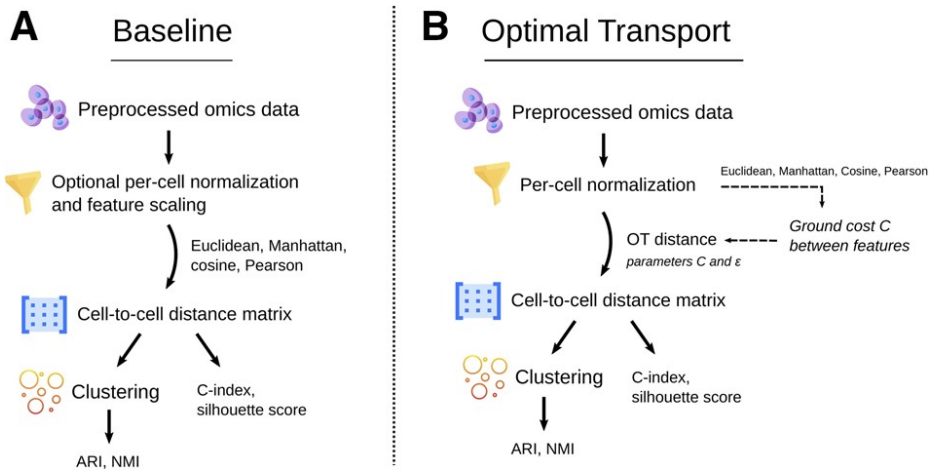
1. Euclidean distance (L2): $\|\mathbf{u} - \mathbf{v}\|_2 = (\sum_k (x_k - y_k)^2)^{1/2}$
2. Manhattan distance (L1): $\|\mathbf{u} - \mathbf{v}\|_1 = \sum_k |x_k - y_k|$
3. Cosine similarity: $\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}$ where $\| \cdot \|_2$ is the Euclidean norm

4. Pearson correlation: $\text{corr}(\mathbf{u}, \mathbf{v}) = \cos(\mathbf{u} - \bar{\mathbf{u}}, \mathbf{v} - \bar{\mathbf{v}})$ where $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$ are the mean of the values of and $\mathbf{u}$ and $\mathbf{v}$ respectively.

For cosine similarity and Pearson correlation, we used the distance-like formulation $1 - \cos(\mathbf{u}, \mathbf{v})$ and $1 - \text{corr}(\mathbf{u}, \mathbf{v})$. Of note, these are not strictly speaking distances, in particular, they do not respect the triangular inequality. The baseline metrics have been computed using functions from the Python package SciPy (`scipy.spatial.distance`): `euclidean`, `cityblock`, `cosine`, `correlation`.

Baseline metrics have been computed on the input data, after the data preprocessing detailed in the section above. Optional per-cell normalization and feature scaling have been also considered (Figure 2.1), but they resulted to be less performant (Supplementary Table A.1).

Figure 2.1: Workflow for metrics comparison.



The employed procedure, from the input preprocessed data to the performance evaluation is summarized for **(A)** baseline metrics and **(B)** OT, respectively. *Icons from flaticon.com*

2.2.4 Optimal Transport distance

The paragraphs below recall definitions presented in the introduction of this thesis (see Section 1.3.1) in the context of cell-cell distances, namely the Wasserstein distance between histograms (Equation 1.6) and its entropic regularization (Equation 1.7).

OT, as defined by Monge [Monge, 1781] and Kantorovich [Kantorovich, 1942], aims at finding a coupling between two probability distributions that minimizes transportation cost. The classical OT distance, also known as the Wasserstein distance, between two distributions $\mathbf{a} = (a_1 \dots a_m)$ and $\mathbf{b} = (b_1 \dots b_m)$ is defined as the minimal cost of transportation to morph $\mathbf{a}$ into $\mathbf{b}$. Given and discrete probability distributions, their OT distance is thus defined as

$$\mathrm{W}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbb{R}_+^{m \times m}} \langle \mathbf{P}, \mathbf{C} \rangle = \min_{\mathbf{P} \in \mathbb{R}_+^{m \times m}} \sum_{k,l} P_{k,l} C_{k,l} \tag{2.1}$$

such that the coupling $\mathbf{P}$ satisfies $\sum_l P_{k,l} = a_k$ and $\sum_k P_{k,l} = b_l$.

According to the coupling $\mathbf{P}$ the mass in the discrete probability distribution is thus moved from one bin to another one in order to transform $\mathbf{a}$ into $\mathbf{b}$. $\mathbf{C}$ is called *ground cost* and it encodes the penalty for moving a unit of mass from one bin to another one. Hence, should be chosen in such a way that similar bins k and l have a low cost $C_{k,l}$.

We propose the use of the OT distance to capture cell-cell similarity in different single-cell omics data. For simplicity, let us consider a scRNA-seq dataset $\mathbf{A}$; the same concepts apply to other single-cell omics. Given a pair of cells i and j, as done for baseline metrics, we consider their expression profiles, corresponding to the vectors $\mathbf{u} = \mathbf{A}[\cdot, i] = (u_1 \dots u_m)$ and $\mathbf{v} = \mathbf{A}[\cdot, j] = (v_1 \dots v_m)$. Given that Equation 2.1 is defined only for discrete probability distributions, we transform $\mathbf{u}$ and $\mathbf{v}$ into two discrete probability distributions $\frac{\mathbf{u}}{\|\mathbf{u}\|_1}$ and $\frac{\mathbf{v}}{\|\mathbf{v}\|_1}$. In the following, we refer to such a transformation as *per-cell normalization*. After transformation, we can apply Equation 2.1 and compute the OT distance $\mathrm{W}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$ which corresponds to evaluating the minimal cost required to transform the gene expression discrete probability distribution ($\mathbf{a}$) of the first cell into the gene expression discrete probability distribution ($\mathbf{b}$) of the second cell. Based on Equation 2.1, OT computes the distance between a pair of cells ($\mathbf{a}, \mathbf{b}$) by taking into account the joint gene expression activity present in the two cells and encouraging genes to exchange mass according to the coupling $\mathbf{P}$. This exchange of mass is expected to happen in between genes related to each other, such as genes involved in the same regulatory program. In consequence, we expect the OT similarity between cells to not be driven by specific genes, but by the overall activity of regulatory programs (see Supplementary Text A.2).

As discussed above, the OT distance is parametrized by the ground cost $\mathbf{C}$. The ground cost reflects the cost of moving a unit of gene expression from a gene to another gene. The choice of the ground cost plays a central role in the final performances of OT and, in our case, there is no straightforward choice. Also ground costs based on prior information (e.g. pathways) could be employed. However, dataset/tissue specific ground costs are expected to be able to be more performant. It is for this reason that for all couples of genes k and l in the single-cell matrix $\mathbf{A}$, we define

$$C_{k,l} = d(\mathbf{A}[k, \cdot], \mathbf{A}[l, \cdot]), \tag{2.2}$$

where d corresponds to a metric among Euclidean and Manhattan distance, Cosine similarity and Pearson correlation. For a comparison of the performances of these different ground costs, see Supplementary Table A.2. Of note, nonlinear distances have also been employed to compute the ground cost in applications of OT to single-cell [Bellazzi et al., 2021, Yang et al., 2020]. We here chose linear distances as they do not require to be adjusted based on the omics under analysis.

Given that the number of features (e.g. genes, peaks) of single-cell data is in the order of tens of thousands, the classical OT problem would be computationally intractable. Indeed, solving Equation 2.1 relies on costly linear programming methods [Peyré et al., 2019]. Hence, we considered the entropic regularization of the classical OT distance, also called Sinkhorn divergence [Cuturi, 2013, Genevay et al., 2019]. The entropic-regularized OT distance between two histograms $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$\mathrm{W}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbb{R}_+^{m \times m}} \langle \mathbf{P}, \mathbf{C} \rangle + \varepsilon \langle \mathbf{P}, \log \mathbf{P} - 1 \rangle \tag{2.3}$$

with $\mathbf{P}$ such that $\sum_l P_{k,l} = a_k$ and $\sum_k P_{k,l} = b_l$.

The first term of Equation 2.3 corresponds exactly to Equation 2.1 with coupling and ground cost.

The additional term corresponds to the entropic regularization. Therefore, if the regularization parameter ε is set to zero, Equation 2.3 corresponds exactly to Equation 2.1 and classical OT is obtained. Increasing values of ε correspond to a more diffused coupling. From a biological perspective, we expect the introduction of the entropic regularization to allow to control for the systematic noise due to the stochasticity of gene expression at single-cell level and for the presence of dropouts, as motivated by the tests reported in Supplementary Table A.3 and Supplementary Text 3. At the same time, the entropic regularization allows a faster execution of the algorithm thus opening to the possibility of analyzing single-cell datasets bigger than those that can be analyzed with the classical OT (Supplementary Table A.3 and 3). Finally, the entropic regularization, encouraging exchanges of mass in between features, allows OT to give more importance to the relationships between features (e.g. genes), which further motivates its application to complex data like single-cell data.

The parameter ε thus plays a central role in the final performances of Sinkhorn divergence and should be carefully chosen. The advantage of the formulation at Equation 2.3 is that $W^{\varepsilon}_{\mathbf{C}}$ can be efficiently computed on a GPU, thereby coping with the high dimensionality of single-cell data. Not only is entropy pivotal to scale the algorithm, but it is also important to break the curse of dimensionality which makes classical OT distance extremely hard to estimate from high-dimensional single-cell data. This phenomenon, analyzed theoretically in [Genevay et al., 2019] is supported by our analysis (Supplementary Table A.2).

An issue with Equation 2.3 for $\varepsilon > 0$ is that, given two cells having the same expression distribution ($\mathbf{a} = \mathbf{b}$), $W^{\varepsilon}_{\mathbf{C}} > 0$. We thus used the debiased Sinkhorn divergence [Feydy et al., 2019] in order to ensure a distance equal to zero for identical cells

$$\overline{W^{\varepsilon}_{\mathbf{C}}}(\mathbf{a}, \mathbf{b}) = W^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) - \tfrac{1}{2}(W^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{a}) + W^{\varepsilon}_{\mathbf{C}}(\mathbf{b}, \mathbf{b})). \tag{2.4}$$

For sake of simplicity in the rest of the paper, we will use the term *OT distance* to refer to the debiased Sinkhorn divergence.

As discussed above, the OT distance depends on two main parameters: the regularization parameter ε and the ground cost $\mathbf{C}$. For every dataset, we performed a grid search varying ε among $1000, 5, 1, 0.5, 0.1, 0.05, 0.01$ and the ground cost among Euclidean, Manhattan, Cosine and Pearson correlation. As shown in Supplementary Table A.2, the best performances on average across datasets and ground costs were obtained for the regularization parameter ε set at 0.5. We thus suggest 0.5 as default ε for future users. In contrast, the best performing ground cost $\mathbf{C}$ varied depending on the analyzed data. In scRNA-seq simulated data and in single-cell DNA methylation data, Pearson correlation achieved the best performances, while on scRNA-seq and scATAC-seq data, cosine similarity performed the best. The performances presented in Section 2.3 thus correspond to this choice of ε and $\mathbf{C}$. The computation of the OT distance has been implemented using the Python package PyTorch and run on a GPU. Computation times are listed in Supplementary Table A.4.

2.2.5 Performance evaluation

For all single-cell datasets, simulated and real, ground-truth labels are available. For the real single-cell omics, the ground-truth labels correspond to cell types, defined through clustering in the original publication, or to the cell lines from which the cells have been extracted.

We first use the C-index and Silhouette score to evaluate to which extent the various metrics detect

the similarity between cells associated with the same label, as well as the difference between cells with different labels (Figure 2.1). The C-index [Hubert and Schultz, 1976] is an internal clustering evaluation index. Given a cell-to-cell distance matrix, the C-index measures if the closest pairwise distances correspond to cells belonging to the same cluster. It is defined as

$$C_{\text{index}} = \frac{S - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}} \tag{2.5}$$

where, if n is the number of intracluster pairwise distances, S_{min} is the sum of the n smallest distances if all pairs of cells are considered, S_{max} is the sum of the n largest distances out of all pairs and S is the sum of distances over all pairs of cells from the same cluster.

Of note, the C-index is always in the interval $[0, 1]$ and it should be minimum in the case of a perfect clustering. To make the results easily readable, we consider $1 - C_{\text{index}}$, so that the best performances are obtained by maximizing the score. Concerning the implementation, we used our own Python implementation of the C-index.

As a complementary evaluation, we further considered the Silhouette score, defined as

$$S(\mathbf{u}) = \frac{E(\mathbf{u}) - e(\mathbf{u})}{\max(E(\mathbf{u}), e(\mathbf{u}))} \tag{2.6}$$

where $E(\mathbf{u})$ is the average distance between the cell $\mathbf{u}$ and the other cells of the same cluster and $e(\mathbf{u})$ is the average distance between $\mathbf{u}$ and cells in the closest different clusters. The distance used to compute $E(\mathbf{u})$ and $e(\mathbf{u})$ varies among the benchmarked metrics (OT, Euclidean, Manhattan, Cosine similarity and Pearson correlation). The global Silhouette score is then obtained by averaging $S(\mathbf{u})$ over all cells. We used the Silhouette score implementation of the Python package scikit-learn [Pedregosa et al., 2011].

To further test the quality of the inferred cell-to-cell distance matrices, we used them as inputs for a clustering algorithm and assessed the agreement between the inferred clusters and the ground-truth labels (Figure 2.1). We considered clustering algorithms for which applications to single-cell data have been already proposed and directly applicable to the cell-to-cell distance matrices. We thus selected hierarchical clustering, with complete linkage, and spectral clustering [Von Luxburg, 2007, Zheng et al., 2019], both implemented in scikit-learn. Regarding hierarchical clustering, we chose complete linkage in place of average linkage, used in other single-cell clustering works [Guo et al., 2015b], because this approach provided better performances for both baseline and OT distances (Supplementary Table A.5). Concerning spectral clustering, since it requires in input an affinity matrix, we converted the inferred distance matrices $\mathbf{D}$ to affinity matrices $1 - \mathbf{D}/\|\mathbf{D}\|_\infty$, and run the clustering algorithm with default parameters. Both clustering algorithms require the specification of the number of clusters in which the cells should be partitioned. In addition, we considered the typical single-cell clustering workflow [Luecken and Theis, 2019] composed of: *(i)* dimensionality reduction (DR) (PCA), *(ii)* kNN graph construction based on Euclidean distance and *(iii)* Leiden/Louvain clustering [Blondel et al., 2008, Traag et al., 2019] of the obtained graph using Scanpy default parameters [Wolf et al., 2018]. In case of OT, we only applied the two last steps of the workflow. Indeed, PCA relies on Euclidean distance and would thus negatively affect the results of OT. The considered clustering algorithms depend on different parameters. To apply spectral and hierarchical clustering, the user needs to set a desired number of clusters (k). On the opposite, the results of Leiden and Louvain depend on the resolution (r), which indirectly affects the number of clusters. We thus set these parameters (k, r) in an unsupervised way, by optimizing the Silhouette score [Rousseeuw, 1987]. For

spectral and hierarchical clustering, we varied the number of clusters (k between 3 and 25), while for Leiden and Louvain we varied the resolution parameter (r between 0.25 and 1.5) and chose the values maximizing the Silhouette score of the clustering. Thereby, we can test how frequently the number of ground-truth labels present in the data are captured by the different metrics. Of note, the overall behavior observed when the number of clusters is optimized is in good agreement with the results obtained by fixing the number of clusters to the number of ground-truth labels (Supplementary Table A.5). To evaluate the quality of the obtained clusters, we used Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI; Figure 2.1).

Given U, clustering inferred from a distance matrix and V, ground-truth labels, the ARI is defined as

$$\mathrm{ARI}(U,V) = \frac{\mathrm{RI}(U,V) - E(\mathrm{RI})}{\max(\mathrm{RI}) - E(\mathrm{RI})} \tag{2.7}$$

where $\mathrm{RI}(U,V)$ is the Rand index, i.e. the fraction of pairs of samples that are either in the same group or in different groups in both U and V, $E(\mathrm{RI})$ is the expected Rand index between U and a random V, and $\max(\mathrm{RI})$ is the largest possible Rand index between U and any V.

To consider a complementary score, we also used the NMI, defined as

$$\mathrm{NMI}(U,V) = \frac{2\mathrm{MI}(U,V)}{\mathrm{H}(U) + \mathrm{H}(V)} \tag{2.8}$$

where $\mathrm{MI}(U,V)$ is the mutual information between U and V i.e. $\mathrm{MI}(U,V) = \mathrm{H}(U) - \mathrm{H}(U|V)$ and $\mathrm{H}(\cdot)$ denotes entropy. To compute ARI and NMI, we used the corresponding scikit-learn implementations.

2.3 Results

Debiased entropic-regularized OT distance (see Section 2.2), henceforth called OT distance, is here proposed as a metric to infer cell-cell similarity across different single-cell omics data. The performances of the OT distance are then benchmarked with respect to state-of-the-art metrics, henceforth called baseline metrics, namely the Euclidean and Manhattan distances, Cosine similarity and Pearson correlation.

The benchmark is performed in three main contexts Table 2.1. First, simulated scRNA-seq data are considered. Data composed of different numbers of cells are generated to test whether the metrics scale also to high-dimensional data, as the currently available single-cell data, and if the number of cells impacts performances. Then the overlap and size of the clusters underlying the simulated scRNA-seq data are also varied to challenge the various metrics in detecting less clear and rare groupings of cells. Second, four real scRNA-seq data, profiled from CRC and cell lines, are considered. Finally, we further challenged the distance measures on other single-cell omics: scATAC-seq and single-cell DNA methylation. See Section 2.2 and Table 2.1 for details concerning the data.

In all three contexts, ground-truth labels were available for all cells. In simulated scRNA-seq data, the ground-truth labels have been imposed during the data simulation (see Section 2.2). In real single-cell omics, the ground-truth labels are extracted from the original publications. For the four profilings of cell lines (Liu scRNA, Li cell lines, Liu scATAC and Leukemia scATAC), labels correspond to the cell line of origin and thus reflect strong transcriptional/epigenetic differences. In contrast, in the

case of the neuronal single-cell DNA methylation (scMethylation mouse and scMethylation human) and scRNA-seq from CRC samples (Li Tumor and Li NM), the ground-truth labels reflect clusters previously identified in the data, based on the activity of predefined markers. These last applications are clearly more challenging, as much weaker differences exist between different cell types or states.

As summarized in Figure 2.1, we first tested with C-index and Silhouette score whether the various metrics are able to detect the similarity between cells belonging to the same group. We then applied different unsupervised clustering algorithms to the computed distance matrices and tested the quality of the resulting clusters using the ARI and NMI.

2.3.1 Comparing OT against baseline metrics based on cell-cell similarity detection

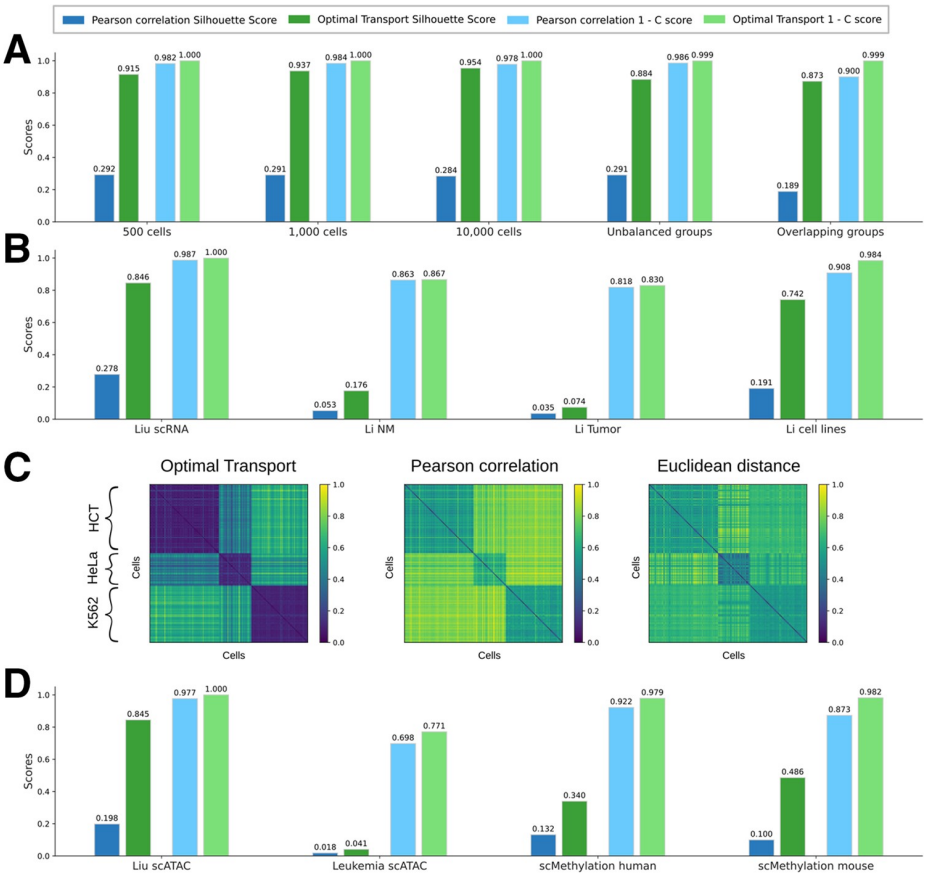
Figure 2.2 summarizes the results of the comparison between OT and baseline metrics. Pearson correlation outperformed the other baseline metrics on all data types. We thus used it as representative of baseline metrics in Figure 2.2 and in the following. The results of alternative baseline metrics are available in Supplementary Table A.1. Of note, better performances of Pearson correlation with respect to other state-of-the-art metrics had been previously observed [Kim et al., 2019].

In all simulated data, OT outperforms all baseline metrics (Figure 2.2A), both in terms of C-index and Silhouette score. In particular, the results of OT are not impacted by the number of cells, given that it shows a consistently performant behavior for 500, 1,000 and 10,000 cells. These results also suggest that OT is scalable to high-dimensional datasets, a crucial feature for the analysis of single-cell data. Finally, OT achieves superior performance with highly unbalanced clusters, which is more realistic as biological samples are often composed of a mixture of rare and common populations of cells, and also with overlapping clusters, which reflects the scenario of subpopulations of cells sharing similar transcriptional patterns, as cell populations tracked over different developmental phases.

We then considered four scRNA-seq datasets: two of them correspond to cancer cell lines [Li et al., 2017, Liu et al., 2019], while the remaining two correspond to colorectal tumor tissue and matched normal mucosa [Li et al., 2017]. All metrics tend to perform better in cancer cell lines than CRC samples. This result is most probably the consequence of the stronger transcriptional difference existing between cell lines. Overall, in all the four scRNA-seq datasets, OT outperformed baseline metrics (Figure 2.2B). The improvement provided by OT in cell lines is important, especially according to Silhouette score (+0.6 Silhouette score). To show to which extent C-index and Silhouette score reflect a clear clustering structure in the cell-to-cell distance matrices, we focused on the smallest dataset, Liu scRNA [Liu et al., 2019]. Figure 2.2C reports the cell-to-cell distance matrices obtained for this dataset with OT distance, Pearson correlation and Euclidean distance. Cells in the matrices are sorted based on their cell line of origin. OT powerfully detected the similarity between cells belonging to the same cancer cell line, showing three clear blocks of cells at distance close to zero. In contrast, the blocks corresponding to the three cell lines are less marked with Pearson correlation. Finally, with Euclidean distance, the values outside the three blocks tend to be less close to one, indicating a less clear separation between cells belonging to different cell lines.

Finally, we challenged the various metrics on other single-cell omics data (Figure 2.2D). Also in this case, OT shows better performance than Pearson correlation. Overall, OT performed better than existing metrics in the detection of cell-cell similarities in all considered datasets. See Supplementary Table A.1 for the performances of other state-of-the-art baseline metrics.

Figure 2.2: Comparison of OT against Pearson correlation in cell-cell similarity inference.



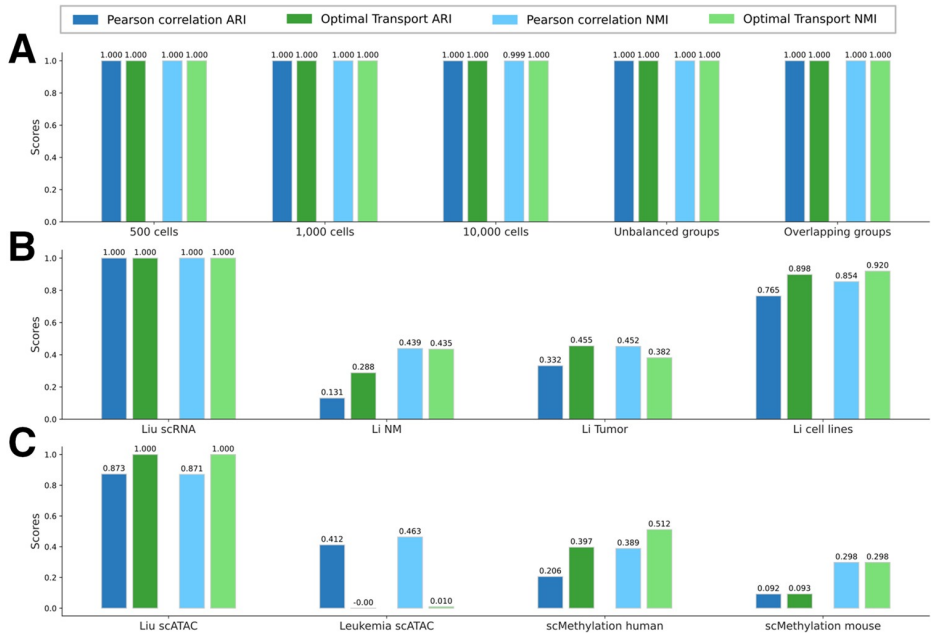
Barplots for C-index and Silhouette score are reported for **(A)** simulated scRNA-seq data composed of 500, 1,000 and 10,000 cells, with unbalanced groups and overlapping clusters; **(B)** four scRNA-seq datasets; **(D)** two single-cell DNA methylation and two scATAC-seq data. Examples of the distance matrices obtained with OT, Pearson correlation and Euclidean distance in Liu scRNA-seq are reported in **(C)**

2.3.2 Comparing OT against baseline metrics based on hierarchical clustering

Hierarchical clustering was applied to the cell-to-cell distance matrices computed with OT and Pearson correlation. In all datasets, the number of clusters was optimized based on the Silhouette score (Section 2.2). Table 2.1 summarizes the number of clusters obtained across all datasets and compares them with those defined in the original publications. The resulting clusters were then compared with the ground-truth labels based on ARI and NMI (Figure 2.3).

In simulated data, all the five datasets contain three underlying clusters. As described in Table 2.1,

Figure 2.3: Comparison of OT against Pearson correlation in hierarchical clustering.



Barplots for ARI and NMI are reported for **(A)** simulated scRNA-seq data composed of 500, 1,000 and 10,000 cells, with unbalanced groups and overlapping clusters; **(B)** four scRNA-seq datasets; **(C)** two single-cell DNA methylation and two scATAC-seq data

in four out of five cases, both OT and Pearson correlation identified the correct number of clusters. The only exception is represented by the dataset composed of 10,000 cells, where OT correctly detects the presence of three clusters, while Pearson correlation overclustered the data, by subdividing the three groups into five clusters. According to ARI and NMI, both Pearson correlation and OT give rise to perfect results (Figure 2.3A). Pearson correlation also led to optimal ARI and NMI scores in the dataset of 10,000 cells, as overclustering is not captured by these scores.

Turning to real scRNA-seq data, in Liu scRNA data, both Pearson correlation and OT detected the correct number of clusters Table 2.1. Concerning the two CRC datasets, for Li Tumor, containing seven cell types, OT only detected three clusters while Pearson predicted the presence of six clusters, thus missing only the rare population of mast cells, represented by one cell. In contrast, for Li NM, Pearson correlation dramatically overclustered the data and inferred 23 clusters, while OT suggested the presence of nine clusters. Finally, for Li cell lines, composed of seven cell lines, OT suggested the presence of eight clusters, while Pearson identified 10. Overall, in all scRNA-seq datasets, excepting [Liu et al., 2019], no distance captured exactly the numbers of clusters reported in the corresponding publications. However, the number of clusters inferred by OT tend to be closer to the ground-truth, while Pearson correlation tended to highly overcluster the data. Regarding the NMI and ARI scores, in two out of four datasets, OT performed better than Pearson correlation (Figure 2.3B), while for Liu

scRNA and Li Tumor both OT and Pearson achieved comparable performances.

In other single-cell omics data, OT still outperforms Pearson correlation in the majority of the datasets. Regarding the numbers of clusters, in Liu scATAC, both Pearson correlation and OT correctly detected the presence of three clusters. In Leukemia scATAC, composed of six cell lines, OT and Pearson correlation predicted 3 and 25 clusters, respectively. Finally, in methylation data, both OT and Pearson underestimated the real number of clusters. As shown in Figure 2.3C, according to ARI and NMI values, OT showed better performances in all datasets except Leukemia data. Pearson correlation performed well according to ARI and NMI for Leukemia data, but this performance is obtained considering 25 clusters instead of the five real clusters presumably present in the data. This is a practical demonstration of the interest of inferring the optimal number of clusters based on the Silhouette score, rather than fixing it to the value reported in their original publication.

The improvement provided by OT for cell-cell similarity inference (Figure 2.2) has an impact on clustering performances. Of note, these results are in agreement with those observed when the number of clusters is fixed (Supplementary Table A.5). Finally, similar results are obtained when substituting hierarchical clustering with spectral clustering (Supplementary Text A.4 and Supplementary Figure A.2), suggesting that the choice of clustering algorithm does not affect conclusions.

2.3.3 Comparing a typical single-cell clustering workflow against its counterpart based on OT

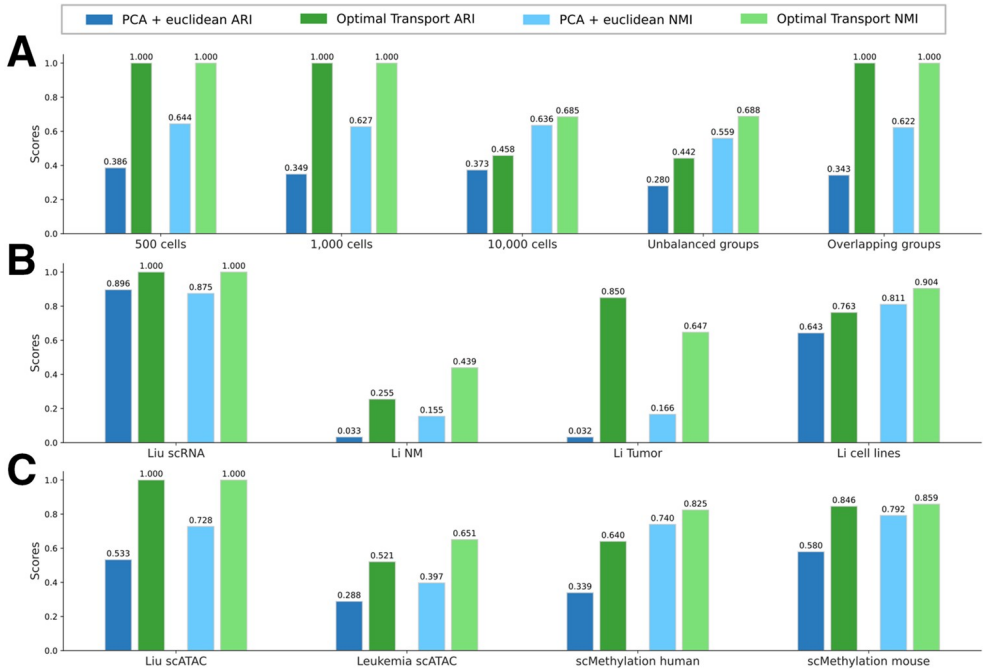
Clustering in single-cell is classically performed following a typical workflow composed of: *(i)* DR (PCA), *(ii)* kNN graph construction based on Euclidean distance and *(iii)* Leiden/Louvain clustering of the obtained graph [Luecken and Theis, 2019]. We here compare the results obtained with this typical single-cell clustering workflow with respect to its counterpart based on OT (see Section 2.2 for details). Figure 2.4 and Table 2.1 report the results obtained once the resolution parameter for the Leiden clustering is optimized based on the Silhouette score (Section 2.2). Of note, the results are not affected by the choice of the resolution parameter (Supplementary Figure A.3).

In simulated data, the estimation of the number of clusters is less precise than with hierarchical and spectral clustering. Overall, the typical clustering workflow tends to more frequently overcluster the data Table 2.1. Regarding ARI and NMI scores, OT shows better performances, reaching perfect score for three out of five datasets (Figure 2.4A).

For scRNA-seq data, both the typical clustering workflow and OT correctly predicted the correct amount of clusters present in Liu scRNA. The two CRC datasets contain seven cell types, corresponding to our ground-truth labels. For Li Tumor, both the typical clustering workflow and OT predicted the presence of two clusters. In contrast, for Li NM, the typical clustering workflow inferred four clusters, while OT suggested the presence of six clusters. Finally, for Li cell lines, OT suggested the presence of nine clusters, while the typical clustering workflow correctly identified seven clusters. Concerning the NMI and ARI scores, OT outperformed the typical clustering workflow in all four scRNA-seq datasets (Figure 2.4B).

In other single-cell omics, OT correctly identified three clusters for Liu scATAC data, while the typical clustering workflow found seven clusters. Regarding Leukemia scATAC data, involving six cell lines, both OT and the typical clustering workflow under-clustered the data. Finally, in methylation data OT tend to be closer to the correct number of clusters. Regarding ARI and NMI (Figure 2.4C), OT always outperforms the typical clustering workflow and obtains perfect performances in Liu ATAC.

Figure 2.4: Comparison of a typical single-cell clustering workflow against its counterpart based on OT.



Barplots for ARI and NMI are reported for **(A)** simulated scRNA-seq data composed of 500, 1,000 and 10,000 cells, with unbalanced groups and overlapping clusters; **(B)** four scRNA-seq datasets; **(C)** two single-cell DNA methylation and two scATAC-seq data

Of note, similar results are obtained when substituting the Leiden clustering with the Louvain clustering (Supplementary Text A.5 and Supplementary Figure A.4), indicating that the choice of clustering algorithm should not affect conclusions.

2.3.4 Open-source implementation and distribution

To foster the reproducibility of all the results presented in this work, we provide a Python package and Jupyter notebook covering all the analyses performed, both available at https://github.com/cantinilab/ot-scOmics, together with all the preprocessed single-cell data used in this study. Since computing OT distances is computationally intensive, the code is designed to be run on a GPU, taking advantage of the PyTorch library. For users who do not have access to a GPU, extensive explanations are provided to run the Jupyter notebooks on the Google Collaboratory platform (http://colab.research.google.com/). Note that our code also allows for CPU computations.

2.4 Discussion

In this study, we assessed the potential of OT to infer cell-cell similarities from single-cell omics data. We extensively benchmarked OT performances against state-of-the-art metrics. Interestingly, OT outperformed alternative metrics in capturing cell-cell similarities in all the 13 considered datasets. The biological relevance of this improvement was assessed by performing cell clustering. In all cases, the use of OT distance resulted in improved clustering results. Of note, different clustering algorithms have been used to test whether the observed performances were affected by the choice of the algorithm. Finally, we further challenged the metrics to detect cell-cell similarity in other single-cell omics: scATAC-seq and single-cell DNA methylation data. The improvement provided by OT is conserved also when other single-cell omics are considered, despite the high sparsity and the close-to-binary nature of scATAC-seq data [PE de Souza et al., 2020, Xiong et al., 2019].

This work presents an empirical choice of ground cost for each considered modality, but an interesting perspective would be to unify the approach across omics. For instance, one could use probabilistic models tailored to each single-cell modality to define a cost between features. Alternatively, one may base the ground cost on deviations from a "null model." Recent methods have defined a null model for scRNA-seq counts and treated Pearson residuals as the biological signal [Lause et al., 2021]. In this case, the ground cost could be a distance between Pearson residuals, and this approach could be generalized to other omics.

Single-cell datasets composed of a number of cells higher than that here employed are currently available (e.g. single-cell atlases). The scope of our analysis is to test the performances of OT and baseline measures in inferring cell-cell similarity from single-cell data. In this context, we are interested in reconstructing the distance matrices comparing all possible couples of cells present in the dataset. Such computations are quadratic in the number of cells, both in terms of memory and time complexity, thus preventing us to test performances on single-cell atlases for both baseline and OT distances.

An important addition to the current benchmark would be to compare OT-based clustering with other nonlinear approaches, such as kernel-based clustering with a Gaussian kernel, and to report both the computational load and performances. Indeed, this work compares different values of ε, including the Maximum Mean Discrepancy limit (for large ε), but it would be interesting to evaluate the impact of OT compared to less computationally intensive nonlinear methods.

The benchmark could also be extended with more complex simulated settings designed to better understand limitations of the proposed method. For instance, this work did not explicitly test performances for different numbers of features, and it would be interesting to understand if the method is robust to increased dimensionality. In addition, since high numbers of features induce a higher computational load, this would allow to relate performance and computation times as a function of the number of features. Other interesting aspects to evaluate in simulated settings would be the level of correlation between the features, sparsity levels, and number of clusters. This may help to understand why the best-performing ground costs differ for real and simulated scRNA-seq, and across modalities.

Our evaluation of metrics performances is dependent on the ground-truth labels associated with the input datasets. In cell lines, the ground-truth is well established and wide transcriptional differences exist between different cell lines. Biological samples instead consist of cell types and states having less pronounced transcriptional differences. What is a ground-truth in this case is thus less

clear. We here used as ground-truth the clusters identified in the original publication of each dataset. However, such labels could be improved. In addition, given that the original publications employ Euclidean distance or Pearson correlation to define the labels, the usage of such labels as ground-truth is expected to advantage these metrics compared to alternative ones.

Of note, DR is frequently applied to reduce the noise in single-cell data before cell clustering and further downstream analyses. Here, we only applied DR in the context of Leiden/Louvain clustering for Euclidean distance, but not for the OT distance. Indeed, the most popular DR approaches, such as PCA, diffusion maps, NMF and ICA, rely on Euclidean geometry and would thus not be a good choice prior to OT computation. The good performances provided by OT in this work suggest that further efforts should be devoted to design DR methods for single-cell data based on other metrics.

Finally, we applied OT to different single-cell omics in isolation. However, different omics data presumably provide complementary information on individual cellular states. Combining different single-cell omics with appropriate metrics thus represent a critical challenge in computational biology. Our results suggest that OT could be a valuable metric for the integration of multiomics data.

Data availability Data from [Liu et al., 2019] can be accessed through the supplementary data 3 and 4 of the original article or at the link github.com/cantinilab/momixnotebook/tree/master/data/single-cell. Data from [Li et al., 2017, Corces et al., 2016, Luo et al., 2017] can be accessed through the Gene Expression Omnibus (GEO) accession numbers: GSE81861, GSE65360 and GSE97179 respectively. The code to reproduce the analyses is available at github.com/cantinilab/OT-scOmics.

Table 2.1: Summary of the datasets used for our benchmark and their clustering results

Dataset name	Data type	Reference	# of cells	# of features after preprocessing	# of clusters in ground-truth	Detected # of clusters for hierarchical clustering (OT)	Detected # of clusters for hierarchical clustering (Pearson)	Detected # of clusters for Leiden clustering (OT)	Detected # of clusters for Leiden clustering (PCA + Euclidean)
500 cells	Simulated scRNA-seq data	Splatter	500	5,000	3	3	3	3	11
1,000 cells	Simulated scRNA-seq data	Splatter	1,000	5000	3	3	3	3	12
10,000 cells	Simulated scRNA-seq data	Splatter	10,000	5000	3	3	5	9	12
Unbalanced clusters	Simulated scRNA-seq data	Splatter	1,000	5000	3	3	3	5	8
Overlapping clusters	Simulated scRNA-seq data	Splatter	1,000	5000	3	3	3	3	12
Liu scRNA	scRNA-seq	[Liu et al., 2019]	206	10,000	3	3	3	3	3
Li Tumor	scRNA-seq	[Li et al., 2017]	364	10,000	7	3	6	2	2
Li NM	scRNA-seq	[Li et al., 2017]	266	10,000	7	9	23	6	4
Li cell lines	scRNA-seq	[Li et al., 2017]	561	10,000	7	8	10	9	7
Liu scATAC	scATAC-seq	[Liu et al., 2019]	206	10,000	3	3	3	3	7
Leukemia scATAC	scATAC-seq	[Corces et al., 2016]	391	7602	6	3	25	4	3
scMethylation mouse	Single-cell DNA methylation	[Luo et al., 2017]	3377	10,000	16	3	3	14	18
scMethylation human	Single-cell DNA methylation	[Luo et al., 2017]	2740	10,000	21	3	7	16	32

Chapter 3

Unsupervised ground metric learning using Wasserstein Singular Vectors

Abstract

Defining meaningful distances between samples in a dataset is a fundamental problem in machine learning. Optimal Transport (OT) lifts a distance between *features* (the "ground metric") to a geometrically meaningful distance between *samples*.

However, there is usually no straightforward choice of ground metric. Supervised ground metric learning approaches exist but require labeled data. In absence of labels, only ad-hoc ground metrics remain. Unsupervised ground metric learning is thus a fundamental problem to enable data-driven applications of OT.

In this paper, we propose for the first time a canonical answer by simultaneously computing an OT distance between *samples* and between *features* of a dataset. These distance matrices emerge naturally as positive singular vectors of the function mapping ground metrics to OT distances. We provide criteria to ensure the existence and uniqueness of these singular vectors. We then introduce scalable computational methods to approximate them in high-dimensional settings, using stochastic approximation and entropic regularization. Finally, we showcase *Wasserstein Singular Vectors* on a single-cell RNA-sequencing dataset.

This chapter was published as

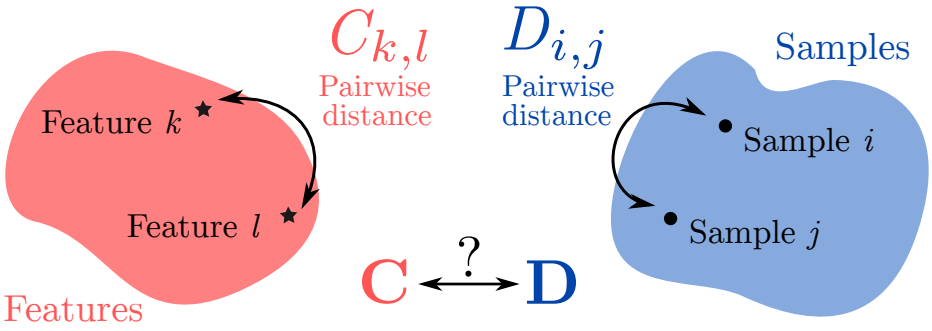
Geert-Jan Huizing, Laura Cantini, Gabriel Peyré, *Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors*, Proceedings of the 39th International Conference on Machine Learning, PMLR, 2022.
https://proceedings.mlr.press/v162/huizing22a.html

Contents

3.1 Introduction 50
3.1.1 Previous works 52
3.1.2 Contributions 53
3.2 Unsupervised Wasserstein metric learning 53
3.2.1 Wasserstein Singular Vectors 54
3.2.2 Theoretical properties 54
3.2.3 Numerical illustration on translated histograms 56
3.3 Large scale stochastic power iterations 58
3.4 Parallelization with entropic regularisation 59
3.4.1 Sinkhorn divergence map 60
3.4.2 Connection with PCA when $\varepsilon \to \infty$ 61
3.5 Metric learning for single-cell genomics 61
3.6 Discussion 63

3.1 Introduction

Figure 3.1: Unsupervised ground metric learning



How to jointly define a distance **C** between features and a distance **D** between samples? Wasserstein Singular Vectors define natural Wasserstein distances (**C**, **D**) in an unsupervised manner.

Machine learning tasks like information retrieval and classification require a notion of distance between samples in a dataset $\mathbf{U} \in \mathbb{R}^{m \times n}$. In particular, we will study the case of single-cell RNA-sequencing data (scRNA-seq).

Discrete histograms In scRNA-seq data, the sample $\mathbf{u}_i \in \mathbb{R}_+^m$ represents the expression value of all genes in the i-th cell. This motivates two key assumptions in this paper: (i) samples are positive,

which is natural when quantifying presence and quantity of physical objects (ii) samples can be normalized to discrete histograms, which is natural when the distribution over features (e.g. genes) is more important than the total mass. Indeed, gene expression in scRNA-seq data is usually normalized in some way as part of preprocessing [Luecken and Theis, 2019].

Optimal Transport distances Optimal Transport (OT) [Monge, 1781, Kantorovich, 1942] offers a geometrically meaningful distance between discrete probability distributions, and has recently emerged as a useful tool for machine learning applications [Frogner et al., 2015, Rolet et al., 2016] (see Section 1.3.1). Contrarily to the Euclidean distance, OT does not compare distributions bin by bin. Instead, OT lifts a ground pairwise distance matrix $\mathbf{C} \in \mathbb{R}_+^{m\times m}$ between the m features to the "Wasserstein" OT distance between normalized samples $\mathbf{a}_i \overset{\text{def.}}{=} \mathbf{u}_i/\|\mathbf{u}_i\|_1 \in \Sigma_m \overset{\text{def.}}{=} \{\mathbf{a} \in \mathbb{R}_+^m \text{ s.t. } \sum_k a_k = 1\}$. It optimizes a transport plan $\mathbf{P} \in \mathbb{R}^{m\times m}$ encoding the displacement of mass between the two m-dimensional histograms $\mathbf{a}_i, \mathbf{a}_j$. As defined in Section 1.3.1, the Wasserstein distance reads

$$\mathrm{W}_{\mathbf{C}}(\mathbf{a}_i, \mathbf{a}_j) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathbb{R}_+^{m\times m}} \sum_{k,l} P_{k,l}C_{k,l} \text{ s.t. } \begin{cases} \mathbf{P}\mathbf{1}_m = \mathbf{a}_i, \\ \mathbf{P}^\top \mathbf{1}_m = \mathbf{a}_j. \end{cases}$$

From supervised to unsupervised ground metric learning The crucial aspect of successful application of OT in machine learning is the design of a metric $\mathbf{C}$ which encodes the geometric relationships between features. In a supervised setting, one might take advantage of so-called *ground metric learning* methods [Cuturi and Avis, 2014]. In an unsupervised setting, one usually resorts to some ad-hoc choice of ground cost. For instance the Word Mover Distance [Kusner et al., 2015] uses Euclidean distances on Word2Vec embeddings [Mikolov et al., 2013]. Similarly, the Gene Mover Distance [Bellazzi et al., 2021] uses Euclidean distances on Gene2Vec embeddings [Du et al., 2019] (see Section 1.3.1).

In this work, we take a radically different route, by requiring that $C_{k,l}$ is itself a Wasserstein distance between histograms $\mathbf{b}_k$ and $\mathbf{b}_l \in \Sigma_n$. The most intuitive case is to consider $\mathbf{a}_i$ (resp. $\mathbf{b}_k$) to be the normalized column i (resp. row k) of a dataset $\mathbf{U} \in \mathbb{R}_+^{m\times n}$. We will thus refer to $\mathbf{a}_i$ as a *sample* and $\mathbf{b}_k$ as a *feature*.

An intuitive way to motivate our method is to consider a bootstrapping approach. Given some initial (for instance random) metric $\mathbf{D}$, one can compute $C_{k,l} = \mathrm{W}_{\mathbf{D}}(\mathbf{b}_k, \mathbf{b}_l)$. But there is no reason to stop here, and the metric $\mathbf{D}$ can be "improved" by updating $D_{i,j} = \mathrm{W}_{\mathbf{C}}(\mathbf{a}_i, \mathbf{a}_j)$. By continuing this process of successively updating $\mathbf{C}$ and $\mathbf{D}$, one could hope to reach a limit where the pair of distance matrices $(\mathbf{C}, \mathbf{D})$ satisfies the following fixed point equation

$$\underbrace{C_{k,l}}_{\text{Distance between features}} = \tfrac{1}{\lambda}\mathrm{W}_{\mathbf{D}}(\underbrace{\mathbf{b}_k, \mathbf{b}_l}_{\text{Features}}), \quad \underbrace{D_{i,j}}_{\text{Distance between samples}} = \tfrac{1}{\gamma}\mathrm{W}_{\mathbf{C}}(\underbrace{\mathbf{a}_i, \mathbf{a}_j}_{\text{Samples}}), \tag{3.1}$$

where $(\lambda, \gamma) \in \mathbb{R}_+^2$ are scaling factors. This corresponds to casting ground metric learning as a nonlinear singular vectors problem. In this work, we study theoretical properties (including existence and uniqueness) as well as practical relevance of Wasserstein Singular Vectors for machine learning.

The idea of jointly analyzing samples and features bears resemblance to Principal Component Analysis (PCA). Indeed, the *nonlinear* singular vector problem outlined in Equation 3.1 mirrors a *linear*

singular vector problem: $\mathbf{Au} = \lambda\mathbf{v}, \mathbf{A}^\top\mathbf{v} = \lambda\mathbf{u}$. Section 3.4 provides a theoretical link between PCA and Wasserstein Singular Vectors with entropic regularization.

3.1.1 Previous works

The paragraphs below give a summary background on OT and metric learning, which are covered in more detail in Section 1.3.1.

Optimal Transport While the initial proposal of Monge [Monge, 1781] formulates the OT problem as an optimal matching problem, its modern and tractable formulation by Kantorovich [Kantorovich, 1942] is a linear program detailed in the Introduction. Besides its use to define matchings and couplings between distributions, the main feature of OT is that the transportation value induces a geometric distance on the space of probability distributions. This "Wasserstein" distance is thus parameterized by the underlying ground cost between pairs of points. We refer to the monographs [Villani, 2003, Santambrogio, 2015] for a detailed account of the theory of OT, and [Peyré et al., 2019] for its computational aspects. OT distances have been used for applications as diverse as image retrieval [Rubner et al., 2000], brain imaging [Gramfort et al., 2015, Janati et al., 2020], natural language processing [Kusner et al., 2015, Yurochkin et al., 2019], and generative models [Arjovsky et al., 2017, Tolstikhin et al., 2017]. In recent years, many applications of OT to single-cell biology have been proposed [Hashimoto et al., 2016, Schiebinger et al., 2019, Bellazzi et al., 2021, Huizing et al., 2022, Tong et al., 2022].

Entropic regularization Entropic regularization of OT allows to scale to high-dimensional machine learning problems. It approximates OT distances using Sinkhorn's algorithm, which has quadratic complexity and streams well on GPU architectures. Entropic regularization was put forward in the seminal paper by Cuturi [Cuturi, 2013], who also emphasizes the smoothing effect, which is crucial when using Sinkhorn as a loss function to train deep learning models. Another benefit of this regularization is that it suffers less from the curse of dimensionality, as proved in [Genevay et al., 2019, Mena and Weed, 2019]. This approach is also pivotal to scale our unsupervised metric learning method to tackle high-dimensional problems, for instance in genomics.

Metric Learning Metric learning is most often framed as the supervised problem of minimizing (resp. maximizing) the distance between points in a same (resp. different) class. Existing approaches are reviewed in [Kulis et al., 2012, Bellet et al., 2013]. It is necessary to restrict the class of distances to make the problem tractable. The most common option is arguably to consider the class of Mahalanobis distances, which generalize the Euclidean distance and are equivalent to computing a vectorial embedding of the data points. See for instance [Xing et al., 2002, Weinberger et al., 2006, Davis and Dhillon, 2008]. One can apply these methods for histogram data, or use instead of Euclidean distances more adapted discrepancies on the simplex, such as Chi-squared [Noh, 2012, Yang et al., 2015] and geodesic distances [Le and Cuturi, 2015]. These methods however fail to capture the geometric nature of the problem, where histograms correspond to discrete distributions viewed as sums of localized Dirac masses.

OT Ground Metric Learning This geometry is leveraged in [Cuturi and Avis, 2014] by introducing the problem of supervised OT ground metric learning and developing a nearest-neighbor based algorithm to solve it. This approach is further refined in [Wang and Guibas, 2012], which drops the triangular inequality constraint (as we do in our approach). It is possible to restrict the class of ground metrics, for instance using Mahalanobis [Xu et al., 2018, Kerdoncuff et al., 2021] or geodesic

distances [Heitz et al., 2020] to develop more efficient learning schemes. [Zen et al., 2014] simultaneously perform ground metric learning and matrix factorization, and this finds applications to NLP [Huang et al., 2016]. Metric learning can also be performed through adversarial optimization, where the metric is maximized over to perform generative model training [Genevay et al., 2018], discriminant analysis [Flamary et al., 2018] and to define robust transportation distances [Paty and Cuturi, 2019, Niles-Weed and Rigollet, 2019]. Note that when imposing only convex constraints, adversarial ground metric learning is a concave maximization problem which finds applications in the modeling of crowd congestion phenomena [Benmansour et al., 2010]. Another related question is the inverse problem of estimating a ground cost from the observation of matchings or couplings [Galichon and Salanié, 2020, Stuart and Wolfram, 2020, Li et al., 2019, Paty and Cuturi, 2020]. This supervised metric learning problem can be regularized using sparsity or low-rank constraints, as explained in [Dupuy et al., 2019, Carlier et al., 2023]. Finally, "hierarchical OT" [Yurochkin et al., 2019, Abrishami et al., 2020] uses OT to define the ground cost of a matching problem, using an intermediate level of meta-features.

3.1.2 Contributions

Our main contribution is the introduction in Section 3.2 of *Wasserstein singular vectors* as the positive singular vectors of monotone homogeneous "distance maps". The associated theoretical contributions, Theorem 3.2.3 and Theorem 3.2.5, state conditions ensuring the existence and uniqueness of such singular vectors.

Our second set of contributions allows to scale the method to large datasets. We first introduce in Section 3.3 a stochastic algorithm similar in spirit to Projected Stochastic Gradient Descent. Theorem 3.3.1 guarantees a convergence rate of $\mathcal{O}(\log t/\sqrt{t})$ under certain conditions. We then explain in Section 3.4 how to scale and parallelize this method even further by leveraging entropic regularization through the Sinkhorn algorithm. Proposition 3.4.4 shows that in the large regularization limit, our method computes metrics associated to 1-D and 2-D embeddings along the leading principal component axes.

Section 3.5 demonstrates the potential of Wasserstein Singular Vectors compared to ad-hoc applications of Optimal Transport, by studying a single-cell RNA-sequencing dataset.

A Python package implementing all algorithms in this paper is available at github.com/gjhuizing/wsingular. Optimal Transport distances were computed using the open-source POT library [Flamary et al., 2021]. Appendix B.1 lists the experiments' computation times and resources.

Notations

We denote $\mathcal{D}_m \subset \mathbb{R}_+^{m\times m}$ the set of pairwise distance matrices. In other words, $\mathbf{C} \in \mathcal{D}_m$ if (i) $C_{k,l} = 0 \iff k = l$, (ii) $C_{k,l} \leq C_{k,s} + C_{s,l}$ (iii) $C_{k,l} = C_{l,k}$. Its closure $\overline{\mathcal{D}_m} \overset{\text{def.}}{=} \{\mathbf{C} \in \mathbb{R}_+^{m\times m} \text{ s.t. } \mathbf{C} = \mathbf{C}^\top, \operatorname{diag}(\mathbf{C}) = 0\}$ is a set of pseudo-distances.

3.2 Unsupervised Wasserstein metric learning

This section introduces the singular vectors of the Wasserstein distance map. Fortunately, this nonlinear singular vector problem enjoys many desirable properties.

3.2.1 Wasserstein Singular Vectors

Wasserstein distance map The following map lifts a ground metric $\mathbf{C} \in \mathcal{D}_m$ to a pairwise distance matrix $\Phi_{\mathbf{A}}(\mathbf{C}) \in \mathcal{D}_n$. The map $\mathbf{D} \in \mathcal{D}_n \mapsto \Phi_{\mathbf{B}}(\mathbf{D}) \in \mathcal{D}_m$ is defined similarly.

$$\Phi_{\mathbf{A}}(\mathbf{C})_{i,j} \overset{\text{def.}}{=} \mathrm{W}_{\mathbf{C}}(\mathbf{a}_i, \mathbf{a}_j) + \tau \|\mathbf{C}\|_\infty R(\mathbf{a}_i - \mathbf{a}_j) \tag{3.2}$$

Role of regularization The norm R operates as a regularizer to enforce strict positivity of the computed distances by padding the matrix outside of the diagonal, and the $\|\mathbf{C}\|_\infty$ scaling ensures homogeneity of the map (see Proposition 3.2.2). In numerical examples below, we take $R = \|\cdot\|_1$. One could consider any other norm, or even a matrix with zeros on the diagonal and ones elsewhere. In practice, our method can be applied in the unregularized setting $\tau = 0$, but some of the theoretical claims require $\tau > 0$. In addition, tuning the hyperparameter τ may be beneficial in practice, since it interpolates between Wasserstein and L1 distances.

Wasserstein singular vectors With this notation, our ground metric learning solves for a pair $\mathbf{C} \in \mathcal{D}_m$ and $\mathbf{D} \in \mathcal{D}_n$ of *Wasserstein singular vectors* satisfying

$$\exists (\lambda, \gamma) \in (\mathbb{R}_+^*)^2 \text{ s.t. } \Phi_{\mathbf{B}}(\mathbf{D}) = \lambda \mathbf{C},\ \Phi_{\mathbf{A}}(\mathbf{C}) = \gamma \mathbf{D}, \tag{3.3}$$

which corresponds to Equation 3.1 when $\tau = 0$. The case $m = n$ and $\mathbf{A} = \mathbf{B}$ corresponds to the computation of an eigenvector $\mathbf{C} = \mathbf{D}$ of $\Phi_{\mathbf{A}}$ with eigenvalue $\lambda = \gamma$.

Power iterations algorithm The de-facto standard algorithm to extract singular vectors are "power iterations"

$$\mathbf{C}_{t+1} \overset{\text{def.}}{=} \frac{\Phi_{\mathbf{B}}(\mathbf{D}_t)}{\|\Phi_{\mathbf{B}}(\mathbf{D}_t)\|_\infty}, \quad \mathbf{D}_{t+1} \overset{\text{def.}}{=} \frac{\Phi_{\mathbf{A}}(\mathbf{C}_{t+1})}{\|\Phi_{\mathbf{A}}(\mathbf{C}_{t+1})\|_\infty}. \tag{3.4}$$

The complexity of performing a single power iteration is $\mathcal{O}(n^2 m^2 (n \log(n) + m \log(m)))$, since the computation of a single Wasserstein distance in $\mathbb{R}_+^n$ is $\mathcal{O}(n^3 \log n)$ [Bonneel et al., 2011]. To cope with large scale datasets, we propose in Section 3.3 and 3.4 to use stochastic optimization and entropic regularization.

Remark 3.2.1. *A remarkable property of this algorithm is that, in cases where the singular vector is unique (which is observed in practice and proved below for large τ), even if the initialization $\mathbf{D}_{t=0}$ is chosen arbitrarily, it converges toward a distance matrix (so in particular it satisfies the triangular inequality at convergence).*

3.2.2 Theoretical properties

Properties of the Wasserstein distance map Non-linear singular vectors problems are notoriously difficult to study. Fortunately, as explained in the Proposition 3.2.2, the Wasserstein distance map is a so-called topical mapping (i.e. positive and monotone) [Lemmens and Nussbaum, 2012]. These mappings can be thought as non-linear generalizations of Markov chains. Problem Equation 3.3 is thus an instance of a non-linear Perron-Frobenius problem, for which, contrarily to generic problems, existence and uniqueness of positive solutions is in general rather the rule than the exception. This explains in large part the practical success of our approach.

The following proposition lists some useful properties of the Wasserstein distance map $\Phi_{\mathbf{A}}$.

Proposition 3.2.2. *(i) $\Phi_{\mathbf{A}}$ is positively 1-homogeneous and monotone. (ii) $\Phi_{\mathbf{A}}$ is continuous on $\mathcal{D}_m$ (iii) $\Phi_{\mathbf{A}}$ is $(1+2\tau k_R)$-Lipschitz on $\mathcal{D}_m$ with regard to the $\|\cdot\|_\infty$ norm, where the constant $k_R > 0$ is such that $R(\cdot) \leq k_R\|\cdot\|_1$.*

Proof. (i) 1-homogeneity and monotony of $\Phi_{\mathbf{A}}$ follows from the definition. (ii) Note that $\Phi_{\mathbf{A}}$ is a vector-valued concave function (each coordinate being an infimum of linear forms) and hence is continuous on $\mathbb{R}^{m\times m}$. Actually, as we now show, it is Lipschitz for ℓ^∞. (iii) Let us prove that $\Phi_{\mathbf{A}}$ is Lipschitz continuous for the ℓ^∞ norm on $\mathcal{D}_m$. Firstly, since $R(\mathbf{a}_i - \mathbf{a}_j) \leq k_R\|\mathbf{a}_i - \mathbf{a}_j\|_1$ and $|\|\mathbf{C}\|_\infty - \|\mathbf{C}'\|_\infty| \leq \|\mathbf{C}-\mathbf{C}'\|_\infty$, we have

$$|\Phi_{\mathbf{A}}(\mathbf{C})_{i,j} - \Phi_{\mathbf{A}}(\mathbf{C}')_{i,j}| \leq |\mathrm{W}_{\mathbf{C}}(\mathbf{a}_i,\mathbf{a}_j) - \mathrm{W}_{\mathbf{C}'}(\mathbf{a}_i,\mathbf{a}_j)| + 2\tau k_R\|\mathbf{C}-\mathbf{C}'\|_\infty \tag{3.5}$$

Secondly, with $\mathcal{U}(\mathbf{a},\mathbf{a}')$ the set of valid couplings,

$$|\mathrm{W}_{\mathbf{C}}(\mathbf{a},\mathbf{a}') - \mathrm{W}_{\mathbf{C}'}(\mathbf{a},\mathbf{a}')| = |\min_{\mathbf{P}\in\mathcal{U}(\mathbf{a},\mathbf{a}')}\langle\mathbf{P},\mathbf{C}\rangle - \min_{\mathbf{P}'\in\mathcal{U}(\mathbf{a},\mathbf{a}')}\langle\mathbf{P}',\mathbf{C}'\rangle| \leq \|\mathbf{C}-\mathbf{C}'\|_\infty. \tag{3.6}$$

Indeed, $|\min(u) - \min(v)| \leq \max|u-v|$ and $\|\mathbf{P}\|_1 = 1$. So $\Phi_{\mathbf{A}}$ is $(1+2\tau k_R)$-Lipschitz. □

Existence of singular vectors The following proposition ensures the existence of positive (i.e. true distances) singular vectors. Its proof can be found in Appendix B.2. Note that the ℓ^∞ bound of Proposition 3.2.2 implies that all singular values are smaller than $1+2\tau k_R$.

Theorem 3.2.3. *When $\tau > 0$, there exist positive singular vectors $(\mathbf{C},\mathbf{D}) \in \mathcal{D}_m \times \mathcal{D}_n$ solving the problem Equation 3.3.*

Existence in the case $\tau = 0$ Extending Theorem 3.2.3 to the unregularized case is an open problem, and is out-of-reach using classical non-linear Perron-Frobenius theorems such as [Akian et al., 2016, Akian et al., 2018], which do not apply. The following remark exhibits solutions for a special case of the unregularized problem.

Remark 3.2.4 (Block-diagonal matrices)**.** *Let us consider the case $\tau = 0$ and a dataset $\mathbf{U} = \mathrm{diag}(\mathbf{U}_p)$, a block-diagonal matrix where $\mathbf{U}_p \in \mathbb{R}_+^{m_p\times n_p}$. Let $\mathbf{A}$ and $\mathbf{B}$ be its normalizations along columns and rows respectively. Then all matrices $(\mathbf{C},\mathbf{D}) \in \overline{\mathcal{D}_m}\times\overline{\mathcal{D}_n}$ of the form $\mathbf{C} = (c_{p,q}\mathbf{1}_{m_p\times m_q})_{p,q}$ and $\mathbf{D} = (c_{p,q}\mathbf{1}_{n_p\times n_q})_{p,q}$ for the same $c_{p\neq q} \in \mathbb{R}_+^*$ and $c_{p,p} = 0$ are dominant singular vectors with associated singular value 1. Indeed, the optimal transport plans for these ground costs also follow this block structure.*

Uniqueness of singular vectors If the dataset is too sparse and $\tau = 0$, one cannot hope to have uniqueness of the leading singular vectors. In fact, the previous remark exhibits an infinity of singular vectors when $\mathbf{U}$ is block-diagonal. It does not seem obvious to guarantee uniqueness by an a priori condition depending only on $(\mathbf{A},\mathbf{B})$. The following proposition gives an a posteriori way to check the uniqueness of singular vectors inside $\mathcal{D}_m \times \mathcal{D}_n$. In the numerical simulations of Section 3.2.3, we checked *a posteriori* that the computed singular vectors are indeed unique.

Theorem 3.2.5. *Let $(\mathbf{C},\mathbf{D}) \in \mathcal{D}_m \times \mathcal{D}_n$ a pair of Wasserstein Singular Vectors. We consider $\mathbf{P}(\mathbf{a}_i,\mathbf{a}_j)$ and $\mathbf{P}(\mathbf{b}_k,\mathbf{b}_l)$ optimal coupling solutions of the OT problems for the costs $\mathbf{C}$ and $\mathbf{D}$ respectively. These optimal couplings induce a graph on $\{(i,j),(k,l)\}$ by linking*

$$(k,l) \to (i,j) \text{ when } \mathbf{P}(\mathbf{a}_i,\mathbf{a}_j)_{k,l} > 0$$
$$(i,j) \to (k,l) \text{ when } \mathbf{P}(\mathbf{b}_k,\mathbf{b}_l)_{i,j} > 0.$$

If there exist optimal couplings such that this graph is strongly connected, then $(\mathbf{C},\mathbf{D})$ are the unique Wasserstein singular vectors.

Proof. Let us consider $\Phi : (\mathbf{C}, \mathbf{D}) \mapsto (\Phi_{\mathbf{B}}(\mathbf{D}), \Phi_{\mathbf{A}}(\mathbf{C}))$ which maps $\mathcal{D}_m \times \mathcal{D}_m$ to itself. We use Theorem 7.5 of [Akian et al., 2016]. It requires that the semi-differential of Φ at $(\mathbf{C}, \mathbf{D})$ has itself a unique positive eigenvector in $\mathcal{D}_m \times \mathcal{D}_n$. In fact, this eigenvector is also $(\mathbf{C}, \mathbf{D})$. From the envelope theorem, upper-gradients of the concave function $\mathbf{C} \mapsto \mathrm{W}_{\mathbf{C}}(\mathbf{a}_i, \mathbf{a}_j)$ are the elements $\mathbf{P}(\mathbf{a}_i, \mathbf{a}_j)$ (so if the optimal coupling is unique, then this map is differentiable). The semi-differential Ψ of Φ is then a block anti-diagonal matrix defining the graph detailed in the statement of the theorem. The (linear) Perron-Frobenius theorem for positive linear operators ensures the existence of a unique positive eigenvector of Ψ if this graph is connected. □

Convergence of power iterations In the case of linear positive maps, Perron-Frobenius theory ensures the convergence of Equation 3.4 toward the unique positive singular vectors at a linear rate. Unfortunately, this result does not hold in general for the case of non-linear maps, and $\Phi_{\mathbf{A}}$ is only non-expansive (and not necessarily contracting). The following proposition, proved in Appendix B.3, states that for large enough regularization, uniqueness and linear convergence are maintained.

Proposition 3.2.6. *For τ large enough, the singular vectors are unique and the power iterations Equation 3.4 converge linearly for $\|\cdot\|_\infty$. When $\tau \to \infty$, the singular vectors converge to $C_{k,l} \propto R(\mathbf{b}_k - \mathbf{b}_l)$ and $D_{i,j} \propto R(\mathbf{a}_i - \mathbf{a}_j)$.*

The numerical simulations of Section 3.2.3 suggest that uniqueness and linear rates always hold in practical cases.

3.2.3 Numerical illustration on translated histograms

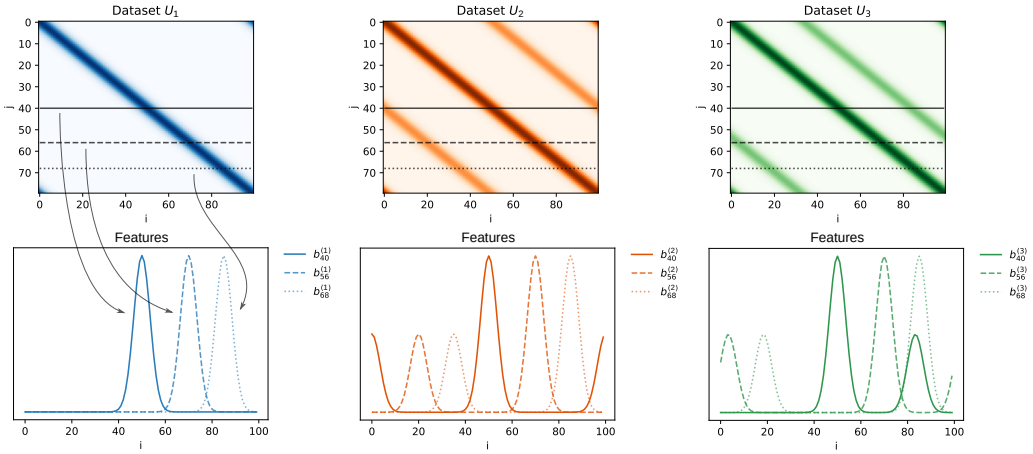
Motivation In general, the singular vectors $\mathbf{C}$ and $\mathbf{D}$ may be complex to interpret. The aim of this section is to present a simple numerical illustration, where the singular vectors $\mathbf{C}$ and $\mathbf{D}$ boil down to periodic 1-D functions, in order to verify that the learned metrics are reasonable. For this, we consider samples $\mathbf{a}_i \in \Sigma_m$ that are translations of a template $\mathbf{a}_0$.

Generating translated histograms We generate three synthetic datasets $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3 \in \mathbb{R}^{m \times n}$ by translating three different templates (see Figure 3.2). We use $n = 100$ samples and $m = 80$ features. We define the datasets by $[\mathbf{U}_p]_{k,i} \overset{\text{def.}}{=} h_p(k/m - i/n)$, and $\mathbf{A}_p$ (resp. $\mathbf{B}_p$) is obtained by normalizing $\mathbf{U}_p$ along columns (resp. rows). The templates h_1, h_2, h_3 are three different periodic functions on the 1-D torus (we use periodic boundary conditions): $h_1(x) \propto \exp\left(-|x|^2/\sigma^2\right)$, $h_2(x) \propto \exp\left(-|x|^2/\sigma^2\right) + \frac{1}{2}\exp\left(-|x+1/2|^2/\sigma^2\right)$, and $h_3(x) \propto \exp\left(-|x|^2/\sigma^2\right) + \frac{1}{2}\exp\left(-|x+1/3|^2/\sigma^2\right)$.

Interpreting the Wasserstein singular vectors We compute the Wasserstein singular vectors for different values of τ. By translational invariance of the problem, the singular vectors $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{C}_3$ are of the form $(\mathbf{C}_p)_{k,l} = (\mathbf{c}_p)_{k-l}$ where $\mathbf{c}_p = (\mathbf{C}_p)_{0,\cdot}$ are periodic 1-D functions. The same argument applies to the singular vectors $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$, displayed in Figure 3.3. $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{C}_3$ are symmetric and can be found in Appendix B.4. The learned metrics integrate geometrical properties (symmetries, multi-modalities, etc.) of the input datasets. For the unimodal Gaussian-like distribution ($\mathbf{D}_1$), the learned metric is close to $|\sin(k/m - i/n)|$, but exhibits non-monotonic behavior for multi-modal distributions ($\mathbf{D}_2$ and $\mathbf{D}_3$).

Convergence rates Figure 3.4 reports in logarithmic scale the convergence rate of power iterations according to the Hilbert metric $d_{\mathcal{H}}(\mathbf{D}, \mathbf{D}') \overset{\text{def.}}{=} \|\log(\mathbf{D}/\mathbf{D}')\|_V$ where $\|\mathbf{Z}\|_V \overset{\text{def.}}{=} \max(\mathbf{Z}) - \min(\mathbf{Z})$. This speed is always linear, suggesting that the maps $\Phi_{\mathbf{A}_p}$ are contracting and that the singular vectors are

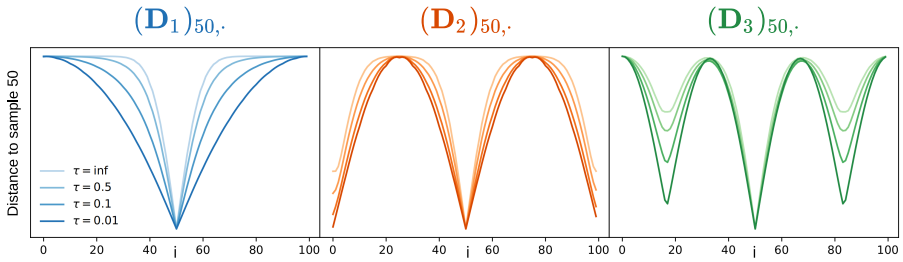
Figure 3.2: Translated histograms on the 1-D torus.



(top) Datasets $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$ generated by the functions $h_1(x) \propto \exp\left(-|x|^2/\sigma^2\right)$, $h_2(x) \propto \exp\left(-|x|^2/\sigma^2\right) + \frac{1}{2}\exp\left(-|x+1/2|^2/\sigma^2\right)$, and $h_3(x) \propto \exp\left(-|x|^2/\sigma^2\right) + \frac{1}{2}\exp\left(-|x+1/3|^2/\sigma^2\right)$.

(bottom) For each dataset, three features are displayed. They correspond to rows of the dataset and are identical up to a translation.

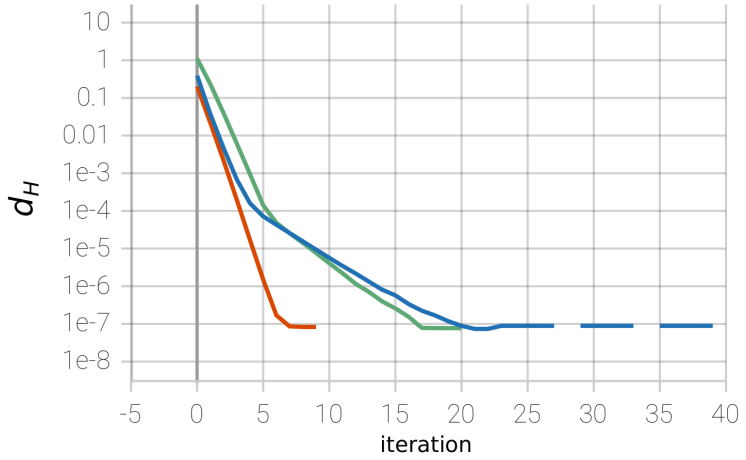
Figure 3.3: Learned singular vectors on the 1-D torus.



By translational invariance of the problem, the singular vectors $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$ are of the form $(\mathbf{D}_p)_{k,l} = (\mathbf{d}_p)_{k-l}$ where $\mathbf{d}_p = (\mathbf{D}_p)_{0,\cdot}$ are periodic 1-D functions. We here display the periodic 1-D functions associated with the singular vectors $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{D}_3$ for varying values of τ.

unique (which is confirmed by running several initializations in $\mathcal{D}_n$, and through condition 3.2.5). The contractance rate (which is the slope of the error curves) is dependent on the geometry of the templates h_p. We also observed a steeper slope for larger values of τ.

Figure 3.4: Convergence of singular vectors for translated histograms on the 1-D torus.



Convergence of the singular vectors $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{D}_3$ along the power iterations for $\tau = 0.1$, according to the $d_{\mathcal{H}}$ metric.

3.3 Large scale stochastic power iterations

As n or m grows, the complexity of the power iterations Equation 3.4 becomes prohibitive. In order to work around this issue we propose a stochastic power iteration scheme similar in spirit to stochastic gradient descent, which updates a single (or several if applied in a mini-batch setting) randomly chosen distance value at each step. This speeds up each iteration and leverages the correlations in the dataset.

Stochastic power iterations For some decreasing step size α_t and scaling factors $\tilde{\lambda}_t, \tilde{\gamma}_t > 0$, we define

$$\mathbf{C}_{t+1} \overset{\text{def.}}{=} \Pi\left((1-\alpha_t)\,\mathbf{C}_t + \alpha_t \tilde{\mathbf{C}}_t\right),$$
$$\mathbf{D}_{t+1} \overset{\text{def.}}{=} \Pi\left((1-\alpha_t)\,\mathbf{D}_t + \alpha_t \tilde{\mathbf{D}}_t\right),$$
$$\text{where} \quad (\tilde{\mathbf{D}}_t)_{i,j} \overset{\text{def.}}{=} \begin{cases} \Phi_{\mathbf{A}}(\mathbf{C}_t)_{i,j}/\tilde{\gamma}_t \text{ if } (i,j) = (i_t, j_t), \\ (\mathbf{D}_t)_{i,j} \text{ otherwise.} \end{cases}$$

We define $\Pi(\mathbf{C}) \overset{\text{def.}}{=} \mathbf{C}/\|\mathbf{C}\|_\infty$ and (i_t, j_t) is is drawn uniformly at random in $\{1, \dots, n\}^2$. It is the index updated at each step. $\tilde{\mathbf{C}}_t$ is computed by an analogous update rule, with $\tilde{\gamma}_t$ replaced by $\tilde{\lambda}_t$.

Convergence of stochastic power iterations and choice of $\tilde{\lambda}_t, \tilde{\gamma}_t$ The following theorem, proved in Appendix B.5, guarantees that for a large enough regularization parameter τ, these iterations converge to a pair of Wasserstein Singular Vectors. However, we observe empirically that these iterations converge even for arbitrary small τ and for $\tilde{\lambda}_t = \tilde{\gamma}_t = 1$. In practice, we suggest updating

several entries of the matrix at each step and estimating $\tilde{\lambda}_t$ and $\tilde{\gamma}_t$ as proposed in Remark 3.3.2. This approach removes the need to set an additional pair of hyperparameters $\tilde{\lambda}_t$ and $\tilde{\gamma}_t$.

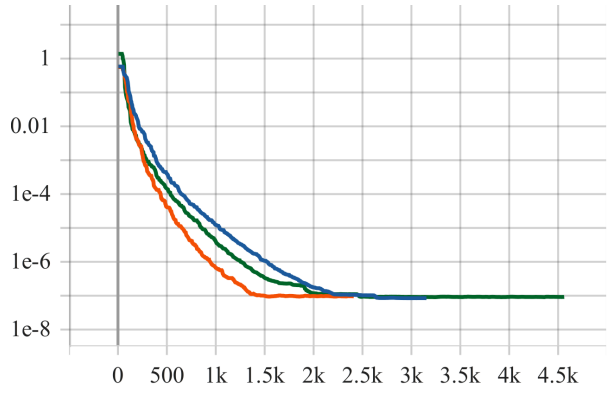
Theorem 3.3.1. *For $\alpha_t = 1/\sqrt{t}$, for constant scaling factors $\tilde{\lambda}_t \leq \tau \min_{k \neq l} R(\mathbf{b}_k - \mathbf{b}_l)$ and $\tilde{\gamma}_t \leq \tau \min_{i \neq j} R(\mathbf{a}_i - \mathbf{a}_j)$, and for τ large enough, the stochastic power iterations defined above converge to a pair $(\mathbf{C}, \mathbf{D}) \in \mathcal{D}_m \times \mathcal{D}_n$ of positive singular vectors with a convergence rate of $\mathcal{O}(\log(t)/\sqrt{t})$.*

Remark 3.3.2 (Adaptive selection of $\tilde{\lambda}_t$ and $\tilde{\gamma}_t$)**.** *Tuning the values of the parameters $\tilde{\lambda}_t$ and $\tilde{\gamma}_t$ is crucial to improving the convergence. Ideally, they should be as close as possible to the (unknown) singular values (λ, γ). Instead of fixing these scaling factors prior to running the algorithm, we propose using an estimation of the singular values. When using mini-batching, i.e. updating several indices $(i, j) \in \mathcal{I}$ at each iteration, one can use a least square estimate,*

$$\tilde{\gamma}_{t+1} = (1 - \alpha_t)\tilde{\gamma}_t + \alpha_t \frac{\sum_{(i,j)\in\mathcal{I}} \Phi_A(\mathbf{C}_t)_{i,j} (\mathbf{D}_t)_{i,j}}{\sum_{(i,j)\in\mathcal{I}} (\mathbf{D}_t)_{i,j}^2},$$

and similarly for $\tilde{\lambda}_{t+1}$. We found that in practice, $(\tilde{\lambda}_t, \tilde{\gamma}_t)$ quickly converges to the singular values (λ, γ).

Figure 3.5: Convergence rates $d_{\mathcal{H}}(\mathbf{D}_t, \mathbf{D}_\infty)$ of stochastic updates.



Numerical illustration on translated histograms Figure 3.5 illustrates the convergence of stochastic power iterations in practice by comparing the approximated Wasserstein Singular Vectors for the synthetic experiments of Section 3.2.3 with the true singular vectors obtained using the non-stochastic power iterations Equation 3.4. We used the approach outlined in Remark 3.3.2. As expected, the stochastic power iterations yield the same result as classical power iterations.

3.4 Parallelization with entropic regularisation

To further speed up the method, we propose to use the entropic regularization of Optimal Transport [Cuturi, 2013], as introduced in the introduction of this thesis (see Section 1.3.1).

Sinkhorn's algorithm Entropic OT can be computed efficiently in $\mathcal{O}(n^2/\eta^2)$ using Sinkhorn's algorithm (detailed in Appendix B.6) at the expense of an approximation of order η [Altschuler et al., 2017]. Beside speeding up the computation of OT, this enables embarrassingly parallel computations of the distance map on GPUs [Cuturi, 2013] and also reduces the curse of dimensionality which plagues OT [Genevay et al., 2019].

Remark 3.4.1. *The theoretical guarantees listed above apply to entropic OT with Euclidean ground costs, but not necessarily to the setting described in this paper. However, in practice we observed that the benefits of Sinkhorn's algorithm were maintained in our situation.*

3.4.1 Sinkhorn divergence map

Sinkhorn divergence As covered in more detail in Section 1.3.1, entropy regularized OT is defined

$$\mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}_i, \mathbf{a}_j) \stackrel{\text{def.}}{=} \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}_i, \mathbf{a}_j)} \langle \mathbf{P}, \mathbf{C} \rangle + \varepsilon \langle \mathbf{P}, \log \mathbf{P} - 1 \rangle.$$

This quantity suffers from a bias, which is removed by using instead the Sinkhorn divergence [Genevay et al., 2018]

$$\overline{\mathrm{W}_{\mathbf{C}}^{\varepsilon}}(\mathbf{a}_i, \mathbf{a}_j) \stackrel{\text{def.}}{=} \mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}_i, \mathbf{a}_j) - \tfrac{1}{2}\left(\mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}_i, \mathbf{a}_i) + \mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}_j, \mathbf{a}_j)\right).$$

This debiasing is crucial to ensure that $\overline{\mathrm{W}_{\mathbf{C}}^{\varepsilon}}(\mathbf{a}_i, \mathbf{a}_i) = 0$, and it also reduces the approximation error to $|\overline{\mathrm{W}_{\mathbf{C}}^{\varepsilon}} - \overline{\mathrm{W}_{\mathbf{C}}}| \sim \varepsilon^2$ [Chizat et al., 2020].

Sinkhorn divergence map Similarly to the map defined in Equation 3.2, we define the Sinkhorn divergence map as

$$\Phi_{\mathbf{A}}^{\varepsilon}(\mathbf{C} \in \mathcal{D}_m)_{i,j} \stackrel{\text{def.}}{=} \overline{\mathrm{W}_{\mathbf{C}}^{\varepsilon'}}(\mathbf{a}_i, \mathbf{a}_j) + \tau \|\mathbf{C}\|_{\infty} R(\mathbf{a}_i - \mathbf{a}_j) \text{ with } \varepsilon' = \varepsilon \times \|\mathbf{C}\|_{\infty},$$

and similarly for $\Phi_{\mathbf{B}}^{\varepsilon}$. It reduces to Equation 3.2 when $\varepsilon = 0$.

Sinkhorn singular vectors By analogy with Equation 3.3 we define *Sinkhorn singular vectors* as

$$(\mathbf{C}, \mathbf{D}) \text{ such that } \exists (\lambda, \gamma) \in {\mathbb{R}_+^*}^2 \text{ s.t. } \Phi_{\mathbf{B}}^{\varepsilon}(\mathbf{D}) = \lambda \mathbf{C}, \;\; \Phi_{\mathbf{A}}^{\varepsilon}(\mathbf{C}) = \gamma \mathbf{D}, \tag{3.7}$$

Remark 3.4.2 (Positivity property)**.** *The map Φ^{ε} is 1-homogeneous and monotone, but it is unclear that it always maps onto positive matrices. It is proved in [Feydy et al., 2019] that it is the case if $e^{-\mathbf{C}/(\varepsilon\|\mathbf{C}\|_{\infty})}$ is a positive kernel (i.e. has positive eigenvalues). While it is unclear that such a condition is maintained during power iterations, we observed numerically that it is still the case in practice. We show below that this is true in the limit $\varepsilon \to +\infty$.*

Choice of entropic regularization Similarly to the hyperparameter τ, one may benefit from calibrating ε according to the problem. However, differently from the choice of τ, larger values of ε speed up the computation of the matrix $\Phi_{\mathbf{A}}^{\varepsilon}(\mathbf{C})$, which may play a role in the choice of ε. The following section seeks to explain the limit behaviour when $\varepsilon \to \infty$.

3.4.2 Connection with PCA when $\varepsilon \to \infty$

Maximum Mean Discrepancy limit For simplicity, let us consider the case $\tau = 0$. We show in Proposition 3.4.3 below that when $\varepsilon \to \infty$, our method operates over the set of squared Euclidean matrices

$$\mathbf{C} \in \mathcal{K}_m \subset \mathcal{D}_m \iff \exists (\mathbf{u}_k \in \mathbb{R}^d)_{k=1}^m, C_{k,l} = \|\mathbf{u}_k - \mathbf{u}_l\|_2^2.$$

Note that these matrices can be equivalently defined as conditionally negative matrices with zero diagonal, see [Schölkopf and Smola, 2002].

Proposition 3.4.3. *One has* $\Phi_{\mathbf{A}}^{\infty} : \mathcal{K}_m \to \mathcal{K}_n$ *where*

$$\Phi_{\mathbf{A}}^{\infty}(\mathbf{C}) \overset{def.}{=} \lim_{\varepsilon \to \infty} \Phi_{\mathbf{A}}^{\varepsilon}(\mathbf{C}) = (-\tfrac{1}{2}\langle \mathbf{C}(\mathbf{a}_k - \mathbf{a}_l), \mathbf{a}_k - \mathbf{a}_l \rangle)_{k,l}.$$

This property shows that in the large ε limits $\Phi_{\mathbf{A}}^{\infty}$ is actually a linear map which computes Maximum Mean Discrepancies [Gretton et al., 2012] (a.k.a. Euclidean distances between probability distributions).

Connection with PCA For the sake of simplicity in the exposition, let us now assume $\mathbf{A} = \mathbf{B}^\top$. In this case, Equation 3.3 is a classical linear singular vectors problem. While in general, ensuring existence of positive singular vectors is non-trivial, the following proposition, proved in Appendix B.7, shows that this is the case for $\Phi_{\mathbf{A}}^{\infty}$.

Proposition 3.4.4 (Connection with PCA)**.** *Let us denote* $\tilde{\mathbf{A}} \overset{def.}{=} \mathbf{A} - \mathbf{A}\mathbf{1}_m\mathbf{1}_m^\top/m$ *the centered matrix. For any pair* $(\mathbf{u}, \mathbf{v})$ *of singular vectors of* $\tilde{\mathbf{A}}$ *with singular value* λ*,*

$$\mathbf{C} = ((v_k - v_l)^2)_{k,l} \in \mathcal{K}_m, \ \mathbf{D} = ((u_i - u_j)^2)_{i,j} \in \mathcal{K}_n$$

are singular vectors of $(\Phi_{\mathbf{A}}^{\infty}, \Phi_{\mathbf{B}}^{\infty})$*, with singular value* $2\lambda^2$*.*

This proposition shows that for $\varepsilon = +\infty$ a set of positive singular vectors is obtained as simply squared Euclidean distances over 1-D principal component embeddings of the data. Entropic regularization thus draws a link between our novel set of OT-based metric learning techniques and classical dimensionality reduction methods. This frames Sinkhorn singular vectors as a well-posed problem regardless of the value of ε.

3.5 Metric learning for single-cell genomics

scRNA-seq data Single-cell RNA sequencing (*scRNA-seq*) is a high-throughput sequencing technology enabling the measurement of gene expression levels at single-cell resolution [Stegle et al., 2015] (see Section 1.1.1). The analysis of scRNA-seq data has offered unprecedented insights in cellular heterogeneity and disease mechanisms [Tanay and Regev, 2017, Yuan et al., 2017]. scRNA-seq data can be represented as a matrix of integer expression levels with genes on rows and cells on columns. One of the main uses of scRNA-seq is to identify cell populations through clustering or visualization. But these tasks rely on some notion of distance between cells. The most popular clustering and visualization tools, in particular Scanpy [Wolf et al., 2018] and Seurat [Stuart et al., 2019], rely on Euclidean distances on PCA embeddings of cells. Embeddings can also be provided by deep learning models like scVI [Gayoso et al., 2022]. A good metric on the space of genes is also important because the phenotype of a cell is determined by the joint activity of all its expressed genes.

Optimal Transport distances between cells In order to take advantage of the biological relationships between genes, OT distances between cells have recently been proposed. The Gene Mover Distance [Bellazzi et al., 2021] is defined similarly to the Word Mover Distance [Kusner et al., 2015]: the authors use as a ground cost the Euclidean distance between precomputed Gene2Vec [Du et al., 2019] embeddings. [Huizing et al., 2022] use a Sinkhorn divergence with a cosine distance between genes (i.e. vectors of cells) as a ground cost. In the present paper we compute OT distances using the Python package POT [Flamary et al., 2021].

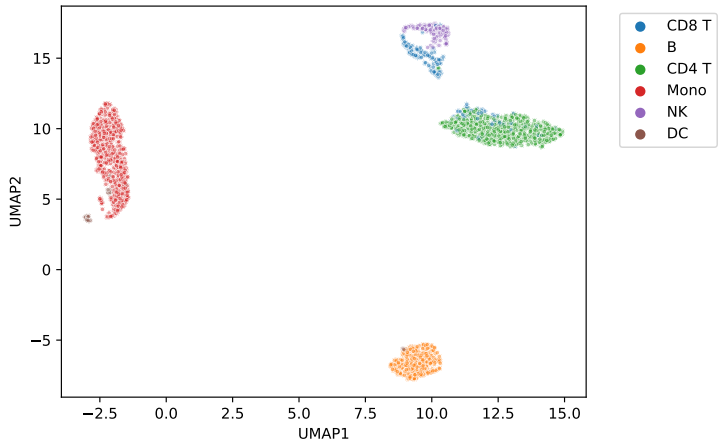
Dataset A commonly analyzed scRNA-seq dataset is the "PBMC 3k" dataset produced by 10X Genomics, obtained through the function `pbmc3k` of Scanpy [Wolf et al., 2018]. Details on preprocessing and cell type annotation are given in Appendix B.8. The processed dataset contains $m = 1030$ genes and $n = 2043$ cells, each belonging to one of 6 immune cell types: 'B cell', 'Natural Killer', 'CD4+ T cell', 'CD8+ T cell', 'Dendritic cell' and 'Monocyte'. The cell populations are heavily unbalanced. In addition, for each cell type we consider the set of canonical marker genes given by Azimuth [Hao et al., 2021], i.e. genes whose expression is characteristic of a certain cell type.

Evaluating distance matrices between cells In order to evaluate the quality of pairwise distance matrices between cells, we use the cell type annotation. We compute the Silhouette score, which evaluates whether cells from the same cell type are closer than to other cell types. We use the function `silhouette_score` of Scikit-learn [Pedregosa et al., 2011]. We compare (i) Euclidean distances on PCA embeddings (ii) Euclidean distances on Kernel PCA embeddings using Scikit-learn's implementation with `kernel='rbf'` (iii) Euclidean distances on scVI [Gayoso et al., 2022] embeddings using default values (iv) Gene Mover Distance (v) Sinkhorn divergence ($\varepsilon = .1$) with a cosine distance between genes as a ground cost (vi) Wasserstein Singular Vector $\mathbf{D}$ ($\tau = 0.001, \varepsilon = 0.1$), approached by 15 power iterations, which we found in this case to lead to better results than the stochastic power iterations. Note that for large regularisation ε, as shown in Section 3.4, the Wasserstein Singular Vectors are themselves (squared) Euclidean distances on PCA embeddings.

Evaluating distance matrices between genes Similarly, we compute Silhouette scores based on marker genes for each cell type, to assess whether genes that are markers for the same cell type are closer than to markers for other cell types. We compare (i) Euclidean distances on the original data (ii) Euclidean distances between gene2vec embeddings (iii) Wasserstein Singular Vector $\mathbf{C}$ ($\tau = 0.001, \varepsilon = 0.1$), approached by 15 power iterations, which we found in this case to lead to better results than the stochastic power iterations.

Results The results in Table 3.1 and Table 3.2 suggest that our method improves over all considered baselines. Figure 3.6 and Figure 3.7 shows the UMAP projection of the cells and the genes in the dataset. The Wassersein Singular Vectors clearly outperform the other metrics in terms of Silhouette scores, both in the context of cells and of genes. Interestingly, in the case of marker genes we outperform the Euclidean distance on Gene2Vec embeddings, which are meant to contain "semantic" information about genes. These scores are also validated by the UMAP projection, where cells and marker genes cluster according to cell type. These results motivate further research in the biological implications of Wasserstein Singular Vectors. Let us highlight that we compute distances between cells or genes, but that we do not produce embeddings like PCA or scVI. In addition, scVI can handle complex tasks like the removal of unwanted sources of variation [Gayoso et al., 2022] which we do not consider in this article.

Figure 3.6: 2-D UMAP projection of the cells using the singular vector D.



Visualization of the singular vector **D**, a pairwise distance matrix between the cells of a scRNA-seq dataset, obtained for $\tau = 0.001, \varepsilon = 0.1$. We project the matrix **D** to a 2-D UMAP representation of the cells, which are colored by cell type.

Table 3.1: Silhouette score for cells

Method	Silhouette
PCA / ℓ^2	0.238
Kernel PCA / ℓ^2	0.241
scVI embedding / ℓ^2	0.168
Sinkhorn	0.003
Gene Mover Distance	0.066
WSV (ours)	**0.348**

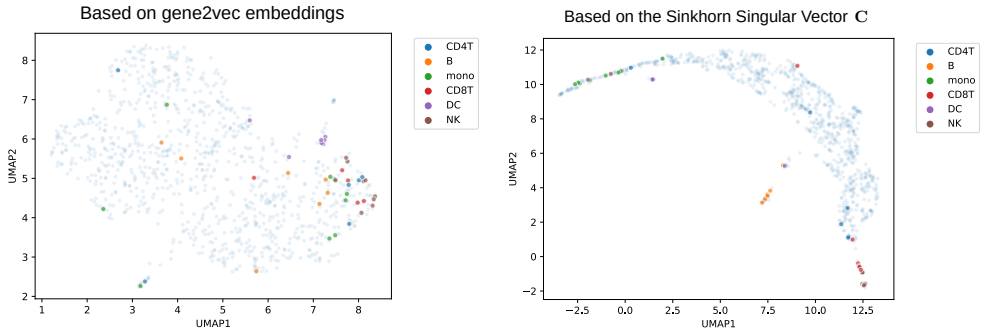
Table 3.2: Silhouette score for marker genes

ℓ^2	Gene2Vec / ℓ^2	WSV (ours)
-0.005	0.0186	**0.136**

3.6 Discussion

Wasserstein Singular Vectors define a pair of "intrinsic" ground metrics associated to a given dataset. This elegantly solves the problem of unsupervised ground metric learning without resorting to ad hoc embeddings. Numerical results on single-cell RNA sequencing suggest that these metrics encode salient geometric structures of the data. This opens several avenues for future works, in particular an in-depth theoretical analysis when $\tau = 0$ and $\varepsilon > 0$. Our method can be extended to unbalanced optimal transport [Liero et al., 2015, Chizat et al., 2018], which has proved useful to increase the robustness of the metric for the analysis of biological sequencing datasets [Schiebinger et al., 2019].

Figure 3.7: 2-D UMAP projection of marker genes, using the computed distances.



Visualization of the genes from a scRNA-seq dataset. Marker genes are colored by associated cell type, and other genes are faded out. **(left)** 2-D UMAP embeddings based on Euclidean distances on Gene2Vec [Du et al., 2019] embeddings; **(right)** 2-D UMAP embeddings based on the Sinkhorn Singular Vector $\mathbf{C}$, obtained for $\tau = 0.001, \varepsilon = 0.1$.

Lastly, our initial results regarding stochastic approximation of Wasserstein Singular Vectors would greatly benefit from further developments enabling better convergence rates.

Chapter 4

Paired single-cell multi-omics data integration with Mowgli

Abstract

The profiling of multiple molecular layers from the same set of cells has recently become possible. There is thus a growing need for multi-view learning methods able to jointly analyze these data.

We here present Multi-Omics Wasserstein inteGrative anaLysIs (Mowgli), a novel method for the integration of paired multi-omics data with any type and number of omics. Of note, Mowgli combines integrative Nonnegative Matrix Factorization and Optimal Transport, enhancing at the same time the clustering performance and interpretability of integrative Nonnegative Matrix Factorization. We apply Mowgli to multiple paired single-cell multi-omics data profiled with 10X Multiome, CITE-seq, and TEA-seq. Our in-depth benchmark demonstrates that Mowgli's performance is competitive with the state-of-the-art in cell clustering and superior to the state-of-the-art once considering biological interpretability.

Mowgli is implemented as a Python package seamlessly integrated within the scverse ecosystem and it is available at github.com/cantinilab/mowgli.

This chapter was published as

Geert-Jan Huizing, Ina Maria Deutschmann, Gabriel Peyré, Laura Cantini, *Paired single-cell multi-omics data integration with Mowgli.* Nature Communications 2023. doi.org/10.1038/s41467-023-43019-2

Contents

4.1 Introduction **66**
4.2 Results **67**
4.2.1 Mowgli: a new tool for paired single-cell multi-omics data integration 67
4.2.2 Mowgli's cell embedding and clustering outperform the state-of-the-art in controlled settings derived from cell lines data 69
4.2.3 Mowgli's cell embeddings and clusterings are competitive with the state-of-the-art in complex and heterogeneous datasets 71
4.2.4 Mowgli improves the biological interpretability of the state-of-the-art by providing cell-type specific factors in TEA-seq data 73
4.2.5 Mowgli identifies relevant subpopulations of immune cells in TEA-seq data . 74
4.3 Discussion **77**
4.4 Methods **78**
4.4.1 Notations 78
4.4.2 Optimal transport 78
4.4.3 Mowgli 79
4.4.4 Benchmarked methods 81
4.4.5 Choice of the number of latent dimensions for all methods 81
4.4.6 Data generation 82
4.4.7 Data preprocessing 82
4.4.8 Data analysis 83
4.4.9 Evaluation metrics 83
4.4.10 Biological interpretation 84
4.4.11 Data availability 84
4.4.12 Code availability 85

4.1 Introduction

Single-cell sequencing technologies, providing a quantitative and unbiased characterization of cellular heterogeneity, are revolutionizing our understanding of the immune system, development, and complex diseases [Rajewsky et al., 2020, Potter, 2018, Papalexi and Satija, 2018]. A new frontier in single-cell sequencing technologies is represented by multi-omics single-cell sequencing, allowing for the simultaneous profiling of multiple molecular readouts (e.g. transcriptome, chromatin accessibility, surface proteins) from the same cell [Lee et al., 2020, Stoeckius et al., 2017, Clark et al., 2018, Swanson et al., 2021, Liu et al., 2019, Angermueller et al., 2016, Cao et al., 2018, Chen et al., 2019, Mimitou et al., 2019]. Examples of these cutting-edge sequencing technologies are CITE-seq, simultaneously measuring RNA and surface protein abundance by leveraging oligonucleotide-conjugated antibodies [Stoeckius et al., 2017], and 10x Genomics Multiome platform, quantifying RNA and chromatin accessibility by microdroplet-based isolation of single nuclei.

Multi-omics single-cell sequencing platforms provide us with complementary molecular readouts from exactly the same set of cells, called in the following paired multi-omics data. The joint analysis of such data offers the exciting opportunity to understand how different molecular facets of a cell collaboratively define the cell's function, morphology, and state [Miao et al., 2021]. Several multi-view

learning methods, jointly analyzing paired multi-omics data by taking into account their shared and complementary information, have thus been recently developed [Argelaguet et al., 2020, Hao et al., 2021, Gayoso et al., 2021, Ashuach et al., 2023, Zuo and Chen, 2021, Duren et al., 2022, Singh et al., 2021, Wang et al., 2020, Jin et al., 2020, Kim et al., 2020]. These methods, differently from unpaired integration ones [Welch et al., 2019, Stuart et al., 2019], take advantage of the known correspondences between cells across modalities. State-of-the-art multi-view learning methods for single-cell multi-omics integration are based on integrative Matrix Factorization [Argelaguet et al., 2020, Duren et al., 2022, Jin et al., 2020], k-nearest neighbors [Hao et al., 2021], or variational autoencoders [Gayoso et al., 2021, Ashuach et al., 2023, Zuo and Chen, 2021, Gong et al., 2021, Lotfollahi et al., 2022]. Integrative Matrix Factorization (integrative MF) and variational autoencoders perform dimensionality reduction, jointly embedding the high-dimensional multi-omics cellular profilings into a shared lower-dimensional latent space by leveraging common cells/observations [Miao et al., 2021, Stanojevic et al., 2022]. Integrative MF, due to its linear nature, defines a latent space with a natural biological interpretation, but it is too simple to catch complex biological processes [Miao et al., 2021, Stanojevic et al., 2022]. On the other hand, non-linear methods, as variational autoencoders, have shown great potential in clustering cells, but despite recent works on the subject [Ainsworth et al., 2018, Svensson et al., 2020], they inherently lack biological interpretability. Improving integrative MF methods is thus crucial to striking a balance between interpretability and performance.

We here propose Multi-Omics Wasserstein inteGrative anaLysIs (Mowgli github.com/cantinilab/mowgli), a novel integrative MF method for single-cell multi-omics data combining integrative Nonnegative Matrix Factorization [Lee and Seung, 1999] (integrative NMF) with Optimal Transport [Monge, 1781] (OT). On one hand, Mowgli employs integrative NMF, popular in computational biology due to its intuitive representation by parts and further enhances its interpretability [Lee and Seung, 1999]. On the other hand, Mowgli enhances the clustering performances of integrative MF by taking advantage of OT, which we have previously shown to better capture similarities between single-cell omics profiles [Huizing et al., 2022]. We extensively benchmarked Mowgli with respect to the state-of-the-art in the integration of several paired multi-omics data profiled with CITE-seq [Stoeckius et al., 2017], 10X Genomics Multiome, and TEA-seq [Swanson et al., 2021] platforms. Of note, while we focused on the integration of the currently available omics data, Mowgli can deal with paired multi-omics datasets with any type and number of omics, without any statistical assumption on the data. The performed in-depth comparison showed that Mowgli's embedding and clustering quality outperform the state-of-the-art in controlled settings derived from real multi-omics data and are competitive with the state-of-the-art in more complex real multi-omics data. Of note, the latter are affected by the lack of an absolute ground truth annotation on most real datasets. Finally, Mowgli was shown to improve the state-of-the-art in terms of biological interpretability through an in-depth biological analysis of TEA-seq data.

4.2 Results

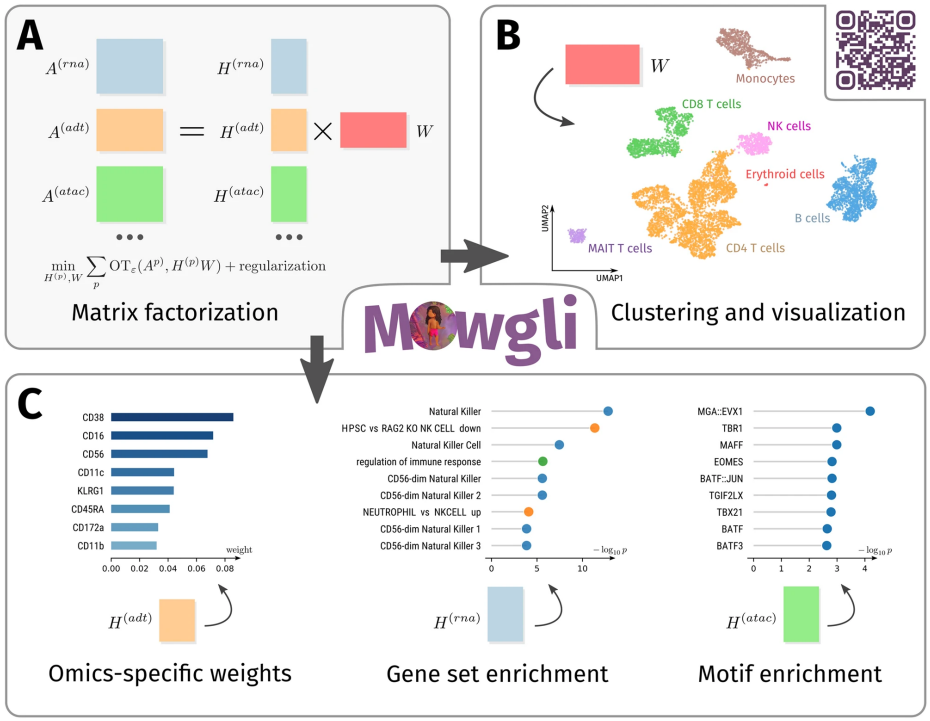
4.2.1 Mowgli: a new tool for paired single-cell multi-omics data integration

We developed Multi-Omics Wasserstein inteGrative anaLysIs (Mowgli), a new tool for paired single-cell multi-omics data integration (github.com/cantinilab/mowgli).

Mowgli is based on integrative Matrix Factorization (integrative MF). Starting from D omics matrices $\mathbf{A}^{(p)} \in \mathbb{R}^{m_p \times n}$ with $p \in [1 \dots D]$, sharing the same columns (the cells) but having different features (e.g. genes, peaks), Mowgli jointly decomposes them into the product of omic-specific dictionaries

$\mathbf{H}^{(p)} \in \mathbb{R}^{m_p \times d}$ and a shared embedding $\mathbf{W} \in \mathbb{R}^{d \times n}$ with $d \ll m_p$ and $d \ll n$ (Figure 4.1A). As a standard nomenclature, in the following we will call d the number of latent dimensions, the columns of $\mathbf{H}^{(p)}$ loadings and the rows of $\mathbf{W}$ factors [Stanojevic et al., 2022, Stein-O'Brien et al., 2018].

Figure 4.1: Overview of Mowgli.



(A) Schematic visualization of Mowgli, an NMF-based model with an Optimal Transport loss; **(B)** the matrix $\mathbf{W}$ of Mowgli can be used for cell clustering and visualization; **(C)** The dictionaries $\mathbf{H}^{(p)}$ of Mowgli contain omics-specific weights for each latent dimension, which can be used for the biological characterization of the latent dimensions through gene set enrichment or motif enrichment analysis.

In line with state-of-the-art MF methods for multi-omics integration [Cantini et al., 2021], the cell embedding $\mathbf{W}$ can be used to visualize and cluster the cells (Figure 4.1B) [Van der Maaten and Hinton, 2008, McInnes et al., 2018, Traag et al., 2019, Blondel et al., 2008]. The dictionaries $\mathbf{H}^{(p)}$ instead enable biological interpretation via gene set enrichment analysis [Mootha et al., 2003], motif enrichment analysis [Korhonen et al., 2017], or by identifying markers among the top weights (Figure 4.1C).

The main innovation of Mowgli is to perform integrative MF by combining integrative Non-Negative Matrix Factorization (integrative NMF) with Optimal Transport (OT). It thus solves the optimization problem:

$$\min_{\mathbf{H}^{(p)},\mathbf{W}} \sum_p \mathrm{W}_{\mathbf{C}}^{\varepsilon}\left(\mathbf{H}^{(p)}\mathbf{W}, \mathbf{A}^{(p)}\right) - \alpha_p \mathrm{E}\left(\mathbf{H}^{(p)}\right) - \beta \mathrm{E}(\mathbf{W}) \text{ with omics } p = 1, \ldots, D \tag{4.1}$$

In computational biology, integrative NMF is usually applied with an Euclidean reconstruction term between $\mathbf{H}^{(p)}\mathbf{W}$ and $\mathbf{A}^{(p)}$ [Duren et al., 2022, Jin et al., 2020, Welch et al., 2019]. We here introduce instead a reconstruction term based on entropy-regularized Optimal Transport (OT) (see Equation 4.1 and Methods), which unlike Euclidean or Kullback-Leibler losses leverages a notion of similarity between features. This choice is justified by the improved performance that we have previously observed once using OT to compare single-cell omics profiles [Huizing et al., 2022]. Of note, outside of biology, OT has been already used in the reconstruction loss of NMF for the factorization of single matrices [Rolet et al., 2016, Qian et al., 2016, Schmitz et al., 2018] and single tensors [Zhang, 2021].

In addition, as in [Rolet et al., 2016], we add to the optimization problem (Equation 4.1) two entropic regularization terms $\alpha_p \mathrm{E}\left(\mathbf{H}^{(p)}\right)$ and $\beta \mathrm{E}(\mathbf{W})$ (see Methods). These terms ensure that the loadings and embeddings are positive distributions and they control their sparsity (see Methods), a crucial feature to further enhance the known NMF's representation by parts and sparsity properties [Lee and Seung, 1999]. α_p and β are the coldness parameters of softmax functions (see Methods) and thus offer a natural way to adjust sparsity. For instance, as β approaches 0, cells will be assigned to only one factor. As instead β increases, cells will be a combination of several factors. For all details on the mathematical formulation of Mowgli see Methods.

Of note, Mowgli is implemented as an open-source Python package seamlessly integrated into the classical Python single-cell analysis pipeline (github.com/cantinilab/mowgli). Users can thus take advantage of scverse tools like Scanpy and Muon for preprocessing and downstream analysis [Wolf et al., 2018, Bredikhin et al., 2022]. In addition, Mowgli provides a user-friendly visualization of top genes and enriched gene sets, thus helping biological interpretability.

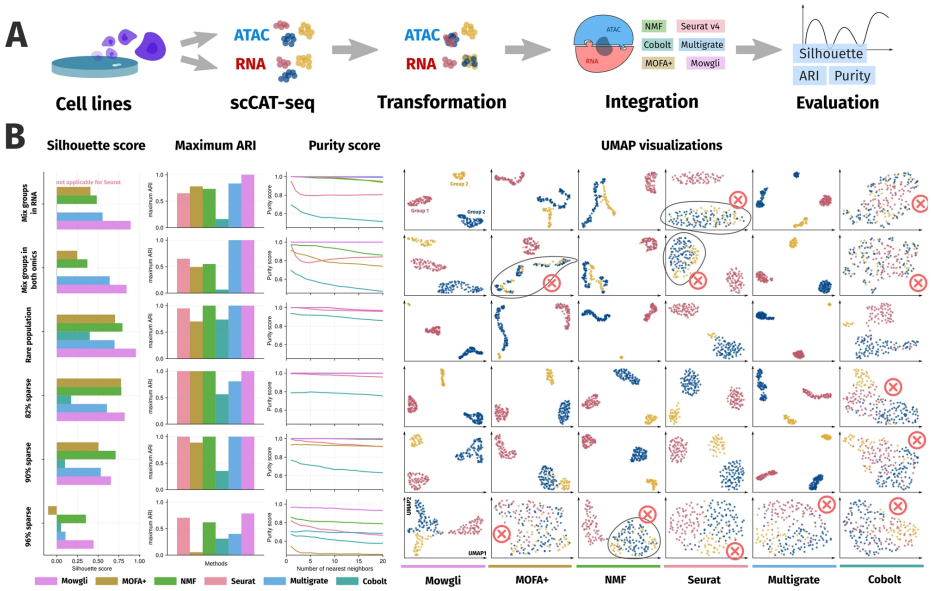
In the following, we extensively benchmark Mowgli against the state-of-the-art: Seurat v4 [Hao et al., 2021], Cobolt [Gong et al., 2021], Multigrate [Lotfollahi et al., 2022], and MOFA+ [Argelaguet et al., 2020]. Although several methods exist [Argelaguet et al., 2020, Hao et al., 2021, Gayoso et al., 2021, Ashuach et al., 2023, Zuo and Chen, 2021, Duren et al., 2022, Singh et al., 2021, Wang et al., 2020, Jin et al., 2020, Kim et al., 2020], we here focused on the leading methods for paired data integration that could be applied to the multiple combinations of single-cell omics data here considered. In addition, an integrative NMF baseline is also considered (see Methods), to further compare Mowgli with the standard integrative NMF.

4.2.2 Mowgli's cell embedding and clustering outperform the state-of-the-art in controlled settings derived from cell lines data

We first focused on evaluating Mowgli's embedding and clustering performance in controlled settings derived from cell lines data. To represent a panel of realistic scenarios with different distributions of cells across three groups, we applied different transformations to a simple dataset composed of three cancer cell lines profiled with scCATseq (see Figure 4.2A). The scCATseq dataset provides a joint profiling of scRNA and scATAC from HCT116, HeLa-S3, and K562 cell lines [Liu et al., 2019]. Unlike simulated data, this solution allows us to avoid making assumptions about the distribution of the data. Indeed, generating simulated data following a Gaussian distribution, for instance, would favor methods that approximate single-cell data with this same distribution.

The four scenarios in our panel represent distinct realistic challenges of multi-omics integration: (i)

Figure 4.2: Cell embedding and clustering benchmark in controlled settings derived from cell lines data.



(A) Schematic representation of the benchmarking process; **(B)** The first three columns of this panel are devoted to silhouette scores, Adjusted Rand Indices (ARIs), and purity scores for the different methods on six controlled settings derived from cell lines data. The following six columns provide UMAP visualizations for the six benchmarked methods (Mowgli, MOFA+, NMF, Seurat v4, Multigrate, Cobolt) on six controlled settings derived from cell lines data. Different colors in these UMAP plots correspond to the three groups of cells imposed in the dataset.

Mixed in RNA contains two cell populations that are mixed in scRNA, but well separated in scATAC; (ii) *Mixed in both* contains two cell populations mixed in scRNA and well separated in scATAC and two cell populations mixed in scATAC and well separated in scRNA; (iii) *Rare population* presents a population with much fewer cells than the others and (iv) *82% sparse*, *90% sparse*, and *96% sparse* contain data with increasing percentages of dropouts (82-96%). Scenarios (i) and (ii) test the ability of methods to take into account the complementarity of different omic data. Scenario (iii) tests the ability of the methods to recover rare populations. Finally, scenario (iv) tests the robustness of the methods to dropout noise, while staying in a realistic range of dropouts for single-cell data. For details on the generation of these datasets see Methods.

We benchmarked Mowgli, Seurat v4, MOFA+, integrative NMF, Cobolt, and Multigrate based on natural metrics for embedding and clustering performance: silhouette score, Adjusted Rand Index (ARI), and purity score (see Methods). The silhouette score cannot be computed for Seurat because it requires an embedding of the cells not provided by this method. In addition, we computed UMAP visualizations for the different methods and datasets [McInnes et al., 2018].

As shown in Figure 4.2B, overall, Mowgli provides superior performance over the current state-of-the-art according to all metrics. Indeed, in all datasets except *90% sparse*, Mowgli has a performance greater or equal to that of other methods. In the *90% sparse* dataset, integrative NMF has a better silhouette score than Mowgli but the same ARI and purity score. Cobolt performs poorly in *Mixed in RNA* and *Mixed in both* according to all three metrics.

These performances are confirmed by looking qualitatively at the UMAP plots in Figure 4.2B. In *Mixed in RNA* and *Mixed in both* Seurat v4 and Cobolt confuse populations when individual omics are not sufficient to identify the three groups. Regarding dropouts, one of the most challenging features of single-cell data [Lähnemann et al., 2020], Mowgli shows the highest resilience with respect to the state-of-the-art. Indeed, while a sparsity of 96% is still coherent with realistic data [Qiu, 2020], MOFA+, Seurat v4, Cobolt, and Multigrate confuse the three populations in the *96% sparse* dataset. On the opposite, Mowgli correctly separates the three groups of cells in *96% sparse*.

All methods except Seurat are parametrized by a number of latent dimensions. For each method, we choose the overall best-performing number of latent dimensions (see Methods). Tuning the number of latent dimensions for each metric and dataset does not change our conclusions (see Supplementary Figure C.1).

4.2.3 Mowgli's cell embeddings and clusterings are competitive with the state-of-the-art in complex and heterogeneous datasets

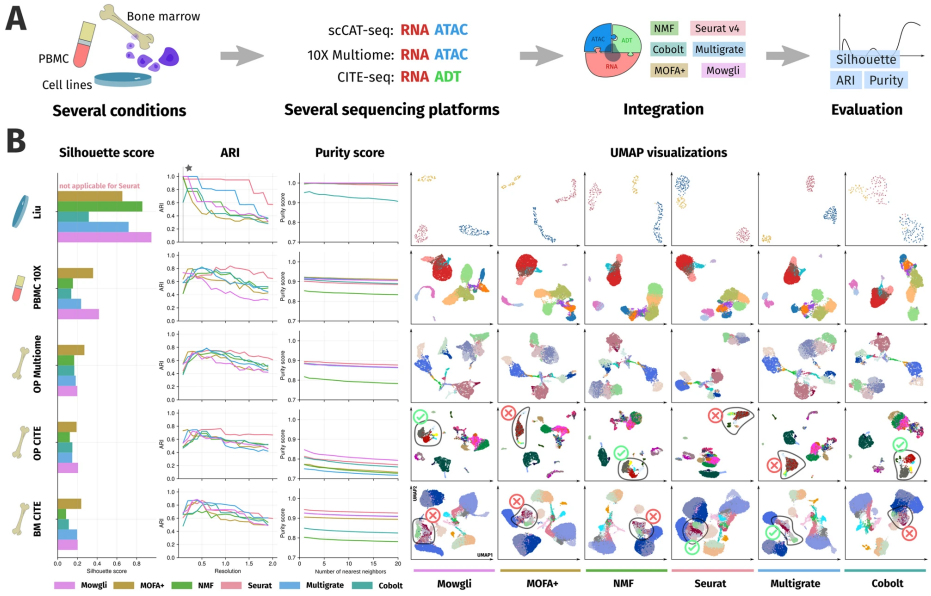
We then benchmarked Mowgli, Seurat v4, MOFA+, integrative NMF, Cobolt, and Multigrate based on their embedding and clustering performance on five paired single-cell multi-omics datasets (see Figure 4.3A). Of note, these data have been already largely used to benchmark single-cell multi-omics integrative methods [Hao et al., 2021, Lance et al., 2022, Luecken et al., 2022]. The chosen datasets span different sequencing technologies, modalities, tissues, and sizes: (i) *Liu* is a scCAT-seq cell lines dataset by [Liu et al., 2019] (ii) *PBMC 10X* is a 10X Multiome human PBMC dataset from 10X Genomics (iii) *OP Multiome* is a 10X Multiome human bone marrow dataset from Open Problems [Luecken et al., 2021] (iv) *OP CITE* is a CITE-seq human bone marrow dataset from Open Problems [Luecken et al., 2021] (v) *BM CITE* is a CITE-seq human bone marrow dataset from Stuart et al. [Stuart et al., 2019]. *BM CITE* is the larger dataset here considered, with 29,803 cells. Supplementary Table C.1 lists the modalities, numbers of cells, and numbers of cell types for each dataset. For details on data preprocessing, see Methods.

We benchmarked Mowgli, Seurat v4, MOFA+, integrative NMF, Cobolt, and Multigrate based on the same natural metrics used in the previous section. Since these metrics require a ground truth annotation, we used the cell-type annotations available from the original publications of these data. In Liu, the ground truth annotations are based on the cell line of origin and are thus well-defined. On the contrary, the annotations of the other datasets were computationally derived, thus affecting this benchmark for all methods. Supplementary Note C.1 illustrates this on specific subpopulations of CD8 T-cells in the *BM CITE* and *OP CITE* datasets.

As displayed in Figure 4.3B, Mowgli can handle large single-cell datasets and deliver embedding and clustering performances competitive with the state-of-the-art, especially considering the lack of absolute ground truth annotations on most datasets here employed.

In particular, according to the silhouette score, Mowgli outperforms other methods in Liu, *PBMC 10X*, and *OP CITE*. MOFA+ performs best in the other two datasets. In terms of ARI across resolutions,

Figure 4.3: Cell embedding and clustering benchmark in complex and heterogeneous datasets.



(A) Schematic representation of the benchmarking process; **(B)** The first three columns of this panel are devoted to silhouette scores, Adjusted Rand Indices (ARIs), and purity scores for the different methods on five complex paired single-cell multi-omics data already largely used to benchmark integrative methods. The following six columns provide UMAP visualizations for the six benchmarked methods (Mowgli, MOFA+, NMF, Seurat v4, Multigrate, Cobolt) on the same data. Different colors in these UMAP plots correspond to the different ground truth cell type annotations provided with the data.

Seurat v4 performs best in *PBMC 10X*, *OP Multiome*, and *OP CITE*. Mowgli and Seurat v4 perform comparably in the *BM CITE* dataset. In the *Liu* dataset, only MOFA+, Multigrate, and Mowgli reach a maximum ARI of 1. Of note, in Liu, ARIs should be compared only at low resolution, as higher resolutions lead to overclustering. In terms of purity score, Mowgli outperforms other methods in the *OP CITE* dataset, and it is comparable to MOFA+ and Multigrate in the *PBMC 10X* dataset. Finally, the purity scores of all methods are comparable in the *Liu* dataset, except for Cobolt which is less performant.

The UMAP plots in Figure 4.3B give a qualitative intuition of the described performance. In *OP CITE*, only integrative NMF, Cobolt, and Mowgli correctly separate subpopulations of B cells (Figure 4.3B circled). In *BM CITE*, MAIT T-cells and subpopulations of CD8+T-cells (Figure 4.3B circled) are more neatly separated in Seurat v4 than in other methods. However, as explained in Supplementary Note C.1, the annotation pipeline of *BM CITE* might favor Seurat v4.

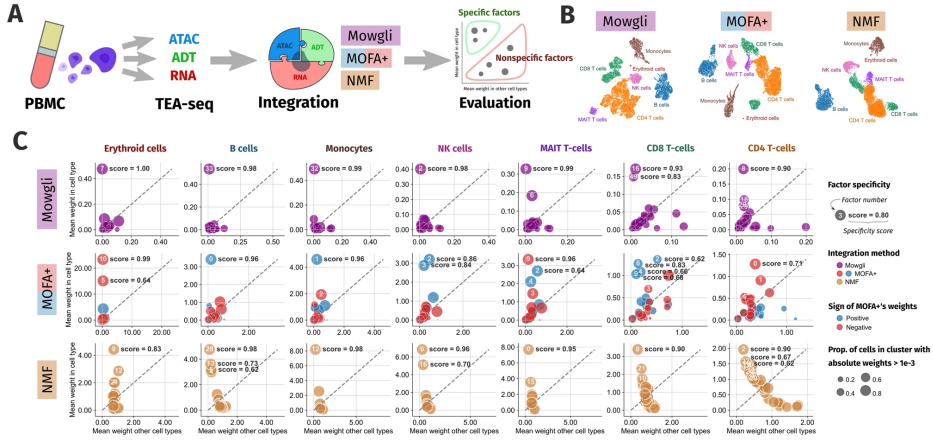
As in the previous section, we chose the overall best-performing number of latent dimensions for each method (see Methods). Tuning the number of latent dimensions for each metric and dataset

does not favor one method over the others, which supports Mowgli's competitiveness with the other methods (see Supplementary Figure C.1).

4.2.4 Mowgli improves the biological interpretability of the state-of-the-art by providing cell-type specific factors in TEA-seq data

We benchmarked Mowgli with respect to MOFA+ and integrative NMF based on its biological interpretability (see Figure 4.4A). Indeed, MOFA+ is the leading single-cell multi-omics integration tool providing user-friendly biological interpretability of its latent dimensions [Argelaguet et al., 2020]. At the same time, integrative NMF can be considered as a baseline with respect to Mowgli.

Figure 4.4: Evaluation of biological interpretability in TEA-seq data.



(A) Schematic representation of the evaluation process on biological interpretability; **(B)** UMAP visualization of Mowgli's, MOFA+'s, and integrative NMF's embeddings. The colors correspond to a marker-based cell-type annotation of the cells; **(C)** average weights within and outside of a cell type are plotted for each factor of Mowgli (violet), MOFA+ (red for the negative part and blue for the positive one), and integrative NMF (orange). For each cell type, the best specificity scores are reported in bold.

For this benchmark, we considered a TEA-seq dataset of human PBMCs, corresponding to the paired profiling of scRNA-seq, scATAC-seq, and surface proteins [Liu et al., 2019]. This dataset allows us to test the methods on more than two omic datasets, thus taking into account more complementary layers of molecular regulation.

First, MOFA+, integrative NMF, and Mowgli were independently applied for the integration of the three omics constituting the TEA-seq data. As the dataset was not provided with an annotation of the cells, we separately clustered the embeddings obtained from Mowgli, integrative NMF, and MOFA+ and annotated them based on gene and protein markers (see Supplementary Figure C.2, see Figure 4.4B). We identified in this way coarse immune cell types: CD4 T-cells, CD8 T-cells, B cells, Natural Killer (NK) cells, MAIT T-cells, Monocytes, and Erythroid cells. Of note, the cell type annotations

obtained with Mowgli, integrative NMF, and MOFA+ agree at 94% and match an independent RNA-based annotation obtained through Azimuth (see Supplementary Figure C.3). All three methods are thus able to recover the expected cell types through clustering of their embeddings.

To then test the biological interpretability of Mowgli, integrative NMF, and MOFA+, we evaluated the specificity of the associations between their factors and the identified immune cell types. The underlying assumption we are making here is that an interpretable method should provide factors that are not broadly active in all the cells, but selectively associated with a cell type. Indeed, characterizing a cell type that results from a combination of many factors is a daunting task. On the contrary, having cell type-specific factors makes the biological characterization of the associated cell type straightforward. To evaluate such specificity, for each cell type, we plotted how Mowgli, integrative NMF, and MOFA+'s factors are distributed according to their mean weight within the cell type and their mean weight outside the cell type (Figure 4.4C). Factors specific to a cell type should have a high average weight within the cell type and a low average weight outside the cell type, thus falling in the upper left corner of the plots. As MOFA+'s factors are not constrained to be positive and their positive and negative parts could be associated with different biological information, we split each factor into two parts, as done in MOFA+'s interpretation tools [Argelaguet et al., 2020]. In addition, we quantified the performance of each factor with a specificity score, also reported in bold in Figure 4.4C, and defined in the Methods section.

As shown in Figure 4.4C, while MOFA+ and integrative NMF tend to associate multiple factors to the same cell type, Mowgli frequently defines clear one-to-one associations between factors and cell types. In addition, the specificity score of such factors is higher in Mowgli than in MOFA+ and integrative NMF. This is particularly striking in NK cells, CD8 T-cells, and CD4 T-cells, where both MOFA+ and integrative NMF seem to aggregate information from many factors whereas Mowgli is more selective. Of note, as shown in Supplementary Figure C.4, the multiple factors associated by MOFA+ with the same cell type do not necessarily correspond to subpopulations of the same cell type.

As shown in Supplementary Figure C.5, our conclusions regarding the specificity of Mowgli's factors are robust to the choice of hyperparameters.

Overall, the comparison with integrative NMF validates that the use of OT as a loss function and of sparsity-inducing regularization introduced in Mowgli has a concrete effect on biological interpretability.

4.2.5 Mowgli identifies relevant subpopulations of immune cells in TEA-seq data

We finally focused on the biological relevance of the factors identified by Mowgli on the human PBMC TEA-seq data, described in the previous section. Indeed, while in the previous section we only considered coarse immune cell types (e.g. B cells, CD4 T-cells, CD8 T-cells), Mowgli could identify multiple factors able to subset such cell types into relevant subpopulations (see Figure 4.5A and Figure 4.5B; Supplementary Figure C.6; Supplementary Note C.2). For example, Mowgli identifies factors splitting the B cell cluster into two subpopulations: memory and naive B cells. In the same way, Mowgli detects factors associated with CD8 T-cells subpopulations (naive, central memory, and effector memory), monocytes subclusters (classical and non-classical), dendritic cells subpopulations (plasmacytoid and conventional) and Natural Killer (NK) subclusters (CD56dim and CD56bright). The association of the factors with specific immune subpopulations is here made based on top-ranked genes and proteins in effector memory CD8 T cells, naive B cells, memory B cells, and CD56dim NK cells. For all other populations, the association with factors is instead based on the correlation of the factors'

weights with that of known protein markers (see Supplementary Note C.2). Figure 4.5B displays side-by-side the UMAP plots showing the similarity between the distribution of the factors' weights and the activity of the protein markers of their associated immune subpopulations. The UMAP visualizations of all factors and marker proteins are available in Supplementary Figure C.6 and Supplementary Figure C.7.

These same results could not be obtained with MOFA+, due to its lower biological interpretability observed in the previous section. In MOFA+, factors having similar patterns to those observed in Mowgli could be obtained for effector memory CD8 T-cells, memory B cells, non-classical monocytes, and CD56dim NK cells (see Supplementary Figure C.4 and Supplementary Figure C.8). On the contrary, MOFA+'s factors most closely associated with the other immune subpopulations identified by Mowgli have less clear patterns (see Supplementary Note C.2). As a consequence, interpreting with MOFA+ the pathways associated with CD56bright NK cells, for example, would require complexly combining the pathway enrichments obtained from different factors. On the contrary, the same analysis in Mowgli can be easily realized by looking at the pathways enriched in the loadings of its 13th factor.

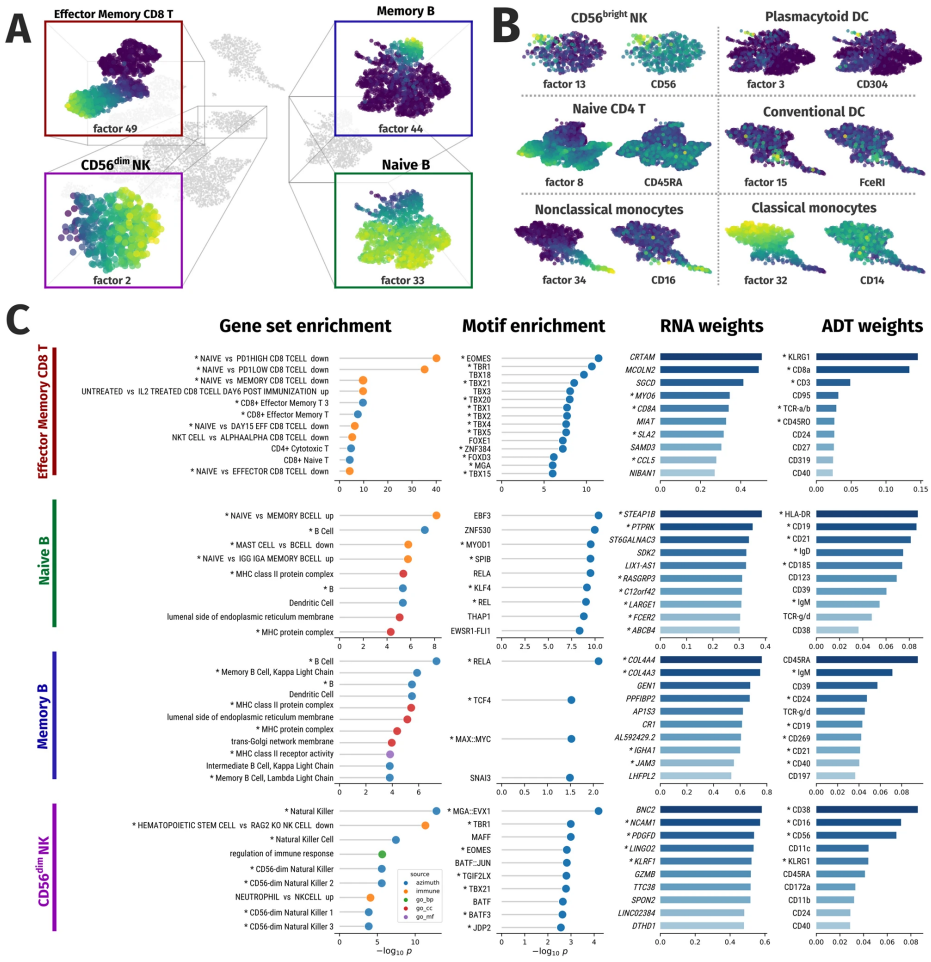
Finally, we looked at the biological information that Mowgli could provide regarding the identified immune subpopulations. For this part, we focused on the factors associated with four immune cell subpopulations: effector memory CD8 T-cells (factor 49), naive B cells (factor 33), memory B cells (factor 44), and CD56dim NK cells (factor 2). For each of these four factors, we considered their associated loadings in $\mathbf{H}^{(rna)}$, $\mathbf{H}^{(adt)}$ and $\mathbf{H}^{(atac)}$ and analyzed the top genes in $\mathbf{H}^{(rna)}$, top proteins in $\mathbf{H}^{(adt)}$, the gene sets enriched in $\mathbf{H}^{(rna)}$ and the motifs enriched in $\mathbf{H}^{(atac)}$ to verify the biological information that could be extracted from the output of Mowgli (see Methods). Figure 4.5C displays the results obtained from this analysis.

For effector memory CD8 T-cells (CD8 TEM cells), corresponding to factor 49, Mowgli could extract two top genes (*CRTAM* and *KLRK1*), known to be essential for CD8+T-cell-mediated cytotoxicity [Lanier, 2015, Boles et al., 2005], two top proteins (CD45RO, TCR-a/b) that are a known memory T cell marker and a T cell receptor, respectively [Prince et al., 1992, Shah et al., 2021]. More interestingly, also several Transcription Factors (TFs) candidate regulators of this subpopulation are identified, among them EOMES and TBX21 (aka T-bet), known to be important for CD8 TEM development [Intlekofer et al., 2005]. In addition, five of the top candidate TF regulators (TBR1, TBX21, TBX4, TBX5 and MGA) target three of the top genes of the same factor (*CCL5*, *CRTAM*, and *IL21R*), thus suggesting a regulatory program possibly important for CD8 TEM cells.

In naive B cells (factor 33), Mowgli identifies as top genes *FCER2* (aka *CD23*), a low-affinity receptor for immunoglobulin E (IgE) with an essential role in the differentiation of B-cells [Pirron et al., 1990] and *MARCH1*, which downregulates the surface expression of major histocompatibility complex (MHC) class II molecules [Bartee et al., 2004]. In the top proteins, we can single out CD19, CD21, and HLA-DR, well-known markers of B cells [Glass et al., 2020]. In addition, the relative weights of IgD and IgM in factor 33 are coherent with the repartition already described for naive B-cells [Glass et al., 2020]. Finally, among the top TF candidate regulators of factor 33, Early B-cell Factors (EBF3 and EBF1) [Lukin et al., 2008] and NF-kB proteins (REL and RELA) stand out as regulators of the top genes of the same factor. Of note, these TFs play an essential role in B-cell development, maintenance, and function [Kaileh and Sen, 2012].

For memory B cells (factor 44), Mowgli extracts as top genes: *IGHA1* and *IGKC*, part of immunoglobulin complexes [Schroeder Jr and Cavacini, 2010] and *JAM3*, belonging to the Immunoglobulin su-

Figure 4.5: Characterization of the immune cell subpopulations identified by Mowgli in TEA-seq data.



(A) UMAP visualization of Mowgli's embedding with focus on four specific immune subpopulations (Effector Memory CD8 T-cells, memory B cells, CD56dim NK cells, naive B cells) for which the UMAP is colored based on factor weights; **(B)** UMAP visualization of Mowgli's embedding colored by factor weight and protein marker weight for other factors corresponding to specific subpopulations of cells; **(C)** Top genes, proteins, gene sets and Transcription Factors (TFs) for the 4 factors visualized in panel a. Stars denote gene sets and markers pertinent for the immune subpopulation associated with the factor and TFs targeting the top genes.

perfamily and already studied in the context of B cell homing and development [Ody et al., 2007, Weber et al., 2007, Doñate et al., 2013]. The top proteins include the well-known B cell markers CD19, CD21, and HLA-DR [Glass et al., 2020]. In addition, as observed before for naive B cells, the relative weights of IgD and IgM in factor 44 are coherent with the repartition already described for memory B-

cells [Glass et al., 2020]. In the top TFs emerging from our motif analysis and targeting the top genes we finally find RELA, TCF4, and MAX::MYC, known to be involved in the transcriptional regulation of memory B cell differentiation [Laidlaw and Cyster, 2021].

Finally, in CD56dim NK cells (factor 2), Mowgli detects at top genes: *NCAM1* (aka *CD56*), the go-to marker for NK cells [Vivier et al., 2008]; *KLRF1* and *KLRD1*, genes of the KLR family of receptors controlling NK cell activity [Roda-Navarro et al., 2000]; *GZMB*, involved in NK-cell mediated cytotoxicity [Su et al., 1994]; *SLAMF7*, mediating NK cell activation [Guo et al., 2015a]. Top proteins include CD56, the canonical marker of NK cells [Vivier et al., 2008], but its weight is lower than that of CD16, which is coherent with the expression profile of CD16+CD56dim NK cells [Vivier et al., 2008]. Regarding TF candidate regulators, we detect EOMES and TBX21 (aka T-bet), which are critical to NK-cell differentiation [Zhang et al., 2021], Maf-F, having a key role in the regulation of NK cell effector functions by IL-27, and JUNB::FOSB, early activator protein (AP)−1 TFs that regulate NK-meditated cytotoxicity [Ponti et al., 2002, Bernard et al., 1999]. Finally, a strong regulatory program seems to emerge here with four of the top candidate TF regulators for factor 2 (MGA::EVX1, EOMES, TBX21, and JDP2) targeting four of the top genes of the same factor (*C1orf21*, *IL18RAP*, *PTGDR*, and *SLAMF7*).

4.3 Discussion

Multiple technologies allowing the multi-omics profiling of the same set of cells are currently available. We thus need integration methods able to jointly learn from multiple omics data profiled on the same cells.

In this article, we introduced Multi-Omics Wasserstein inteGrative anaLysIs (Mowgli), an integrative method for paired multi-omics data that enables rich biological interpretation for any type and number of omics. We then in-depth benchmark Mowgli's cell embedding and clustering performance with respect to the state-of-the-art in controlled settings derived from scCAT-seq profiling of cancer cell lines. Mowgli outperforms in this benchmark the state-of-the-art showing its high potential even in challenging conditions. We then considered more complex and heterogeneous data profiled with CITE-seq and 10x Genomics Multiome technologies. On these data, Mowgli performed comparably with the state-of-the-art, with no method clearly outperforming others. Finally, regarding the biological interpretability, once tested on TEA-seq data, corresponding to paired scRNA, scATAC, and surface protein profiling, Mowgli produces biologically meaningful representations superior to those of the state-of-the-art.

A major limitation affecting this benchmark and all others focused on paired multi-omics integration corresponds to the lack of a high-quality biological annotation of the cells. While in some cases Fluorescence-activated cell sorting (FACS) could represent a clear solution for an independent annotation of the cells, paired multi-omics data with this type of annotation are lacking in the literature.

A limitation of Mowgli is that it does not offer a straightforward approach to defining the number of latent dimensions. The choice of number of latent dimensions (d) is inherently problem-dependent and several values of d should be tested. This problem has been however extensively studied in NMF literature and classical tools like the cophenetic coefficient [Brunet et al., 2004] or the elbow method [Kotliar et al., 2019] may guide the user's choice of d. At the same time, Mowgli displays relative robustness to changes in thus suggesting that small changes in d will not affect its performance. In addition, OT is inherently expensive to compute compared to the Euclidean distance. But the entropic regularization of OT considered here is GPU-friendly, and GPU computations enable Mowgli to

scale to the larger datasets presented in this article. The availability of GPUs is nowadays a standard in research centers and this will be further enhanced in the future, once larger single-cell datasets will be available. Regarding possible extensions of Mowgli, it would be interesting to deal with batch correction once integrating paired multi-omics data. Indeed, most recent large-scale paired multi-omics data are profiled in different centers thus creating batch correction issues.

4.4 Methods

4.4.1 Notations

Let us consider n cells, measured across several modalities. Each modality p has m_p features (e.g. genes). Let us denote $\mathbf{A}^{(p)} \in \mathbb{R}_+^{m_p \times n}$ the dataset for modality p. Additionally, we impose each column of $\mathbf{A}^{(p)}$ to sum to 1, i.e. be a discrete probability distribution.

4.4.2 Optimal transport

The paragraphs below recall definitions presented in the introduction of this thesis (see Section 1.3.1) in the context of cell-cell distances, namely the Wasserstein distance between histograms (Equation 1.6) and its entropic regularization (Equation 1.7).

Optimal Transport (OT), as defined by [Monge, 1781] and [Kantorovich, 1942], aims at comparing two probability distributions $\mathbf{a}$ and $\mathbf{b}$ by computing the minimal cost of transporting one distribution to the other. In the discrete case, the classical OT distance, also known as the Wasserstein distance, between m-dimensional histograms $\mathbf{a} = (a_1, \dots, a_m)$ and $\mathbf{b} = (b_1, \dots, b_m)$ is defined as

$$\mathrm{W}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{k,l} P_{k,l} C_{k,l} \tag{4.2}$$

where $\mathcal{U}(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \in \mathbb{R}_+^{m \times m}$ such that $\sum_l P_{k,l} = a_k$ and $\sum_k P_{k,l} = b_l\}$.

In this discrete case, the coupling $\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})$ is a matrix that represents how the mass in the discrete probability distribution $\mathbf{a}$ is moved from one bin (e.g. gene) to another one in order to transform $\mathbf{a}$ into $\mathbf{b}$. In other words, $P_{k,l}$ is the amount of gene expression transported between genes k and l when transporting the gene expression profile of the cell $\mathbf{a}$ to the gene expression profile of the cell $\mathbf{b}$.

The ground cost $\mathbf{C} \in \mathbb{R}_+^{m \times m}$ is a pairwise distance matrix that encodes the penalty for transporting mass from one feature (e.g. gene) to another. Hence, $\mathbf{C}$ should be chosen in such a way that similar bins (e.g. genes) k and l have a low cost $C_{k,l}$. Indeed, this will favor transporting gene expression between similar genes. Here, for a certain omic p, we define $\mathbf{C}$ in a data-driven way as the matrix of pairwise cosine distances between the features, i.e. the rows in our dataset $\mathbf{A}^{(p)}$. In other words, denoting $\mathbf{x}_k \in \mathbb{R}_+^n$ and $\mathbf{x}_l \in \mathbb{R}_+^n$ two rows in our dataset,

$$C_{k,l} = 1 - \frac{\langle \mathbf{x}_k, \mathbf{x}_l \rangle}{\|\mathbf{x}_k\|_2 \|\mathbf{x}_l\|_2} \tag{4.3}$$

where $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$ is the dot product. This choice of ground cost gave the best results in our previous work [Huizing et al., 2022].

Due to the high dimensionality of single-cell data, we use an approximation of classical OT. The

entropic regularization of OT, described in (Equation 4.4) below is computed using the fast and GPU-enabled Sinkhorn algorithm [Cuturi, 2013].

$$\mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{k,l} P_{k,l} C_{k,l} - \varepsilon \mathrm{E}(\mathbf{P}) \tag{4.4}$$

where the entropy E is defined as $\mathrm{E} : \mathbf{X} \in \mathbb{R}_{+}^{m \times m} \mapsto -\sum_{k,l} X_{k,l} \left(\log X_{k,l} - 1\right)$.

If ε is set to zero, (Equation 4.4) corresponds exactly to classical OT (Equation 4.2). Increasing values of correspond to a more diffused coupling $\mathbf{P}$. In previous work, we showed the entropic regularization of OT to improve similarity inference between single-cell omics profiles compared to classical notions of distance [Huizing et al., 2022].

As explored in that work [Huizing et al., 2022], entropic regularization is expected to control the systematic noise due to technical dropouts and to the stochasticity of gene expression at the single-cell level. In addition, more diffused couplings increase the exchange of mass between features. This enables OT to leverage the relationships between features (e.g. genes), motivating further its application to single-cell data.

4.4.3 Mowgli

We aim to decompose each matrix $\mathbf{A}^{(p)}$ as the product of a matrix $\mathbf{H}^{(p)} \in \mathbb{R}_{+}^{m_p \times d}$ (the modality-specific dictionaries) and $\mathbf{W} \in \mathbb{R}_{+}^{d \times n}$ (the embeddings, shared across modalities). The integer d is the number of dimensions of the latent space and should be small compared to the number of features. We use the entropic regularization of OT as a reconstruction loss to compare $\mathbf{H}^{(p)}\mathbf{W}$ to the reference data $\mathbf{A}^{(p)}$.

In addition, we require the columns of $\mathbf{H}^{(p)}\mathbf{W}$ to sum to 1, i.e. belong to the simplex. We thus impose that the columns of $\mathbf{H}^{(p)}$ sum to one, and that the columns of $\mathbf{W}$ sum to one. Following [Rolet et al., 2016], we use the entropy function E defined previously, with a value of $-\infty$ when columns do not sum to one.

Combining the reconstruction and the entropy terms yields the loss

$$\sum_{p} \left(\sum_{j} \mathrm{W}_{\mathbf{C}^{(p)}}^{\varepsilon} \left(\mathbf{H}^{(p)} \mathbf{w}_j, \mathbf{a}_j^{(p)} \right) - \alpha_p \mathrm{E} \left(\mathbf{H}^{(p)} \right) - \beta \mathrm{E} \left(\mathbf{W} \right) \right). \tag{4.5}$$

The parameters α_p, control the sparsity of the columns of $\mathbf{H}^{(p)}$ and $\mathbf{W}$. In order to make these parameters more comparable across omics and datasets, we define

$$\alpha_p = \frac{1}{\log m_p \times d} \tilde{\alpha}_p \text{ and } \beta = \frac{1}{\log d \times n} \tilde{\beta}. \tag{4.6}$$

We have set the values of $\tilde{\alpha}_{rna} = 0.01$, $\tilde{\alpha}_{adt} = 0.01$, $\tilde{\alpha}_{atac} = 0.1$ and $\tilde{\beta} = 0.001$ since they performed best across datasets and metrics (silhouette score, ARI, purity score) (see Supplementary Figure C.9, C.10, and C.11). Regarding the choice of number of latent dimensions, for *Liu* and datasets derived from Liu, we run the method with 5 factors. For other datasets, we choose 50 factors. These parameters gave the best results overall (see Supplementary Figure C.12).

Similarly to [Rolet et al., 2016], we alternate between minimizing (Equation 4.5) on $\mathbf{H}^{(p)}$ and $\mathbf{W}$. One can show that these smooth minimization problems on $\mathbf{H}^{(p)}$ and $\mathbf{W}$ are equivalent to the following smooth minimization problems on new dual variables $\mathbf{G}^{(p)}$. These problems can be solved using standard optimization methods, and the method of choice is L-BFGS, a limited-memory quasi-Newton method.

- Optimizing $\mathbf{H}^{(p)}$. We solve the following smooth minimization problem:

$$\min_{\mathbf{G}^{(p)}} \sum_j \left(\left(\mathrm{W}^{\varepsilon}_{\mathbf{C}^{(p)}} \right)^{\star} \left(\mathbf{g}_j^{(p)}, \mathbf{a}_j^{(p)} \right) \right) - \alpha_p \times (-\mathrm{E})^{\star} \left(-\mathbf{G}^{(p)} \mathbf{W}^{\top} / \alpha_p \right) \tag{4.7}$$

  Then, we update the primal variable as follows:

$$\mathbf{H}^{(p)} = \mathrm{softmax} \left(-\mathbf{G}^{(p)} \mathbf{W}^{\top} / \alpha_p \right) \tag{4.8}$$

  The column-wise softmax is defined as:

$$\mathrm{softmax} : \mathbf{X} \in \mathbb{R}^{m \times n} \mapsto \frac{\exp\left(X_{i,j}\right)}{\sum_i \exp\left(X_{i,j}\right)} \tag{4.9}$$

- Optimizing $\mathbf{W}$. We solve the following smooth minimization problem:

$$\min_{\mathbf{G}^{(p)}} \sum_p \left(\sum_j \left(\mathrm{W}^{\varepsilon}_{\mathbf{C}^{(p)}} \right)^{\star} \left(\mathbf{g}_j^{(p)}, \mathbf{a}_j^{(p)} \right) \right) - D \times \beta \times (-\mathrm{E})^{\star} \left(-\frac{1}{D \times \beta} \sum_p {\mathbf{H}^{(p)}}^{\top} \mathbf{G}^{(p)} \right) \tag{4.10}$$

  Then, we update the primal variable as follows:

$$\mathbf{W} = \mathrm{softmax} \left(-\frac{1}{D \times \beta} \sum\nolimits_p {\mathbf{H}^{(p)}}^{\top} \mathbf{G}^{(p)} \right) \tag{4.11}$$

Here, $\left(\mathrm{W}^{\varepsilon}_{\mathbf{C}^{(p)}} \right)^{\star}$ and $(-\mathrm{E})^{\star}$ denote the Legendre duals of the $\mathrm{W}^{\varepsilon}_{\mathbf{C}^{(p)}}$ and $-\mathrm{E}$ functions, and their smooth closed form expressions are defined in [Rolet et al., 2016].

An important property of NMF enabling representation by parts is sparsity, but it is not enforced explicitly [Hoyer, 2004]. In classical NMF, the L1 norm can be leveraged to explicitly enforce sparsity [Hoyer, 2004, Le Roux et al., 2015]. In our setting, the simplex constraint renders the L1 penalty ineffective (because it is equal to 1), but we can leverage the entropic penalty of [Rolet et al., 2016] to control sparsity. Indeed, the coefficients α_p and β parametrize *softmax* functions, and hence control the sparsity of distributions. When setting α_p (or β) to zero, the solutions are Dirac masses (extreme sparsity) while when setting it to $+\infty$ the solutions are uniform (no sparsity).

The code is implemented in Python and relies on PyTorch [Paszke et al., 2019] for matrix operations on the GPU and on [Bredikhin et al., 2022] and [Wolf et al., 2018] to handle single-cell multimodal data. The compared running times of Mowgli and other methods is reported in Supplementary Table C.2.

4.4.4 Benchmarked methods

MOFA+ We compare Mowgli to [Argelaguet et al., 2020], a variational inference method analogous to sparse PCA for multi-omic data. We use the R interface MOFA2 with default training parameters. MOFA+ provides a parameter drop_factor_threshold designed to keep only informative factors, but we found that in practice it removed important information. For example, the benchmark in Zuo and Chen18 only kept one factor for [Argelaguet et al., 2020], which is not enough to represent cellular heterogeneity in the data. We thus choose to keep 5 factors for *Liu* and the datasets derived from Liu, and 15 factors for the other datasets. These parameters gave the best results overall (see Supplementary Figure C.13).

Seurat v4 We compare Mowgli to [Hao et al., 2021] which uses Weighted Nearest Neighbors to integrate multi-omics data. We use the R interface Seurat with default parameters.

Cobolt We compare Mowgli to [Gong et al., 2021], a deep neural network approach to integrate multi-omics data. We use the Python package cobolt with a learning rate of 0.001 and 200 epochs. As suggested in the documentation, we use raw counts as input for the RNA and ATAC modalities. For *Liu* and the datasets derived from Liu, we use 5 latent dimensions, and for the other datasets, we use 30 latent dimensions. These parameters gave the best results overall (see Supplementary Figure C.14).

Multigrate We compare Mowgli to [Lotfollahi et al., 2022], another deep neural network approach to integrate multi-omics data. We use the Python package multigrate with 200 epochs. As suggested in the documentation, we use raw counts and a negative binomial loss for RNA and processed data with a mean squared error for ATAC and ADT. For *Liu* and the datasets derived from Liu, we use 15 latent dimensions, and for the other datasets, we use 50 latent dimensions. These parameters gave the best results overall (see Supplementary Figure C.15).

Integrative NMF We implemented a baseline NMF-based integration method by concatenating the features from the different omics and solving the optimization problem with positivity constraints:

$$\arg \min_{\mathbf{H} \in \mathbb{R}_+^{m \times d}, \mathbf{W} \in \mathbb{R}_+^{d \times n}} \|\mathbf{A} - \mathbf{HW}\|_2. \tag{4.12}$$

We implemented this approach using the TorchNMF package. For *Liu* and the datasets derived from Liu, we use 5 latent dimensions, and for the other datasets, we use 30 latent dimensions. These parameters gave the best results overall (see Supplementary Figure C.16).

Note that this is almost equivalent to [Chalise and Fridley, 2017] with $\theta = 1$, which minimizes instead $\sum_p \|\mathbf{A}^{(p)} - \mathbf{H}^{(p)}\mathbf{W}\|_2$. However, on the considered datasets, the intNMF package was too slow to be able to include it in the benchmark.

4.4.5 Choice of the number of latent dimensions for all methods

For each method described above, we selected the number of latent dimensions displaying the overall best performances across evaluation metrics (Silhouette score, ARI, purity score) and datasets. This analysis has been done considering *Liu* and the controlled settings separately from the other datasets (*PBMC 10X, OP Multiome, OP CITE, BM CITE*). Indeed, Liu, composed only of three cell lines, presents much less variation than the other real datasets.

4.4.6 Data generation

Mixed in RNA We simulate a dataset where one modality confuses two populations, while the other can separate them. To do so we replace the RNA profiles of all HCT cells with RNA profiles of random HeLa cells. ATAC profiles are left untouched.

Mixed in both We simulate a dataset where the two modalities each confuse two cell populations, but separate two others. This makes the two omics complementary. To do so we replace the RNA profiles of all HCT cells with RNA profiles of random HeLa cells. Then, we replace the ATAC profiles of all K562 cells with ATAC profiles of random HeLa cells.

Sparse We simulate high dropout noise by randomly replacing 50%, 70% or 90% of the values with zeros. Since the data is already 65% sparse, the final sparsity is 82%, 90% and 96%.

Rare population We simulate the presence of a rare population by keeping only 10 randomly chosen HeLa cells.

4.4.7 Data preprocessing

All preprocessing was performed using the Scanpy [Wolf et al., 2018] and Muon [Bredikhin et al., 2022] Python packages.

RNA preprocessing Quality control filtering of cells was performed on the proportion of mitochondrial gene expression, the number of expressed genes, and the total number of counts (using Muon's filter_obs). Quality control filtering of genes was performed on the number of cells expressing the gene (using Muon's filter_var). Cells were normalized to sum to 10000 (using Scanpy's normalize_total), then log-transformed (using Scanpy's log1p). The top 2500 most variable genes (1500 for the *Liu* dataset) were selected for downstream analysis (using Scanpy's highly_variable_genes with flavor='seurat').

ATAC preprocessing Quality control filtering of cells was performed on the number of open peaks and the total number of counts (using Muon's filter_obs). Quality control filtering of peaks was performed on the number of cells where the peak is open (using Muon's filter_var). In Liu, TEA, and 10X PBMC, cells were normalized to sum to 10000 (using Scanpy's normalize_total), then log-transformed (using Scanpy's log1p). In *OP Multiome*, cells were normalized using TF-IDF (using Muon's tfidf) to follow the preprocessing chosen by its authors. The most variable peaks were selected for downstream analysis (using Scanpy's highly_variable_genes with flavor='seurat'). Due to differences in the data's distribution across datasets, we chose to keep 1500 peaks in Liu, 5,000 peaks in PBMC, and 15,000 peaks in *OP Multiome* and TEA.

ADT preprocessing Since the number of proteins is small and the data is less noisy than RNA or ATAC, no quality control or feature selection was performed. The data was normalized by Center Log Ratio (using Muon's clr).

4.4.8 Data analysis

Gene set enrichment analysis (GSEA) The gProfiler API83 was used through Scanpy's enrich. Custom sources GO:CC, GO: MF, GO: BP, Azimuth, and ImmuneSigDB were retrieved from the Enrichr website [Chen et al., 2013]. Gene sets enriched with adjusted p-values under 0.05 (with Bonferroni correction) were selected for further analysis. To make genes comparable, we normalized rows of the matrix $\mathbf{H}^{(rna)}$ to 1. The 150 top genes for every factor were then used as an unordered input to gProfiler.

Motif enrichment analysis [Stuart et al., 2021] was used to perform Motif Enrichment Analysis, using the JASPAR2022 motif database [Castro-Mondragon et al., 2022]. To make peaks comparable, we normalized rows of the matrix $\mathbf{H}^{(atac)}$ to 1. The 100 top peaks for every factor were used as input to Signac's FindMotifs. The union of the top peaks across factors constitutes the background.

Visualization To visualize the latent representation of cells in MOFA+, integrative NMF, and Mowgli's models, we computed kNN graphs ($k = 20$) with the euclidean distance between the cells' low-dimensional embeddings (using Scanpy's neighbors). We used these graphs to compute 2-D UMAP [McInnes et al., 2018] projections (using Scanpy's umap). For Seurat v4, 2-D UMAP projections based on WNN graphs were performed using Seurat v4's function RunUMAP.

Clustering For Mowgli, integrative NMF, and MOFA+, we clustered datasets using the Leiden algorithm [Traag et al., 2019] with varying resolutions (using Scanpy's leiden). Similarly to UMAP visualization, the inputs of the Leiden algorithm were the previously computed kNN graphs. For Seurat v4, Leiden clustering was performed using Seurat v4's function FindClusters.

4.4.9 Evaluation metrics

Silhouette score For each sample, the silhouette width is defined as $\frac{\text{dist}_{\text{inter}} - \text{dist}_{\text{intra}}}{\max(\text{dist}_{\text{intra}}, \text{dist}_{\text{inter}})}$ where $\text{dist}_{\text{intra}}$ is the mean distance of the sample to other samples of the same cluster and $\text{dist}_{\text{inter}}$ is the mean distance of the samples to samples from the nearest cluster. The silhouette score is the mean of silhouette widths across samples. The silhouette score varies between -1 and 1. We used Scikit-learn's implementation silhouette_score [Pedregosa et al., 2011].

kNN purity score The kNN purity score measures the average proportion of a sample's nearest neighbors that share the sample's cluster annotation. It thus varies between 0 and 1.

Adjusted rand index The Rand Index defines the similarity between a ground truth annotation and an experimental clustering. The ARI is then defined as

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}\left(\text{RI}\right)}{\max\left(\text{RI}\right) - \mathbb{E}\left(\text{RI}\right)} \tag{4.13}$$

and varies between -1 and 1, with 0 representing a random clustering. We used Scikit-learn's implementation adjusted_rand_score. Since the simulated datasets are derived from three cell lines, there is no biological heterogeneity within each of the three groups and clustering at high resolution necessarily leads to overclustering. Figure 4.2 thus displays maximum ARI, which is achieved for low resolutions and is more informative than the ARI across resolutions. ARI across resolutions can be found in Supplementary Figure C.9 to Supplementary Figure C.16.

Specificity MOFA+ was applied with 15 factors, which are enough to represent the data (see Supplementary Figure C.13). Integrative NMF was applied with 30 factors. Mowgli was applied with 50 factors. For all three methods, a coarse Leiden clustering was applied (using Scanpy's leiden). For all three methods, each cluster was assigned a cell type based on the expression of the canonical gene and protein markers (see Supplementary Figure C.2). To confirm this annotation, Azimuth was run on the RNA signal of the dataset (using the Azimuth web tool and the PBMC reference). The agreement of the four independent annotations is confirmed in a Sankey diagram (see Supplementary Figure C.3). Dendritic cells are absent in our manual annotations because of the coarseness of the clustering. Likewise, the ADT signal (see Supplementary Figure C.7) informs us that there is a CD4-CD8- T cell population missed in all four annotations. For each cell type, and for each factor i, we computed (i) the proportion of cells within that cell type with an absolute weight greater than 10^{-3} (ii) u_i, the mean weight for cells within that cell type (iii) v_i, the mean weight for cells outside of that cell type. For each cell type, we then defined a specificity score for factor i as

$$\text{specificity}_i = \frac{u_i - v_i}{\max_j u_j} \tag{4.14}$$

where $\max_j u_j$ is the maximum value of u_j computed over all factors j. The specificity score is thus bounded by 1. See Figure 4.4C for a visualization of this information and see Supplementary Figure C.5 for results across hyperparameters β and numbers of latent dimensions.

4.4.10 Biological interpretation

We added stars in front of biologically interesting elements in Figure 4.5C. The first resource we used was the Human Protein Atlas, from which we programmatically retrieved information about the top proteins and genes. We starred them if they were marked as specific to NK cells, Naive B cells, Memory B cells, or Memory CD8 T cells respectively. In addition, some genes or proteins were starred manually; we discuss those in the Results and refer to the relevant literature.

We starred gene sets if they matched the considered cell types, e.g. MHC II protein complex for B cells. To reduce the noise in the Immune gene sets, we only considered gene sets opposing subtypes of the broad cell type considered, e.g. NAIVE_VS_MEMORY_BCELL_UP.

We starred the TFs if they target one of the top 20 genes. For this, we retrieved TF-gene links from the Regulatory Circuits database [Marbach et al., 2016] and considered the Natural Killer cells, CD19+B cells, and CD8+T cells networks.

4.4.11 Data availability

Regulatory Circuits At the time of writing, the Regulatory Circuits website ww1.regulatorycircuits.org/ is down. We recovered the data from the mirror www2.unil.ch/cbg/regulatorycircuits/FANTOM5_individual_networks.tar.

PBMC We retrieve a 10X Genomics Multiome (RNA+ATAC) dataset with 9320 PBMCs. Data is available at www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0.

Liu We retrieve a scCAT-seq (RNA+ATAC) dataset from [Liu et al., 2019] with 206 cells from three cancer cell lines (HCT116, HeLa-S3, K562). Data is available in the supplementary materials of the original publication.

Simulated data Controlled settings derived from cell lines data were generated from the *Liu* dataset and can be reproduced using the provided reproducibility code (see Code availability).

OP-Multiome We retrieve a Multiome (RNA+ATAC) dataset from the Open Problems challenge [Luecken et al., 2021] and select only the first batch, which contains 6137 BMMCs. The GEO accession number is GSE194122.

OP-CITE We retrieve a CITE-seq (RNA+ADT) dataset from the Open Problems challenge [Luecken et al., 2021] and select only the first batch, which contains 4249 BMMCs. The GEO accession number is GSE194122.

BM-CITE We retrieve a CITE-seq (RNA+ADT) dataset from Stuart et al. [Stuart and Satija, 2019] with 29,803 BMMCs. The GEO accession number is GSE128639.

PBMC TEA-seq We retrieve a recent TEA-seq (RNA+ATAC+ADT) dataset from Swanson et al. [Swanson et al., 2021] with 7084 PBMCs. The GEO accession number is GSE158013.

4.4.12 Code availability

Package The Python package for Mowgli is hosted at github.com/cantinilab/mowgli/. It can be installed easily by running `pip install mowgli`.

Reproducibility Code to reproduce the experiments and figures is available at github.com/cantinilab/mowgli_reproducibility/.

Chapter 5

Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein

Abstract

Spatial transcriptomics is revolutionizing the study of many biological problems. It has provided insights into the role of spatial organization and cellular interactions in contexts such as neurodegenerative diseases and tumor microenvironments. Spatial transcriptomics is also crucial to discovering the mechanisms underlying spatial organization in dynamic biological processes such as development. Thus, inferring cell fate trajectories from spatial transcriptomics profiled at several time points has emerged as a critical goal for computational biology. A promising framework for analyzing sequencing data across time is Wasserstein gradient flow learning. It provides a causal model of differentiation built around a neural network representing the differentiation potential. However, existing gradient flow learning methods are ill-equipped to handle spatial information and cannot analyze spatially resolved transcriptomic data.

Here, we propose a novel method to learn a spatially informed potential. We employ Fused Gromov-Wasserstein, an extension of Optimal Transport (OT) that is invariant to rotations and translations in the spatial coordinates. We benchmark our approach using three large Stereo-seq spatiotemporal atlases, encompassing developmental and regenerative processes across different organisms: mice, zebrafish, and axolotls. We demonstrate superior spatial coherence compared to prevailing approaches which use linear OT. Finally, we provide an in-depth analysis of axolotl neural regeneration and mouse gliogenesis using our proposed method. In these contexts, we recover gene trends for known markers, such as *Nptx1* in Nptx+ excitatory neuron regeneration and *Aldh1l1* in gliogenesis, and additional putative drivers of cellular differentiation.

The contents of this chapter are under preparation for an article submission.

Contents

5.1 Introduction 87
5.2 Results 89
5.2.1 STORIES: a new single-cell trajectory inference method for spatial transcriptomics profiled through time 89
5.2.2 STORIES's predictions are more biologically relevant than the state-of-the-art 91
5.2.3 STORIES identifies trajectories and potential drivers of neuron regeneration in axolotls 94
5.2.4 STORIES identifies trajectories and potential drivers of gliogenesis in the developing mouse dorsal midbrain 95
5.3 Discussion 97
5.4 Methods 99
5.4.1 Data collection 99
5.4.2 Preprocessing 100
5.4.3 Wasserstein gradient flow learning with a quadratic objective 100
5.4.4 Computational Optimal Transport 103
5.4.5 Neural network architecture 103
5.4.6 Neural network training 104
5.4.7 Computional growth rate 104
5.4.8 Evaluation 105
5.4.9 Trajectory inference 106

5.1 Introduction

Spatial transcriptomics technologies are revolutionizing the study of the spatial organization of cells within tissues [Palla et al., 2022]. Techniques based on high-throughput sequencing have enabled the unbiased discovery of gene expression patterns within their spatial context. For instance, recent studies have revealed previously unknown spatial organization at the tumor-microenvironment interface in melanoma [Hunter et al., 2021] and Alzheimer's disease amyloid plaque microenvironment [Chen et al., 2020]. The most widely used spatially-resolved sequencing techniques (e.g. 10X Visium) measure spots larger than the typical cell size. However, recent technological developments based on barcoded arrays like Stereo-seq and HDST have reached single-cell resolution, effectively bridging functional and structural characterizations of the cell [Chen et al., 2022a, Vickovic et al., 2019]. Recent works have leveraged Stereo-seq to produce large spatiotemporal atlases of various biological processes by profiling a system with spatial transcriptomics at several points in time [Chen et al., 2022a, Liu et al., 2022, Wei et al., 2022]. These datasets are ideal for studying cellular dynamics within the tissue during processes such as development and the onset of complex diseases, where cells undergo coordinated transcriptomic changes and spatial reorganization.

Inferring the dynamics of biological processes from single-cell sequencing data requires tailored computational approaches known as trajectory inference methods [Saelens et al., 2019]. Monocle initiated the field of trajectory inference by ordering cells along a pseudotime axis based on their transcriptomic similarities [Trapnell et al., 2014]. While pseudotime represents the progression along a

differentiation process, pseudotime-based methods do not provide a model for the underlying transcriptomic changes, and thus cannot predict a cell's future transcriptomic state [Tritschler et al., 2019]. RNA velocity has thus been proposed to predict local changes in gene expression based on splicing dynamics [La Manno et al., 2018]. However, RNA velocity-based methods do not provide a straightforward notion of pseudotime and rely on simple kinetic models that can misinterpret cell dynamics, for instance in the case of transient boosts in transcription [Bergen et al., 2021].

When several time points are available, one can instead take advantage of methods based on Optimal Transport (OT). Several OT-based methods have been formulated to infer trajectories from snapshots along differentiation. Waddington OT [Schiebinger et al., 2019] infers trajectories by computing probabilistic cell-cell transitions between adjacent time points. However, it delivers neither a notion of pseudotime nor a notion of velocity. Another class of OT-based methods proposes a continuous model of population dynamics by training neural networks representing a generalized notion of velocity [Tong et al., 2020, Huguet et al., 2022], but provides no straightforward notion of pseudotime. A promising OT-based framework for trajectory inference consists in learning a potential function governing a causal model of differentiation [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022]. Framing cellular differentiation as the minimization of a potential function is rooted in systems biology [Huang et al., 2007] and formalizes Waddington's idea of epigenetic landscape [Allen, 2015]. Furthermore, the potential function is a natural alternative to pseudotime, and its gradient yields a rigorous notion of velocity.

OT-based approaches for trajectory inference using spatial transcriptomics through time have recently been developed. For instance, Moscot [Klein et al., 2023] computes cell-cell transitions between adjacent time points using an extension of OT called Fused Gromov-Wasserstein (FGW) [Vayer et al., 2018b], and stVCR [Peng et al., 2024] learns a spatial velocity along with a gene expression velocity. However, these methods do not provide an intuitive ordering of cells along differentiation, limiting their interpretability.

Here, we propose STORIES, a novel trajectory inference method capable of learning a causal model of cellular differentiation from spatial transcriptomics through time using FGW. Unlike Moscot, which uses FGW to connect adjacent time points, we use FGW as a machine learning loss to train a potential function capable of defining each cell's stage of differentiation and capable of predicting its future transcriptomic state. We benchmarked our approach on three large-scale spatiotemporal Stereo-seq atlases, covering mouse development [Chen et al., 2022a], zebrafish development [Liu et al., 2022], and axolotl regeneration [Wei et al., 2022]. Furthermore, we used STORIES for the in-depth analysis of cellular trajectories in axolotl neural regeneration and mouse gliogenesis. We recover gene trends for known markers, such as Nptx1 in Nptx+ excitatory neuron regeneration and Aldh1l1 in gliogenesis. In addition, STORIES uncovers other possible driver genes and transcriptional regulators of cellular differentiation in these contexts, which may be of interest for further biological investigation.

Finally, we provide STORIES as an open-source and user-friendly Python package (https://github.com/cantinilab/stories). It is based on the Scverse ecosystem [Virshup et al., 2023], making it easy to interface STORIES with existing tools for single-cell analysis such as Scanpy and Cellrank [Wolf et al., 2018, Weiler et al., 2023]. In addition, STORIES benefits from the JAX ecosystem for deep learning and OT computation [DeepMind et al., 2020, Cuturi et al., 2022], enabling the fast handling of large datasets.

5.2 Results

5.2.1 STORIES: a new single-cell trajectory inference method for spatial transcriptomics profiled through time

We developed SpatioTemporal Omics eneRgIES (STORIES), a novel tool for single-cell trajectory inference using omics data profiled through spatial and temporal dimensions (https://github.com/cantinilab/stories). STORIES allows studying dynamic biological processes in their spatial context by identifying cell fates, gene trends, and candidate transcriptional regulators (see Figure 5.1).

STORIES is based on the theory of Optimal Transport (OT), a mathematical framework that enables the geometrically meaningful comparison of distributions, using various flavors of the Wasserstein distance [Peyré et al., 2019, Villani et al., 2009]. OT also provides a valuable model for population dynamics: the so-called Wasserstein gradient flows, popularized by Jordan, Kinderlehrer and Otto for their connection with the Fokker-Planck equation [Jordan et al., 1998], and were recently used for trajectory inference in single-cell transcriptomics [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022]. However, existing methods for trajectory inference based on Wasserstein gradient flows are not equipped to deal with spatially-resolved omics data. STORIES introduces key methodological innovations that allow one to address the specific challenges of including spatial information.

As an input, STORIES takes slices of spatial transcriptomics profiled at several time points. For instance, Figure 5.1A displays sections of axolotl brains profiled at different stages during regeneration. STORIES then learns the parameters θ of a neural network J_θ, which assigns a differentiation potential to each cell according to its gene expression profile $\mathbf{x}$ (see Figure 5.1B). The function J_θ formalizes the Waddington epigenetic landscape [Allen, 2015], where undifferentiated cells have a high potential and, as they differentiate, move towards low-potential transcriptomic states, which correspond to mature cell types. The transition to these low-potential attractor states defines a causal model of cellular dynamics capable of predicting future gene expression patterns and suggesting potential driver genes and transcriptional regulators (see Figure 5.1C).

STORIES's potential-based approach provides two interpretable and biologically meaningful outputs: (i) the potential $J_\theta(\mathbf{x})$ naturally orders cells $\mathbf{x}$ along a differentiation process (ii) the vector $-\nabla_{\mathbf{x}} J_\theta(\mathbf{x})$ gives the direction of the evolution of gene expression. On the contrary, pseudotime-based methods [Trapnell et al., 2014, Street et al., 2018] focus only on the first aspect and velocity-based [La Manno et al., 2018, Bergen et al., 2020] methods focus only on the second. Crucially, STORIES also innovates compared to state-of-the-art potential-based methods [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022] by enabling the use of spatial coordinates.

Briefly, STORIES trains the neural network J_θ by predicting a distribution $\rho_{t_k}(\theta)$ of gene expression profiles for each time point t_k where $k \in [1, \dots, K]$. These distributions are then compared to the ground-truth distributions μ_{t_k}, and the parameters θ are updated to improve the predictions. Unlike existing potential-based methods, STORIES allows one to take into account the spatial coordinates of cells when comparing the distributions of gene expression.

Formally, let $\mu_t = \sum_i a_i \delta_{(\mathbf{x}_i, \mathbf{r}_i)}$ denote the empirical distribution of cells at time t, characterized by their gene expression profile $\mathbf{x}_i \in \mathbb{R}^d$, spatial coordinates $\mathbf{r}_i \in \mathbb{R}^2$ and weight $a_i \in \mathbb{R}_+$ where $\sum_i a_i = 1$. Similarly, let us denote $\rho_t(\theta) = \sum_j b_j \delta_{(\mathbf{y}_j, \mathbf{s}_j)}$ the predictions of STORIES at time t. Unlike the gene expression profiles $\mathbf{x}_i, \mathbf{y}_j$, the spatial coordinates $\mathbf{r}_i, \mathbf{s}_j$ are not directly comparable because

the slices are not necessarily aligned between time points. In other words, the spatial coordinates $\mathbf{r}_i, \mathbf{s}_j$ are defined up to an isometry (e.g., a rotation or translation).

Existing potential-based methods train the neural network using a linear OT objective, which is sensitive to isometries. Our approach instead uses a recently developed quadratic extension of OT called Fused Gromov-Wasserstein (FGW) [Vayer et al., 2018b], which renders the model invariant to spatial isometries. The FGW distance, defined below and explained more thoroughly in the Methods section, allows to compare the distributions μ_t and ρ_t directly on gene expression profiles, and up to an isometry on spatial coordinates.

$$\mathcal{FGW}^{\varepsilon}_{\alpha}(\mu_t, \rho_t) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} (1-\alpha)\mathrm{L}(\mathbf{P}) + \alpha \mathrm{Q}(\mathbf{P}) - \varepsilon \mathrm{E}(\mathbf{P}) \tag{5.1}$$

FGW seeks a matrix $\mathbf{P}$ mapping cells from μ_t to ρ_t such that $\mathbf{P}$ minimizes the sum of three terms: (i) the linear term L compares the gene expression coordinates $\mathbf{x}_i, \mathbf{y}_j$ (ii) the quadratic term Q compares pairwise distances $d(\mathbf{r}_i, \mathbf{r}_{i'})$ and $d(\mathbf{s}_j, \mathbf{s}_{j'})$, which are not affected by translating or rotating the tissue (iii) an entropic regularization term E. The parameter $\alpha \in [0, 1[$ denotes the relative weight of spatial information.

Our proposed objective function evaluates the predictions across all timepoints using a debiased version of the FGW distance denoted $\overline{\mathcal{FGW}^{\varepsilon}_{\alpha}}$ (see Methods):

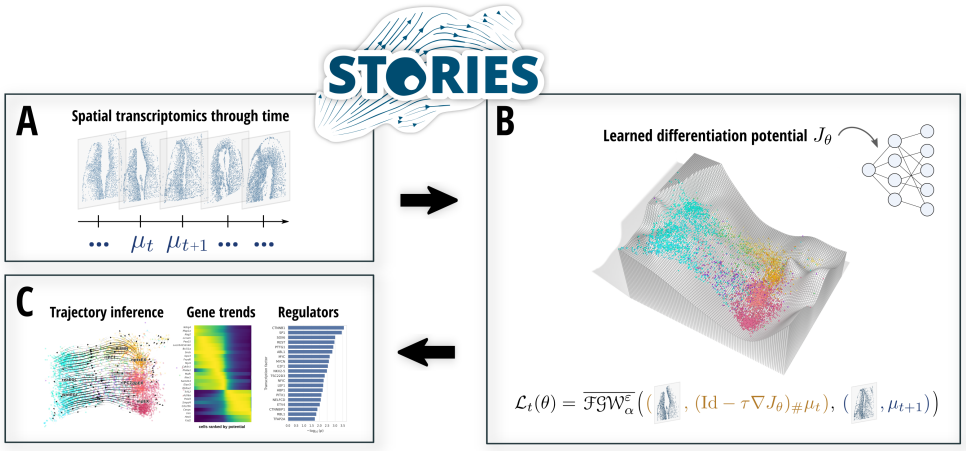
$$\mathcal{L}(\theta) = \sum_{k=0}^{K-1} (t_{k+1} - t_k) \overline{\mathcal{FGW}^{\varepsilon}_{\alpha}}(\mu_{t_{k+1}}, \rho_{t_{k+1}}(\theta)) \tag{5.2}$$

For $\alpha = 0$, Equation 2 corresponds to a model relying purely on linear OT and which does not leverage spatial information, as proposed in the state-of-the-art [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022]. In the following, we refer to this as the linear method. [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022] propose different strategies to make the predictions $\rho_{t_{k+1}}(\theta)$. They vary in terms of teacher-forcing, number of steps between t_k and t_{k+1}, and whether steps are implicit or explicit (see "Discretization" in Methods). The linear method presented here ($\alpha = 0$) incorporates the best-performing choices for our experiments (see "Discretization" in Methods and Supplementary Figure D.1). The linear method thus offers an unbiased way to assess the performance of STORIES compared to the state-of-the-art, as explored in the next section.

STORIES is implemented as an open-source Python package seamlessly integrated into the classical Python single-cell analysis pipeline (https://github.com/cantinilab/stories). Users can thus take advantage of scverse tools like Scanpy, Squidpy, and CellRank for preprocessing and downstream analysis [Wolf et al., 2018, Palla et al., 2021, Weiler et al., 2023]. In addition, STORIES provides a user-friendly visualization of driver genes and enriched transcription factors, thus helping biological interpretability.

In the following sections, we extensively benchmark STORIES against the state-of-the-art [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022] using several large-scale spatiotemporal atlases. To this end, we compare the linear method to our proposed model across different values of the parameter α.

Figure 5.1: Overview of STORIES.



(A) STORIES takes as an input spatial transcriptomics through time; **(B)** STORIES learns the parameters of a neural network representing the differentiation potential of a cell based on its transcriptomic profile. The objective function is based on Fused Gromov-Wasserstein, which leverages both the transcriptomic profile and the spatial coordinates; **(C)** The gradient of the function delivers a velocity that can be used to perform trajectory inference. The potential itself is a natural alternative to pseudotime and allows to study gene trends along differentiation. Finally, STORIES can highlight possible transcription factors regulating differentiation.

5.2.2 STORIES's predictions are more biologically relevant than the state-of-the-art

We assessed the effectiveness of STORIES in predicting cell states over time across three Stereo-seq spatiotemporal atlases: a mouse development atlas [Chen et al., 2022a], a zebrafish development atlas [Liu et al., 2022], and an axolotl brain regeneration atlas [Wei et al., 2022]. Details on data processing are provided in the Methods section.

From each atlas, we created three sets: a training set, an early test set, and a late test set (see Figure 5.2A). The test sets are composed of two time points, and the goal is to use the first time point to predict the second time point's gene expression. The late test set is particularly challenging because its second slice comes from an entirely new time point, which may contain cell states not seen during training. For example, in the zebrafish atlas, fast muscle cells only appear at 24 hours post-fertilization (i.e. hpf), whereas the training set includes slices only up to 18 hpf.

Two scores, defined in the Methods section, have been used for the benchmark. First, the gene expression prediction score evaluates how closely the prediction approaches the real gene expression at the next time point. Second, the spatial coherence score evaluates whether the predictions are

consistent with space. For instance, predicting lung cells from liver cells would yield a bad spatial coherence score.

We compared STORIES to the linear method described in the previous section, which incorporates the best-performing aspects of [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022] for our experiments (see "Discretization" in Methods and Supplementary Figure D.1). Since the linear method is formally equivalent to STORIES with $\alpha = 0$, it offers an unbiased way to assess the usefulness of spatial information compared to the state-of-the-art [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022].

To better evaluate the impact of spatial information, we reported results with a varying weight α, representing the importance of the spatial term compared to the gene expression term (see Figure 5.2B). As expected, smaller values of α perform better in terms of gene expression prediction score, while larger values perform better in terms of spatial coherence score. A good compromise between gene expression and space is obtained for $\alpha = 5 \times 10^{-3}$, which we set as a default value.

For this choice of α, STORIES outperforms the linear method in terms of spatial coherence across all datasets for both early and late test sets (see Figure 5.2B). In addition, STORIES outperforms the linear method in terms of gene expression prediction in the mouse atlas's early test set and the zebrafish atlas's early test set. This demonstrates that STORIES's predictions are biologically relevant on two levels: they are accurate in gene expression and coherent in space.

The biological relevance of STORIES's predictions is confirmed in Figure 5.2C, which compares the training losses of STORIES and the linear method. Since both methods involve matching predictions with a reference population of cells, we compared their matchings for specific cell types (see Methods).

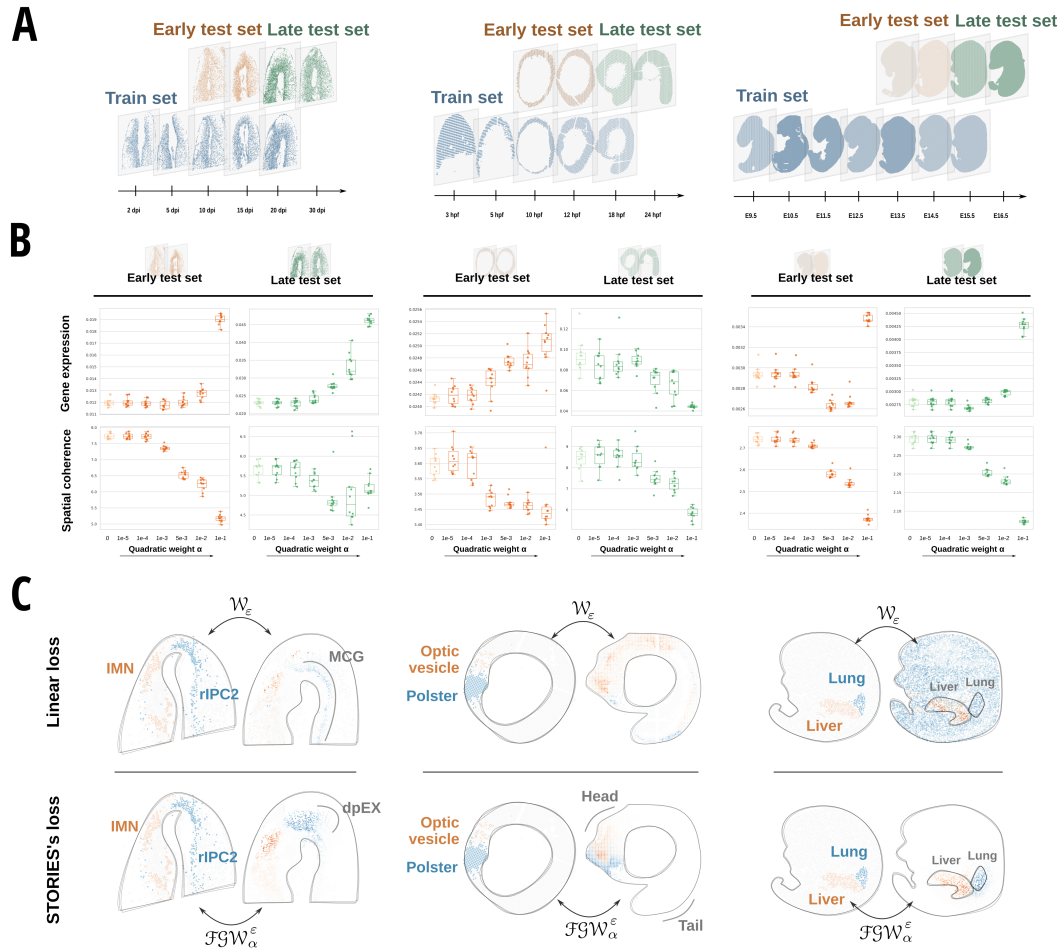
First, in the axolotl atlas, STORIES correctly matches predictions from immature neurons (IMN) with Nptx+ excitatory neurons in the lateral pallium (NptxEX), and predictions from regeneration intermediate progenitor cells (rIPC2) with excitatory neurons in the dorsal pallium (dpEX) [Wei et al., 2022]. The linear method, on the contrary, incorrectly matches rIPC2 predictions with microglial cells (MCG) from a different anatomical region (see Figure 5.2C).

Second, in the zebrafish atlas, STORIES accurately matches predictions from the optic vesicle with cells located around the eye, and predictions from the polster with cells located within the head. In contrast, the linear method incorrectly matches optic vesicle predictions with a broad group of cells across different anatomical regions, and polster predictions with cells from the tail area (see Figure 5.2C).

Third, in the mouse atlas, STORIES correctly matches predictions from liver and lung cells with their respective organs. The linear method, instead, incorrectly matches lung cell predictions with a broad group of cells across organs (see Figure 5.2C), suggesting poor biological coherence of the linear method.

STORIES's superior performance in achieving biologically coherent and accurate gene expression predictions demonstrates the significant benefit of considering spatial information when learning a gradient flow model on spatial transcriptomics data.

Figure 5.2: Benchmark of STORIES on three large datasets.



(A) Visual representation of the three datasets in our benchmark. From left to right, an axolotl brain regeneration dataset, zebrafish development, and mouse development. Slices in each dataset are split into a train set (blue), an early test set (orange), and a late test set (green); **(B)** Gene expression prediction score (top) and spatial coherence score (bottom) in early test sets (orange) and late test sets (green) across the three datasets. Scores are reported for ten seeds, across seven values of quadratic weight parameter α, including the linear method ($\alpha = 0$, light orange/green); **(C)** Visual representation of the optimal transport matching involved in the loss of the linear method (top) and STORIES for $\alpha = 5 \times 10^{-3}$ (bottom), for the three datasets. In each dataset, the left slice displays two cell types and the right slice displays the cells they are matched with at the following time point.

5.2.3 STORIES identifies trajectories and potential drivers of neuron regeneration in axolotls

To further assess the potential of STORIES for trajectory inference in spatial transcriptomics through time, we first focused on axolotl brain regeneration.

We trained STORIES as described in Methods on the subset of cells described in [Wei et al., 2022] as involved in neuron regeneration: Wnt+ and reactive ependymoglial cells (wntEGC and reaEGC), regeneration intermediate progenitor cells (rIPC1 and rIPC2), immature neurons (IMN), Nptx+ lateral pallium excitatory neurons (nptxEX), dorsal pallium excitatory neurons (dpEX), and medial pallium excitatory neurons (mpEX). As shown in Figure 5.3A, STORIES learns an energy landscape consistent with the original publication. Indeed, the potential J_θ assigns a high potential to progenitor states (wntEGC and reaEGC), a medium potential to intermediary states (rIPC1, rIPC2, and IMN), and a low potential to mature states (nptxEX, dpEX, and mpEX).

We computed cell-cell transitions by applying CellRank on the gradient of the trained potential (see Methods), as visualized in Figure 5.3B. These transitions highlight that STORIES not only detect the correct stage of differentiation, but also recovers the three trajectories described in the original publication: wntEGC-mpEX, reaEGC-rIPC2-dpEX, and reaEGC-rIPC1-IMN-nptxEX [Wei et al., 2022]. Importantly, the authors identified these three trajectories by applying Moncole separately on three spatial regions and specifying EGCs as the starting point in each case. In contrast, STORIES achieves the same results without the need to isolate specific spatial regions and to specify the starting point of the trajectory. Indeed, STORIES leverages spatial information to process all regions simultaneously and leverages temporal information to infer progenitor states from the data.

We then narrowed further into the reaEGC-rIPC1-IMN-nptxEX trajectory, which the original publication studies in most detail [Wei et al., 2022]. First, we sought to confirm expected gene trends along this trajectory. The original study [Wei et al., 2022] suggests *Vim*, which encodes a critical cytoskeletal protein, as a marker of reaEGC cells and *Nptx1*, which is involved in synaptic plasticity, as a marker of NptxEX. Accordingly, STORIES recovered a clear decreasing trend for Vim expression along differentiation and a clear increasing trend for *Nptx1* expression (see Figure 5.3C).

Next, we performed unsupervised discovery of gene trends by fitting a spline regression model along the previously mentioned trajectory (see Methods). Figure 5.3D reports the best candidate driver genes across differentiation stages. Interestingly, the early stages of differentiation coincide with high expression of *Hes5*, which is known to maintain stemness in the context of neural differentiation [Ohtsuka et al., 2001], and *Cdc25B*, a cell-cycle regulator key to neuron production [Peco et al., 2012, Bonnet et al., 2018]. Conversely, late stages of differentiation coincide with high expression of the microtubule-associated protein gene *Map1a*, crucial to neural development and regeneration [Nunez and Fischer, 1997], and *L1cam*, shown to promote axonal regeneration [Schäfer and Frotscher, 2012]. STORIES also outputs additional genes that represent possible drivers of neuron regeneration and would require further biological investigation. For instance, STORIES uncovered a trend for late expression of the scarcely studied *Nsg2*, which is thought to be involved in synaptic function and, like *Nptx1*, interacts with AMPA receptors [Chander et al., 2019].

Finally, our analysis revealed possible transcriptional regulators of the differentiation process (see Figure 5.3F) by testing transcription factor (TF) enrichment using the curated literature-based TRRUST [Han et al., 2018] database (see Methods). The most significantly enriched TF, CTNNB1, encodes β-catenin, which has been described as an essential regulator in neuron regeneration in mouse

models [Marchetti, 2018, Duraikannu et al., 2018] and in limb regeneration in axolotl [Kawakami et al., 2006]. Other top TFs include SP1 and MYC, described in the context of neuron regeneration [Miras-Portugal et al., 2016, Zhang et al., 2023, Yang et al., 2022] and computationally retrieved in axolotl limb regeneration [Jhamb et al., 2011]. Additionally, we identify SOX6, MYCN, and REST, which are not widely studied in the context of regeneration but are known regulators in development [Tolchin et al., 2020, Chen and Guan, 2022, Ballas and Mandel, 2005]. Interestingly, a recent study predicted REST as a regulator of neuron regeneration and validated this role in a mouse model [Cheng et al., 2022].

STORIES thus learned a Waddington landscape that captures cell fate trajectories, continuous changes of gene expression, and the regulatory landscape underlying neuron regeneration in axolotls. Our model allowed for the unbiased discovery of potential drivers and mechanisms, possibly relevant for further biological investigations.

5.2.4 STORIES identifies trajectories and potential drivers of gliogenesis in the developing mouse dorsal midbrain

We then sought to highlight STORIES's potential in trajectory inference by studying mouse dorsal midbrain development.

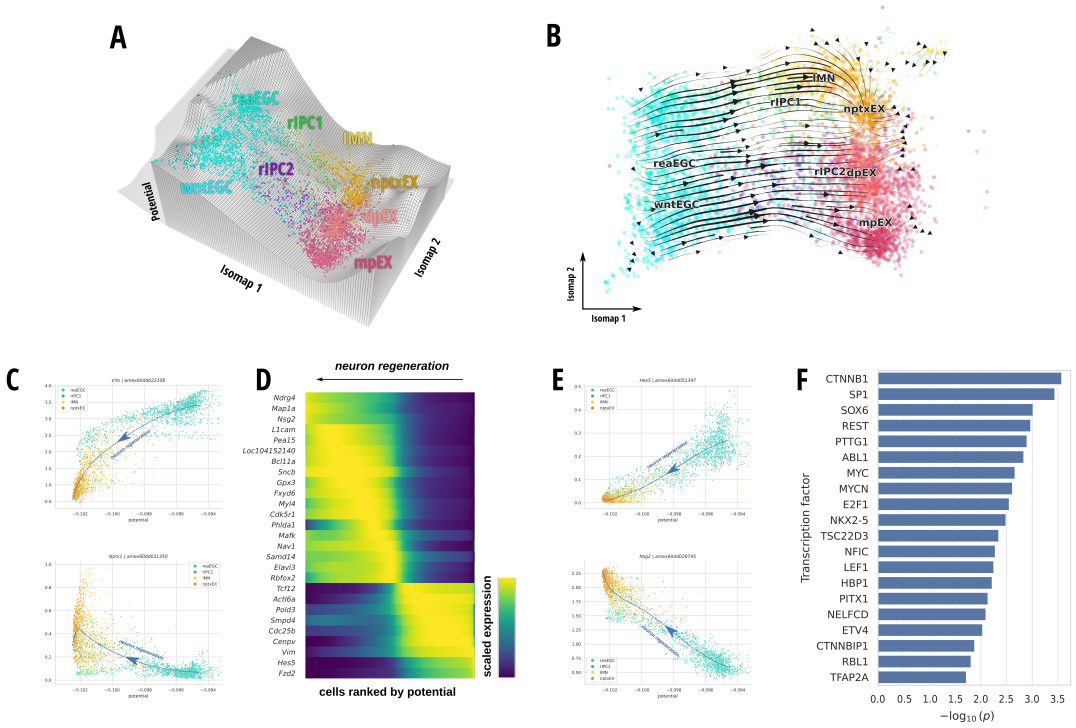
We trained STORIES as described in Methods on the subset of cells described in [Chen et al., 2022a] as exhibiting a branching trajectory: radial glial cells (RGC) differentiating into either neuroblasts (NeuB) or glioblasts (GlioB). As visualized in Figure 5.4A, STORIES learns an energy landscape consistent with the original publication. Indeed, the potential J_θ assigns a high potential to RGC and a low potential to the more differentiated NeuB and GlioB.

We computed cell-cell transitions by applying CellRank on the gradient of the trained potential (see Methods), as visualized in Figure 5.4B. These transitions highlight that STORIES not only detects the correct stage of differentiation but also recovers the expected branching from RGC to glial and neural cell fates [Chen et al., 2022a]. Importantly, the original publication identified this branching using Monocle 3, which required manually setting RGC as the trajectories' starting point. On the contrary, STORIES achieves the same results without manual input by leveraging temporal information to infer the starting point from the data.

Glial cells outnumber neurons in the brain, but their development has been studied less extensively [He and Sun, 2007]. Moreover, understanding gliogenesis is of critical therapeutic importance because of its parallels with glioma, the most common and deadliest form of brain cancer [Laug et al., 2018]. Thus, we focused further on the RGC-GlioB trajectory. We first sought to confirm expected gene trends along this trajectory. The original study identifies *Mki67*, a proliferation marker, as highly expressed in RGC, and *Aldh1l1*, an astrocyte marker, as highly expressed in GlioB [Chen et al., 2022a]. Accordingly, STORIES recovered a decreasing trend for *Mki67* expression along differentiation and an increasing trend for *Aldh1l1* expression (see Figure 5.4C).

Next, we performed unsupervised discovery of gene trends by fitting a spline regression model along the previously mentioned trajectory (see Methods). Figure 5.4D reports the best candidate driver genes across differentiation stages. The early stages of differentiation coincide with a high expression of cell cycle genes *Gmnn*, *Rrm2*, and *Hmgb2* [Tirosh et al., 2016a]. Additionally, we observed a high expression of the alpha-tubulin gene *Tuba1b* in the early stages of differentiation, as previously described in the developing brain [Hausrat et al., 2021]. Conversely, late stages of differentiation

Figure 5.3: Trajectory inference with STORIES in axolotl neuron regeneration.



(A) 3-D representation of the potential landscape learned with STORIES. The x and y axes are Isomap coordinates, and the z-axis is an interpolation of the potential. Colors represent cell types involved in the regeneration process; **(B)** Visual representation of cell-cell transitions computed using CellRank from STORIES's velocity vectors; **(C)** Smoothed gene expression for *Vim* and *Nptx1* along the potential computed by STORIES. The blue line is a spline regression of expression from potential; **(D)** Normalized gene expression regressed using a spline model along the potential computed by STORIES. Genes are ordered by the potential for which they achieve maximum expression; **(E)** Smoothed gene expression for *Hes5* and *Nsg2* along the potential computed by STORIES. The blue line is a spline regression of expression from potential; **(F)** Enrichment scores of Transcription Factors targeting candidate driver genes.

coincide with the high expression of the glutamine synthetase gene *Glul*, a key astrocyte marker [Anlauf and Derouiche, 2013, Jurga et al., 2021]. STORIES also outputs additional genes that represent possible drivers of gliogenesis and would require further biological investigation. For instance, *Glis3* displays an increasing trend along gliogenesis (see Figure 5.4E) but is little studied in this context. However, *Glis3* was recently suggested as a therapeutic target to suppress proliferation in glioma [Liu et al., 2021].

Finally, our analysis revealed candidate transcriptional regulators of the differentiation process (see

Figure 5.4F) by testing transcription factor (TF) enrichment using the curated literature-based TRRUST [Han et al., 2018] database (see Methods). Among the most enriched TFs, SOX4 and NOTCH2 have been studied in gliogenesis [Braccioli et al., 2018, Taylor et al., 2007]. Additionally, STORIES recovers MYC, MYCN, and MAX, which have been studied in the context of glioma [Hirvonen et al., 1994].

This second experiment in the context of gliogenesis confirms STORIES's ability to learn a Waddington landscape that captures cell fate trajectories, continuous gene expression changes, and regulatory processes. In doing so, we highlight known and potential mechanisms that motivate further biological investigations.

5.3 Discussion

Recent technological advances in spatial transcriptomics have enabled the tracking of gene expression at single-cell resolution in the spatial context of the tissue. Large datasets of spatial transcriptomics profiled through time give a unique opportunity to understand dynamic biological processes such as development and disease onset. However, their analysis requires trajectory inference tools tailored to the specific challenges of spatial data.

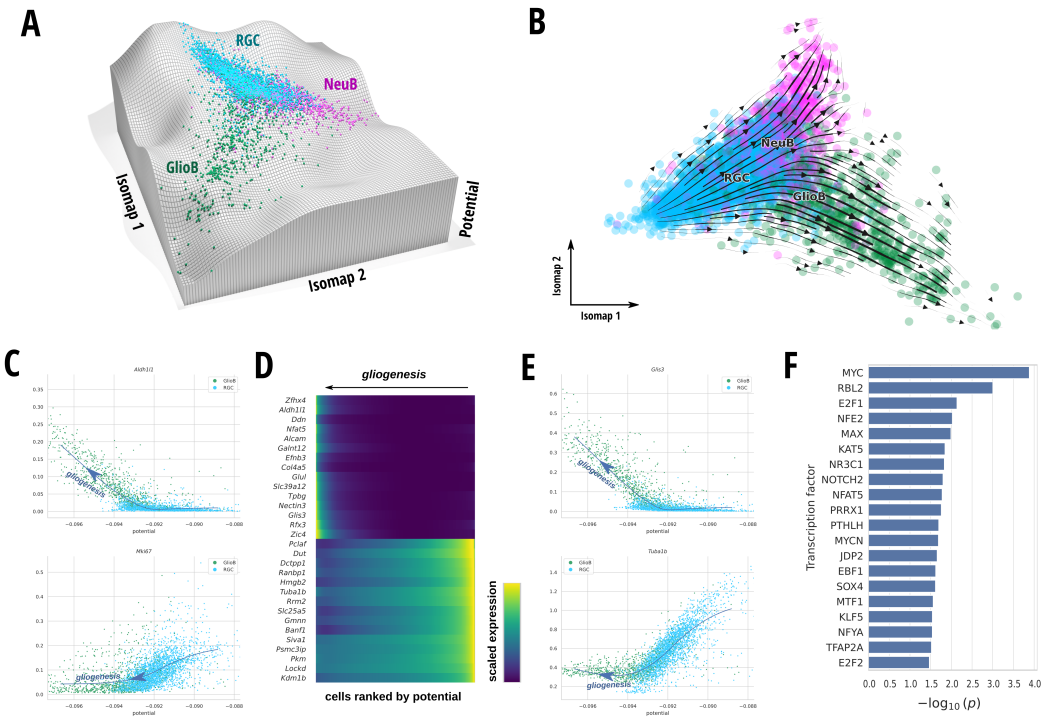
In this article, we proposed SpatioTemporal Omics eneRgIES (STORIES), a novel computational framework for trajectory inference from spatial transcriptomics profiled at several time points. STORIES enables a rich and spatially informed analysis of differentiation trajectories. To evaluate STORIES's performance, we benchmarked it against the state-of-the-art in three large Stereo-seq datasets and highlighted the advantage of considering spatial information in trajectory inference from time-course single-cell data. We then further showcased STORIES's abilities in two concrete settings: axolotl neuron regeneration and mouse gliogenesis.

The most widely used sequencing-based spatial transcriptomics technologies, like 10x Visium, have a spot size larger than the typical cell size [Ståhl et al., 2016]. STORIES seeks to model population dynamics at single-cell resolution and is thus limited to recent technologies like Stereo-seq or Visium HD, for which less data is available. STORIES could leverage deconvolution techniques for low-resolution data [Andersson et al., 2020, Kleshchevnikov et al., 2022] or be adapted for low-throughput imaging-based techniques such as MERFISH [Chen et al., 2015]. In addition, given the fast-paced developments in spatial transcriptomics [Palla et al., 2022], the number of spatiotemporal atlases at single-cell resolution can be expected to increase steadily.

Biological processes such as development, immune response, and the onset of complex disease involve intricate gene regulatory networks (GRNs), cell-cell interactions, and oscillations within a cell state [Weinreb et al., 2018]. Differentiation potentials such as those learned using STORIES arise from simple GRNs and provide an interpretable model of differentiation [Huang et al., 2007]. However, potential energies cannot model complex GRNs, cell-cell communication, or oscillations within a cell state [Weinreb et al., 2018]. Extensions to more complex energy functionals could address these limitations and are an exciting topic for further research. For instance, recent work in gradient flow learning explored interaction energies [Terpin et al., 2024], which could represent cell-cell communication in biology.

Finally, dynamic processes such as development involve coordinated expression changes and tissue reorganization [Zhou et al., 2023]. While STORIES learns from spatial coordinates, it does not predict

Figure 5.4: Trajectory inference with STORIES in mouse gliogenesis.



(A) 3-D representation of the potential landscape learned with STORIES. The x and y axes are Isomap coordinates, and the z-axis is an interpolation of the potential. Colors represent radial glial cells which differentiate into either neuroblasts or glioblasts; **(B)** Visual representation of cell-cell transitions computed using CellRank from STORIES's velocity vectors; **(C)** Smoothed gene expression for *Mki67* and *Aldh1l1* along the potential computed by STORIES. The blue line is a spline regression of expression from potential; **(D)** Normalized gene expression regressed using a spline model along the potential computed by STORIES. Genes are ordered by the potential for which they achieve maximum expression; **(E)** Smoothed gene expression for *Tuba1b* and *Glis3* along the potential computed by STORIES. The blue line is a spline regression of expression from potential; **(F)** Enrichment scores of transcription factors targeting candidate driver genes.

the future positions of cells. Including a spatial component in the energy functional may provide a more comprehensive view of biological processes by predicting cell migration. Relating this to existing models for morphogenesis, such as Alan Turing's reaction-diffusion model [Turing, 1990, Zhu et al., 2010, Economou et al., 2012] is an exciting avenue for further research.

5.4 Methods

5.4.1 Data collection

Zebrafish

For the training and validation sets, we used the five first time points showcased in Figure 2A of [Liu et al., 2022]: 3.3hpf slice 1, 5.25hpf slice 10, 10hpf slice 11, 12hpf slice 8, 18hpf slice 8.

For the test set, we used:

- 10hpf slice 17 and 12hpf slice 5 to evaluate prediction within the time range seen during training. These two slices were studied together in Figure S4A of [Liu et al., 2022].
- 18hpf slice 11 and 24hpf slice 4 to evaluate prediction outside the time range seen during training. These two slices were studied together in Figure S4E of [Liu et al., 2022].

Altogether, this represented 17,920 cells after preprocessing.

Mouse

For the training and validation sets, we used the 7 first time points showcased in Figure 3A of [Chen et al., 2022a]: E9.5 E1S1, E10.5 E1S1, E11.5 E1S1, E12.5 E1S1, E13.5 E1S1, E14.5 E1S1, E15.5 E1S1.

For the test set, we used:

- E13.5 E1S2 and E14.5 E1S2 to evaluate prediction within the time range seen during training. These two slices were studied in Figure S2C of [Chen et al., 2022a].
- E15.5 E1S2 and E16.5 E1S1 to evaluate prediction outside of the time range seen during training. The first slice was studied in Figure S2C and the second in Figure 3A of [Chen et al., 2022a].

Altogether, this represented 794,063 cells after preprocessing.

Dorsal midbrain

We retained one slice per time point. We used the 3 slices showcased in Figure S7 A of [Chen et al., 2022a]: E12.5 (E1S3), E14.5 (E1S3), E16.5 (E1S3). As in the original publication, we subset the analysis to the RGC, NeuB, and GlioB cell types. Altogether, this represented 4,581 cells after preprocessing.

Note that the "Mouse" dataset and the "Dorsal midbrain" dataset originate from the same experiments but have different resolutions (bin 50 vs image-based segmentation), so the same neural network weights cannot be used in both cases.

Axolotl

Benchmark. For the training and validation sets, we used the five first time points showcased in Figure 3B of [Wei et al., 2022]: 2DPI (rep1), 5DPI (rep1), 10DPI (rep1), 15DPI (rep4), 20DPI (rep2). We manually removed spatial outliers in 10DPI (rep1), 15DPI (rep4), and 20DPI (rep2). For the test set, we used:

- 10DPI (rep2) and 15DPI (rep1) to evaluate prediction within the time range seen during training. These two slices were showcased in Figure S9 of [Wei et al., 2022].

- 20DPI (rep3) and 30DPI (rep2) to evaluate prediction outside of the time range seen during training. These two slices were showcased in Figure S9 of [Wei et al., 2022].

As in the original publication, we restricted the analysis to the dorsal part of the injured hemisphere. Altogether, this represented 22,083 cells.

In-depth analysis. For the analysis in Section 5.2.3, we used slices 2DPI (rep1), 5DPI (rep1), 10DPI (rep1), 15DPI (rep4), 20DPI (rep2), and 30DPI (rep2). As before, we manually removed spatial outliers and restricted the analysis to the dorsal part of the injured hemisphere. As in the original publication, we subset the data to the following cell types: nptxEX, reaEGC, wntEGC, dpEX, mpEX, IMN, rIPC1, rIPC2. This represented 5,904 cells.

5.4.2 Preprocessing

For all datasets, we performed the following preprocessing steps.

Cell and gene quality control. Using Scanpy's sc.pp.filter_cells, we removed cells with less than 200 expressed genes. Then, we removed the top 0.1% of cells with the most expressed genes. Finally, we removed genes expressed in less than 3 cells using Scanpy's sc.pp.filter_genes.

Normalization and highly variable gene selection. We applied Scanpy's "Pearson residuals normalization" and selected 10,000 highly variable genes. Scanpy computed highly variable genes for each batch and merged them to avoid selecting batch-specific genes.

Dimensionality reduction. Using Scanpy's sc.tl.pca, we applied Principal Component Analysis (PCA) to reduce the data to 50 dimensions. Section 5.2.3 and Section 5.2.4 used a subset of relevant cell types. This was done after PCA but before batch correction.

Batch correction. We applied Harmony [Korsunsky et al., 2019] on the PCA components to correct the batch effects, using Scanpy's sc.external.pp.harmony_integrate, a wrapper around `harmonypy`.

Visualization. Using Scanpy's sc.tl.umap, we applied UMAP to project the batch-corrected data into two dimensions. In Section 5.2.3 and Section 5.2.4 we applied Isomap instead of UMAP. Indeed, we found that visually, Isomap respected cell type transitions better than UMAP. We used scikit-learn's sklearn.manifold.Isomap.

5.4.3 Wasserstein gradient flow learning with a quadratic objective

The paragraphs below recall definitions presented in the introduction of this thesis (see Section 1.3.1) in the context of population dynamics, namely Wasserstein gradient flows (Equation 1.16), Sinkhorn divergences (Equation 1.9), and Fused Gromov-Wasserstein (Equation 1.14).

Notations

Let us consider a time point $t \in \mathbb{R}$ and $\mu_t \in \mathcal{P}(\mathbb{R}^d)$ a discrete distribution of n cells. We denote $\mu_t = \sum_{i=1}^n a_i \delta_{\mathbf{x}_i}$ where the vector $\mathbf{x}_i \in \mathbb{R}^d$ represents the gene expression of the i-th cell, and the weights $a_i \in \mathbb{R}_+$ are such that $\sum a_i = 1$. In the following, for a function $f : \mathbb{R}^d \to \mathbb{R}^d$, we define the pushforward measure as $f_{\#}\mu_t = \sum_i a_i \delta_{f(\mathbf{x}_i)}$.

Wasserstein gradient flow

Similarly to [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022], we model the evolution of μ_t as a Wasserstein gradient flow for a potential energy. For one single cell $\mathbf{x}$, the Euclidean gradient flow is a $\mathbf{x}_t$ that verifies $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\nabla J(\mathbf{x})$, the continuous counterpart of gradient descent. Wasserstein gradient flows extend this to the space of measures. We say that μ_t is a Wasserstein gradient flow for the energy $\mathcal{F} : \mu \mapsto \sum_{i,j} J(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \in \mathbb{R}$ if its density verifies the continuity equation

$$\frac{\partial \mu(\mathbf{x}, t)}{\partial t} = \operatorname{div}\big(\mu(\mathbf{x}, t)\nabla J(\mathbf{x})\big). \tag{5.3}$$

See [Ambrosio et al., 2005, Santambrogio, 2017] for a rigorous introduction to Wasserstein gradient flows. As in [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022], we do not know the potential $J : \mathbb{R}^d \to \mathbb{R}$ *a priori* and aim instead to learn a neural network J_θ from snapshots $\mu_{t_k} \in \mathcal{P}(\mathbb{R}^d)$ for $t_1 < \cdots < t_k < \cdots < t_K$. See "neural network architecture" for details about J_θ.

Discretization

Our approach boils down to learning J_θ such that for given parameters θ and an initial population $\rho_{t_1} = \mu_{t_1}$, the predicted populations ρ_{t_k} are close to the observed snapshots μ_{t_k}. To make these predictions, existing potential-based methods [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022] differ in three main aspects: number of steps, teacher-forcing, and discretization scheme. As detailed below, we select the best-performing choices for these three aspects, as measured by validation loss in the Zebrafish atlas. The linear method ($\alpha = 0$) thus differs from existing works by combining their best-performing aspects.

Number of steps. [Hashimoto et al., 2016, Yeo et al., 2021] make intermediary predictions between ρ_{t_k} and $\rho_{t_{k+1}}$, i.e. $\tau \ll (t_{k+1} - t_k)$. [Bunne et al., 2022] perform a single step instead, i.e. $\tau = (t_{k+1} - t_k)$. In our experiments, multiple steps did not improve results (see Supplementary Figure D.1A) so we chose the computationally less expensive single-step method.

Teacher-forcing. [Hashimoto et al., 2016, Yeo et al., 2021] predict $\rho_{t_{k+1}}$ from ρ_{t_k}. [Bunne et al., 2022] introduce teacher-forcing, i.e. predicting $\rho_{t_{k+1}}$ from μ_{t_k}. In our experiments, teacher-forcing improved results (see Supplementary Figure D.1B), so we used it throughout this work.

Discretization scheme. To predict $\rho_{t+\tau}$ from an earlier population ρ_t, we used the forward Euler discretization scheme $\rho_{t+\tau} = (\mathrm{Id} - \tau\nabla J_\theta)_{\#}\rho_t$ as in [Hashimoto et al., 2016, Yeo et al., 2021]. In our discrete setting, this corresponds to $\mathbf{x}_{t+\tau} = \mathbf{x}_t - \tau\nabla J_\theta(\mathbf{x}_t)$ for each cell $\mathbf{x}$. [Bunne et al., 2022] propose using a backward Euler scheme to improve stability for large τ. However we did not find this to improve results in our experiments (see Supplementary Figure D.1C), so we used the computationally less expensive forward method.

Pairwise information

Our setting differs from [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022] in that for each cell $\mathbf{x} \in \mathbb{R}^d$ we have access to spatial coordinates $\mathbf{r} \in \mathbb{R}^2$. Spatial coordinates are defined up to an isometry. For instance, a slice of an embryo may be rotated without changing the problem. Consequently, one cannot simply concatenate $\mathbf{x}$ and $\mathbf{r}$ to leverage the spatial coordinates. The next paragraph details how we used the coordinates $\mathbf{r}$ to inform the problem, while still defining $J_\theta : \mathbf{x} \in \mathbb{R}^d \mapsto J_\theta(\mathbf{x}) \in \mathbb{R}$ only on gene expression.

Learning the potential

For each $k \in [1, \dots, K]$, we compared the prediction ρ_{t_k} to the reference snapshot μ_{t_k}.

Linear model. Let us consider two discrete probability distributions $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ and $\rho = \sum_j b_j \delta_{\mathbf{y}_j} \in \mathcal{P}(\mathbb{R}^d)$. To compare μ and ρ, [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022] use the Sinkhorn divergence [Genevay et al., 2018], defined as

$$\overline{\mathcal{W}}_\varepsilon(\mu, \rho) = \mathcal{W}_\varepsilon(\mu, \rho) - \tfrac{1}{2}\big(\mathcal{W}_\varepsilon(\mu, \mu) + \mathcal{W}_\varepsilon(\rho, \rho)\big), \tag{5.4}$$

where $\mathcal{W}_\varepsilon$ is the entropy-regularized Optimal Transport (entropy-regularized OT) [Cuturi, 2013], defined as

$$\mathcal{W}_\varepsilon(\mu, \rho) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j} \|\mathbf{x}_i - \mathbf{y}_j\|_2^2 \, \mathbf{P}_{i,j} - \varepsilon \mathrm{E}(\mathbf{P}). \tag{5.5}$$

Here, $\mathcal{U}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \{\mathbf{P} \in \mathbb{R}_+^{n \times m} \text{ such that } \sum_j P_{i,j} = a_i \text{ and } \sum_i P_{i,j} = b_j\}$ and $\mathrm{E}(\mathbf{P}) = \sum_{i,j} P_{i,j} (\log P_{i,j} - 1)$ is the Shannon entropy. $\overline{\mathcal{W}}_\varepsilon$ is a debiased version of $\mathcal{W}_\varepsilon$, such that $\forall \rho, \overline{\mathcal{W}}_\varepsilon(\rho, \rho) = 0$ [Genevay et al., 2018].

Quadratic model. The Sinkhorn divergence between ρ and μ only compares distributions of gene expression. Instead, we propose a debiased Fused Gromov-Wasserstein (debiased FGW) loss to enforce the spatial coherence of the predictions.

Let us consider $\mu = \sum_i a_i \delta_{(\mathbf{x}_i, \mathbf{r}_i)}$ and $\rho = \sum_j b_j \delta_{(\mathbf{y}_j, \mathbf{s}_j)} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^2)$. Gromov-Wasserstein (GW) is a quadratic extension of OT well suited to compare measures defined up to an isometry [Mémoli, 2011]. [Bunne et al., 2019] used a debiased version of GW to learn a Generative Adversarial Network (GAN). Fused Gromov-Wassertein (FGW) [Vayer et al., 2018a] combines a linear and a quadratic OT term. In our setting, it is natural to use the linear term for gene expression and the quadratic term for spatial coordinates:

$$\mathcal{FGW}_\alpha^\varepsilon(\mu, \rho) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} (1 - \alpha)\mathrm{L}(\mathbf{P}) + \alpha \mathrm{Q}(\mathbf{P}) - \varepsilon \mathrm{E}(\mathbf{P}), \tag{5.6}$$

$$\text{where } \mathrm{L}(\mathbf{P}) = \sum_{i,j} \|\mathbf{x}_i - \mathbf{y}_j\|_2^2 \, \mathbf{P}_{i,j} \tag{5.7}$$

$$\mathrm{Q}(\mathbf{P}) = \sum_{i,j,i',j'} \left| \|\mathbf{r}_i - \mathbf{r}_{i'}\|_2^2 - \|\mathbf{s}_j - \mathbf{s}_{j'}\|_2^2 \right|^2 \mathbf{P}_{i,j} \mathbf{P}_{i',j'}. \tag{5.8}$$

Analogously to [Bunne et al., 2019], we introduce a debiased FGW to ensure the loss vanishes for an exact match:

$$\overline{\mathcal{FGW}_\alpha^\varepsilon}(\mu, \rho) = \mathcal{FGW}_\alpha^\varepsilon(\mu, \rho) - \tfrac{1}{2}\big(\mathcal{FGW}_\alpha^\varepsilon(\mu, \mu) + \mathcal{FGW}_\alpha^\varepsilon(\rho, \rho)\big) \tag{5.9}$$

A weight $\alpha = 1$ corresponds to the debiased entropy-regularized GW introduced in [Peyré et al., 2016]. As $\alpha = 0$, we recover the Sinkhorn divergence $\overline{\mathcal{W}_\varepsilon}$.

Final loss function. For $\alpha \in [0, 1[$, the full objective reads

$$\mathcal{L}(\theta) = \sum_{k=2}^{K} (t_k - t_{k-1}) \overline{\mathcal{FGW}_\alpha^\varepsilon}\big(\mu_{t_k}, \rho_{t_k}\big) \tag{5.10}$$

where $\rho_{t_k} = (f_\theta)_{\#} \mu_{t_{k-1}}$ with $f_\theta : (\mathbf{x}, \mathbf{r}) \mapsto \big(\mathbf{x} - \tau \nabla J_\theta(\mathbf{x}), \mathbf{r}\big)$. In other words, gene expression is predicted with teacher-forcing and a single forward step (see Section 5.4.3), and the spatial coordinates remain unchanged. The loss in Equation 5.10 is optimized using minibatches, as in [Yeo et al.,

2021, Bunne et al., 2022]. A theoretical study of minibatch Optimal Transport for debiased linear and quadratic OT is provided by [Fatras et al., 2021].

Choice of the quadratic weight

To investigate the effect of the relative weight of the linear term in FGW, we reported results for $\alpha \in \{1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}\}$. Choosing $\alpha = 1$ would not allow learning the potential J_θ because it would ignore gene expression. As expected, with a low weight ($\alpha = 10^{-5}$), STORIES behaves as the linear method (see Section 5.2.2).

To compare our approach with the linear model proposed in [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022], we also trained the model with a Sinkhorn divergence, i.e. $\alpha = 0$.

The benchmark in Section 5.2.2 suggests a default value of $\alpha = 5 \times 10^{-3}$, which can be adjusted by the user depending on the dataset. In Section 5.2.3 and Section 5.2.4 we set a value of $\alpha = 10^{-3}$.

5.4.4 Computational Optimal Transport

OTT solvers

We use the OTT package [Cuturi et al., 2022] to solve OT problems in a fast, GPU-enabled, and differentiable manner. In particular, we rely on the Sinkhorn and GromovWasserstein solvers. We set the entropic regularization $\varepsilon = 0.01$.

Linear term: gene expression

The linear OT terms are defined on gene expression space, which we defined as the d first components of the Harmony-aligned Principal Component Analysis. We chose $d = 20$ components as in [Bunne et al., 2022]. Before training the neural network, we normalized the points as $\mathbf{x}_i \leftarrow \frac{\mathbf{x}_i}{\max_j \|\mathbf{x}_j\|_\infty}$ to make the linear and quadratic terms comparable.

Quadratic term: Spatial coordinates

Before training the neural network and for each slice separately, we centered the spatial coordinates and scaled them to unit variance to make linear and quadratic terms comparable.

5.4.5 Neural network architecture

We implemented a Multi-Layer Perceptron (MLP) with two hidden layers of dimension 128 and GeLU activations using Flax [Heek et al., 2023]. This is a similar architecture to [Hashimoto et al., 2016, Yeo et al., 2021, Bunne et al., 2022]. The linear output layer has no bias since it would not influence the values of ∇J_θ. Likewise, a soft activation like GeLU is preferable to the classical ReLU because we manipulate ∇J_θ. Indeed, the derivative of ReLU is simply a unit step function, which is discontinuous and not very expressive.

5.4.6 Neural network training

Data loading

For each time point described in "Data Collection", 75% of the cells were used as training samples and 25% as validation. At each training or validation iteration, a batch containing 1,000 cells per time point was sampled uniformly without replacement. In development, the early time points contain fewer cells than the later time points. If less than 1,000 cells were available for a time point, we used all available cells. To reflect the train/validation split, one in four iterations performs a validation step.

Optimizer

We used Optax's [DeepMind et al., 2020] implementation of the AdamW [Loshchilov and Hutter, 2017] optimizer, with parameters `b1=0.9`, `b2=0.999`, `eps=1e-8`, and `weight_decay=1e-4`. We set the learning rate using Optax's cosine scheduler, with an initial value of `1e-2` and `10,000` decay steps. To ensure convergence, when performing 10 steps, we set the learning rate to `1e-3`. Similarly, when performing an implicit step, we set the learning rate to `1e-4`.

Early stopping

We set the maximum number of iterations to 15,000 but stopped the training when the validation loss had not improved in 150 iterations. We kept the weights associated with the lowest validation loss.

Seeds

We ran every experiment with 10 random seeds: 17158, 20181, 12409, 5360, 21712, 21781, 24802, 13630, 9668, and 651. The random seed reproducibly determines the train/validation split and weight initialization. For the plots in Figure 5.2C, and the analysis in Section 5.2.3 and Section 5.2.4, the experiments correspond to the randomly chosen seed 20181.

5.4.7 Computional growth rate

Weight of marginals

When comparing a prediction ρ_{t_k} to the reference snapshot μ_{t_k}, [Yeo et al., 2021] proposes setting the weights b_j of the prediction $\rho_{t_k} = \sum_j b_j \delta_{\mathbf{y}_j}$ proportionally to a computationally derived growth rate. The motivation is that cells with a larger growth rate should be matched to more descendants. This idea was introduced by Schiebinger [Schiebinger et al., 2019] and recently reimplemented in MOSCOT [Klein et al., 2023]. In the next paragraphs, we follow MOSCOT's implementation.

Proliferation and apoptosis

Computing the growth rate relies on cell-wise proliferation and apoptosis scores. We used Scanpy's sc.tl.score_genes with lists of genes collected from the literature. The gene lists are described in more detail at the end of the section. We calculated gene scores on raw counts, after quality filtering of cells and genes.

Calculating the growth rate

For a given cell $\mathbf{x}_i$, let us call $\text{prol}_i \in \mathbb{R}$ the proliferation score and $\text{apo}_i \in \mathbb{R}$ the apoptosis score. We then define the birth rate $\beta_i \in \mathbb{R}$ and the death rate $\gamma_i \in \mathbb{R}$ as

$$\beta_i = \beta_{\text{min}} + \frac{\beta_{\text{max}}}{1 + \exp\left(-4.0 \times \frac{\text{prol}_i - \beta_{\text{center}}}{\beta_{\text{width}}}\right)}, \quad \gamma_i = \gamma_{\text{min}} + \frac{\gamma_{\text{max}}}{1 + \exp\left(-4.0 \times \frac{\text{apo}_i - \gamma_{\text{center}}}{\gamma_{\text{width}}}\right)} \tag{5.11}$$

We used MOSCOT's default parameters $\beta_{\text{min}} = \gamma_{\text{min}} = 0.3$, $\beta_{\text{max}} = \gamma_{\text{max}} = 1.7$, $\beta_{\text{center}} = 0.25$, $\beta_{\text{width}} = 0.5$, $\gamma_{\text{center}} = 0.1$, $\gamma_{\text{width}} = 0.2$. Finally, we defined the cell's growth rate as

$$g_i = \exp\big(\Delta t \times (\beta_i - \gamma_i)\big) \tag{5.12}$$

where Δt is the time difference between populations. We obtained a_i by normalizing the growth rate

$$a_i = \frac{g_i}{\sum_j g_j}, \quad \text{i.e.} \quad \mathbf{a} = \text{softmax}\big(\Delta t \times (\beta - \gamma)\big). \tag{5.13}$$

In this equation, Δt plays the role of the softmax's inverse temperature. The histogram $\mathbf{a}$ will thus be sharper for large values of Δt. Most slices in our experiments were evenly sampled, so we set a fixed $\Delta t = 1$, which yielded sharp enough weight differences between cell types.

In the benchmark

[Schiebinger et al., 2019, Yeo et al., 2021, Klein et al., 2023] tested gene sets to compute the growth rate in the case of mice and humans, but not of zebrafish and axolotl. We thus kept uniform marginals for the benchmark in Section 5.2.2. In Section 5.2.3 and Section 5.2.4, we analyzed a restricted number of cell types, where we could validate the biological coherence of the computed growth rate and of the learned potential (see details in paragraphs below).

Dorsal midbrain

For Section 5.2.3, we used murine proliferation and apoptosis gene sets from MOSCOT. Proliferation genes come from https://doi.org/10.1038/nature20123 and apoptosis genes from https://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_P53_PATHWAY. See Appendix D.1 and Appendix D.2 for the gene names.

Axolotl neuron regeneration

For Section 5.2.4, we used an NSC axolotl gene set described in the original publication https://www.science.org/doi/10.1126/science.abp9444 to represent proliferation and a human apoptosis gene set from MOSCOT, originally from gsea-msigdb's HALLMARK_APOPTOSIS. See Appendix D.3 and Appendix D.4 for the gene names.

5.4.8 Evaluation

We solve the FGW problem in Equation 5.6 with $\alpha = 10^{-3}$ and $\varepsilon = 10^{-3}$ and report the terms $\text{L}(\mathbf{P})$ and $\text{Q}(\mathbf{P})$ separately. Since $\text{L}(\mathbf{P})$ quantifies the error in terms of gene expression, we call this quantity the **gene expression prediction score**. Similarly, since $\text{Q}(\mathbf{P})$ quantifies the error in terms of spatial coordinates, we call this quantity the **spatial coherence score**.

Figure 5.3C compares the transport plans involved in the linear method and STORIES with $\alpha = 5 \times 10^{-3}$. Formally, let us consider the indicator vector $\mathbf{a} \in \mathbb{R}^n$ where $a_i = 1$ if the i-th cell corresponds to a given cell type, and $a_i = 0$ otherwise. The transport plan $\mathbf{P}$ between the prediction ρ_t and the ground truth μ_t is applied to the indicator, yielding a vector $\mathbf{b} = \mathbf{P}^\top \mathbf{a}$ representing the mass transported from $\mathbf{a}$ towards each cell in the second time point.

5.4.9 Trajectory inference

Gene imputation

In our analysis, the gene expression trends (Figure 5.3 and 15 per group in Figure 5.4) would be negatively affected by the sparsity of gene expression. Cellrank [Weiler et al., 2023] demonstrated good performances in identifying gene expression trends with MAGIC [Van Dijk et al., 2018]. We thus applied MAGIC gene imputation after all other preprocessing steps. We computed the exponentiated Markov transition matrix on the Harmony-aligned PCA space instead of the original PCA. We did not use the imputed signal for tasks other than gene expression trends.

Potential visualization

The neural potential J_θ is a functional defined on the 20-dimensional space of Harmony-aligned principal components. To visualize the potential as a Waddington-like landscape defined on 2 dimensions, we proceed similarly to [Qin et al., 2023].

- First, we compute the potential $J_\theta(\mathbf{x})$ associated with each cell $\mathbf{x}$.
- Then, we use Scipy's [Virtanen et al., 2020] RBF interpolation and the 2-dimensional Isomap coordinates of the cells to define a potential on a 2-dimensional grid.
- The cells are projected on the surface using the interpolator.
- Finally, the maximum value is thresholded.

We rendered the resulting surface and point cloud using Blender's Python API.

Cell-cell transition matrix

We used CellRank's VelocityKernel with a velocity $v(\mathbf{x}) = -\nabla J_\theta(\mathbf{x})$ for a trained J_θ. Based on this kernel, we computed a cell-cell transition matrix, and the trajectory plots Figure 5.3B, Figure 5.4B.

Gene expression trends

We fit the MAGIC-imputed gene expression as a function of the learned potential, using Scipy's [Virtanen et al., 2020] Spline Regression. To plot the gene expression cascades in Figure 5.3D, Figure 5.4D, we order genes by the value of potential for which the maximum regressed expression value is achieved. Genes are then split into equally sized groups illustrating diffrent stages of differentiation (10 groups for Figure 5.3D and 2 groups for Figure 5.4D). Finally, regressed values for the genes with the best regression scores in each group are displayed (3 per group in Figure 5.3D and 15 per group in Figure 5.4D).

Transcription factor enrichment

We perform TF-target enrichment based on the TRRUST dataset, which contains "Activation", "Repression", and "Unknown" links based on a curated literature search [Han et al., 2018]. For each TF in the database, we perform a Wilcoxon rank-sum test comparing the list of regression scores of its target genes, and the list of regression scores of the other genes. Figure 5.3B and Figure 5.4B display the TFs ranked by p-value.

Chapter 6

Discussion

6.1 Summary of contributions

Paired single-cell multimodal sequencing technologies (paired single-cell multiomics and spatially resolved single-cell transcriptomics) have transformed the study of complex biological systems by enabling the unbiased discovery of cellular heterogeneity and dynamic processes.

The novel approaches for paired single-cell multimodal integration described in this thesis fall into two conceptual frameworks based on Optimal Transport.

- First, **joint dimensionality reduction of paired single-cell multi-omics data.** Chapter 2, Chapter 3, and Chapter 4 frame cells as discrete probability distributions over the space of molecular features. Based on this representation, comparing cells using Optimal Transport allows us to capture the complex interplay between features in a ground cost $\mathbf{C}$. Chapter 2 and Chapter 4 use a data-driven choice of $\mathbf{C}$, while Chapter 3 offers a new way to jointly learn a cost between features, and a distance between cells. Chapter 2 and Chapter 3 consider unimodal clustering tasks, and Chapter 4 extends this framework to the case of joint dimensionality reduction for single-cell multi-omics. Biological insights include the in-depth characterization of several precise CD8+ T cell clusters in a dataset with three omics.

- Second, **trajectory inference in spatially resolved single-cell transcriptomics data.** Unlike the previous chapters, Chapter 5 considers distributions of cells defined by their gene expression and spatial coordinates. The proposed method frames cellular differentiation as a Wasserstein gradient flow. Unlike existing approaches, Chapter 5 uses quadratic Optimal Transport to learn the associated energy functional. It thus enables learning from spatial coordinates, which are defined up to an isometry and cannot be compared directly between slices. This method delivers a highly interpretable differentiation potential, which can (i) order cells along a differentiation axis and (ii) predict future transcriptomic states of cells. An in-depth analysis of axolotl neuron regeneration and mouse gliogenesis reveals gene trends and possible transcriptional regulators associated with specific differentiation trajectories.

Although the versatility of OT allows the use of the same formalism in these two directions, they represent two drastically different biological settings. In Chapter 2, Chapter 3, and Chapter 4, the distribution $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ represents a cell, and OT transports mass between genes. In Chapter 5, the distribution $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ represents a population of cells, and OT transports mass between cells.

Open-source and easy-to-use Python packages encourage the adoption of the proposed contributions and the reproducibility of the experiments outlined in this thesis.

6.2 Limitations and perspectives

New frontiers for paired single-cell multimodal data include the emergence of novel techniques like spatial multi-omics [Baysoy et al., 2023, Vandereyken et al., 2023] and the generalization of large-scale atlases [Regev et al., 2017]. Taking full advantage of increasingly complex and large-scale paired multimodal single-cell omics data motivates further work to build upon the OT-based paired multimodal integration methods presented in this thesis. The sections below point out some limitations of the proposed methods and outline exciting opportunities for future work.

Atlas-scale computation of OT distances for single-cell data

Technological advances like combinatorial indexing drive the exponential growth of the number of cells sequenced in single-cell experiments [Svensson et al., 2018]. In addition, coordinated efforts across institutes through ambitious programs like the Human Cell Atlas are delivering large multi-donor datasets across tissues [Regev et al., 2017]. The increasing availability of such large-scale paired multimodal atlases is shifting methodological development towards models that learn from vast volumes of single-cell data [Theodoris et al., 2023, Cui et al., 2024].

The work presented in this thesis represents novel applications of Optimal Transport in the context of paired single-cell multimodal integration. However, OT is inherently computationally expensive to compute (see Section 1.3.2). Below, I outline the principal computational bottlenecks in this thesis and point out exciting opportunities in methodological development to take full advantage of the growing quantity of available data.

Chapter 2 and Chapter 3 involve computing large pairwise OT distance matrices on datasets of thousands of cells. However, scaling this approach to single-cell atlases requires methodological developments to mitigate the computational cost. Two recent works have proposed methods to accelerate experiments from Chapter 2 and Chapter 3 by approximating OT distances. First, [Tong et al., 2022] proposed approximating OT distances with L1 distances on multiscale embeddings and, repeating an experiment from Chapter 2, computed pairwise OT distances between cells in scRNA-seq data with a significant speedup. Second, [Düsterwald and Yamada, 2024] adapted the Wasserstein Singular Vectors proposed in Chapter 3 to a tree-based approximation of OT.

[Courty et al., 2017] and [Haviv et al., 2024] proposed neural network-based approaches to embed point clouds in a low-dimensional space such that Euclidean distances in that space approximate p-Wasserstein distances in the original space. On top of speeding up the computation of pairwise distances, Euclidean distances are compatible with fast approximate nearest-neighbor search algorithms [Dong et al., 2011]. Tools like Scanpy rely on approximate neighbor search for the rapid computation of kNN graphs for downstream analysis of single-cell data [Wolf et al., 2018].

The joint dimensionality reduction approach proposed in Chapter 4 only requires the computation of $\mathcal{O}(n)$ cell-cell distances at each optimization step. However, in cases where the number of cells is too large, one could extend the proposed method to an online strategy for nonnegative matrix factorization [Mairal et al., 2009]. If, instead, the number of features becomes a bottleneck due to

the memory footprint of the cost matrix $\mathbf{C}$, a straightforward extension of the proposed method is to replace $\mathbf{C}$ with a precomputed low-rank factorization.

In Chapter 5, handling large datasets is achieved through mini-batching the Fused Gromov-Wasserstein distance, an approach studied theoretically in [Fatras et al., 2021]. However, some biological settings may necessitate larger batch sizes to accurately represent cellular heterogeneity at each iteration. In this case, low-rank OT [Scetbon and Cuturi, 2022] offers a fast and less memory-intensive approximation of OT. [Klein et al., 2023] has demonstrated the efficacity of low-rank Fused Gromov-Wasserstein for pairwise mapping of whole embryos in spatial transcriptomics.

Opportunities for methodological improvements to gain further biological insights

As single-cell multimodal techniques mature, they enable the identification of rare cell types and facilitate the discovery of complex regulatory mechanisms. For instance, a recently published atlas of the human tonsil measured spatial transcriptomics, 10X Multiome, CITE-seq, and identified 121 cell types, along with differentiation trajectories and the associated molecular mechanisms [Massoni-Badosa et al., 2024]. A key challenge for paired single-cell multimodal integration methods is the ability to model these intricate relationships effectively [Argelaguet et al., 2021, Adossa et al., 2021]. Below, I highlight some opportunities for methodological improvement to the methods proposed in this thesis, with the aim of deriving further biological insights from increasingly complex data.

First, all contributions across this thesis consider discrete probability measures, meaning they have unit mass. In Chapter 2, Chapter 3, Chapter 4, a cell is represented as a histogram on molecular features. In Chapter 5, the weights in a population of cells always have unit sum. This normalization constraint may give disproportionate importance to outliers in the first case and does not account for population growth in the second case. A natural solution to this limitation is to extend the various contributions in this thesis to unbalanced OT, which allows the manipulation of measures of arbitrary mass [Séjourné et al., 2023].

Next, the joint matrix factorization model proposed in Chapter 4 is a linear model, which benefits from solid interpretability but may lack the expressivity required to catch subtle nonlinear relationships between features. To obtain further biological insights, possible future directions for Wasserstein-based dimensionality reduction include nonlinear dictionary learning [Schmitz et al., 2018] or neural networks.

Finally, the potential-based trajectory inference method proposed in Chapter 5 provides an interpretable model of cellular differentiation but assumes cellular trajectories to follow a gradient field. As explained in [Weinreb et al., 2018], such trajectories cannot model fluctuations within a cell state and can only arise from symmetric gene regulatory networks (GRNs), which capture few realistic biological processes. Formally, [Weinreb et al., 2018] frames the elements of the Hessian matrix of $\delta\mathcal{F}$, the energy's first variation, as the weights of a GRN. A promising way to model dynamic processes arising from more realistic GRNs is thus to consider more general classes of functionals $\mathcal{F}$. Generalizing the approach proposed in Chapter 5 to more complex energy functionals could enable to account for cell-cell interactions, spatial information, and conditions. Recently, [Terpin et al., 2024] explored learning an interaction term and a diffusion term along with the potential function.

Unifying ground metric learning, dimensionality reduction, and trajectory inference

In the context of trajectory inference from single-cell data profiled at several time points, emerging technologies have the potential to offer a more comprehensive view of the cell and its evolution across time. For instance, recent datasets have made paired single-cell multiomics data profiled at several time points available [Burkhardt et al., 2022]. In addition, spatial multi-omics technologies are being developed and will allow to characterize cellular differentiation at unprecedented levels. Finally, recent advances like lineage tracing [McKenna et al., 2016], time-resolved transcriptomics [Erhard et al., 2022], and Live-seq [Chen et al., 2022b] enable the partial tracking of the evolution of cells across time. As technological developments offer an increasingly comprehensive view of the cell, methodological developments should take advantage of all aspects of cellular profiling in a unified framework. Below, I draw links between this thesis's contributions and highlight the potential for a unified OT-based framework to model cellular identity across time.

Chapter 5, as all OT-based methods for trajectory inference, uses a Euclidean ground cost between cells to model population dynamics. A natural link with Chapter 2 would be to consider OT distances between cells instead. Computing OT distances using an OT ground cost is sometimes called hierarchical OT and has found applications in the context of computer vision and natural language processing [Dukler et al., 2019, Yurochkin et al., 2019]. However, these applications take place in a static setting, and extending this framework to model population dynamics motivates the development of a theory of hierarchical Wasserstein gradient flows. In practice, scaling a dynamic hierarchical OT framework to large single-cell datasets may be achieved using previously mentioned approaches that embed samples such that Euclidean distances in the embedding space approximate OT distances on the original space [Courty et al., 2017, Haviv et al., 2024]. A related approach has been proposed in the context of single-cell trajectory inference to study flows for a geodesic ground cost: [Huguet et al., 2022] embed cells in a space where Euclidean distances approximate geodesic distances on a graph, allowing them to use the $\mathbb{W}_2$ distance.

Using OT distances as a cost between cells raises the question of ground metric learning in the presence of several time points, as explored for a single time point in Chapter 3. Recently developed single-cell sequencing technologies like lineage tracing [McKenna et al., 2016], time-resolved transcriptomics [Erhard et al., 2022], and Live-seq [Chen et al., 2022b] enable the drawing of links between single-cell profiles at consecutive time points. This additional information provides valuable input for learning a potential function and constitutes an observed coupling $\hat{\mathbf{P}}$ between two populations. A promising research direction is the extension of inverse OT (see Section 1.3.1) to a hierarchical setting, i.e., learning a cost between cells, which is itself an OT distance.

Finally, extending the approach proposed in Chapter 5 to paired multi-omics or spatial paired multiomics data profiled through time could be achieved by operating on a joint embedding of the data, such as the method proposed in Chapter 4. [Huguet et al., 2022] learn dynamics of an embedding of the data but do not allow for several omics, and the embedding does not use the temporal information. [Velten et al., 2022] learn a time-informed multiomics dimensionality reduction but do not model population dynamics. Developing an end-to-end framework for joint dimensionality reduction and trajectory inference is thus critical to take full advantage of emerging paired single-cell multimodal sequencing data. As mentioned in the previous paragraphs, such a unified framework involves OT at the level of cells and the level of features, opening up exciting opportunities at the interface between metric learning, dimensionality reduction, and trajectory inference.

Bibliography

[Abrishami et al., 2020] Abrishami, T., Guillen, N., Rule, P., Schutzman, Z., Solomon, J., Weighill, T., and Wu, S. (2020). Geometry of graph partitions via optimal transport. *SIAM Journal on Scientific Computing*, 42(5):A3340–A3366. (Cited on page 53.)

[Adossa et al., 2021] Adossa, N., Khan, S., Rytkönen, K. T., and Elo, L. L. (2021). Computational strategies for single-cell multi-omics integration. *Computational and Structural Biotechnology Journal*, 19:2588–2596. (Cited on page 110.)

[Ahlmann-Eltze and Huber, 2023] Ahlmann-Eltze, C. and Huber, W. (2023). Comparison of transformations for single-cell rna-seq data. *Nature Methods*, 20(5):665–672. (Cited on pages 10 and 144.)

[Ainsworth et al., 2018] Ainsworth, S. K., Foti, N. J., Lee, A. K., and Fox, E. B. (2018). oi-vae: Output interpretable vaes for nonlinear group factor analysis. In *International Conference on Machine Learning*, pages 119–128. PMLR. (Cited on pages 11, 67, and 145.)

[Akian et al., 2018] Akian, M., Gaubert, S., and Hochart, A. (2018). A game theory approach to the existence and uniqueness of nonlinear perron-frobenius eigenvectors. *arXiv preprint arXiv:1812.09871*. (Cited on page 55.)

[Akian et al., 2016] Akian, M., Gaubert, S., and Nussbaum, R. (2016). Uniqueness of the fixed point of nonexpansive semidifferentiable maps. *Transactions of the American Mathematical Society*, 368(2):1271–1320. (Cited on pages 55 and 56.)

[Allen, 2015] Allen, M. (2015). Compelled by the diagram: thinking through ch waddington's epigenetic landscape. *Contemporaneity*, 4:119. (Cited on pages 22, 88, 89, and 157.)

[Altschuler et al., 2017] Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1961–1971. (Cited on pages 60 and 183.)

[Ambrosio et al., 2005] Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media. (Cited on pages 23, 101, and 158.)

[Amezquita et al., 2020] Amezquita, R. A., Lun, A. T., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2020). Orchestrating single-cell analysis with bioconductor. *Nature methods*, 17(2):137–145. (Cited on pages 9 and 143.)

[Andersson et al., 2020] Andersson, A., Bergenstråhle, J., Asp, M., Bergenstråhle, L., Jurek, A., Fernández Navarro, J., and Lundeberg, J. (2020). Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications biology*, 3(1):565. (Cited on page 97.)

[Angermueller et al., 2016] Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods*, 13(3):229–232. (Cited on page 66.)

[Anlauf and Derouiche, 2013] Anlauf, E. and Derouiche, A. (2013). Glutamine synthetase as an astrocytic marker: its cell type and vesicle localization. *Frontiers in endocrinology*, 4:144. (Cited on page 96.)

[Argelaguet et al., 2020] Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. (2020). Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21:1–17. (Cited on pages 15, 67, 69, 73, 74, 81, and 149.)

[Argelaguet et al., 2021] Argelaguet, R., Cuomo, A. S., Stegle, O., and Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nature biotechnology*, 39(10):1202–1215. (Cited on page 110.)

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR. (Cited on page 52.)

[Ashuach et al., 2023] Ashuach, T., Gabitto, M. I., Koodli, R. V., Saldi, G.-A., Jordan, M. I., and Yosef, N. (2023). Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231. (Cited on pages 15, 67, 69, and 149.)

[Baek and Lee, 2020] Baek, S. and Lee, I. (2020). Single-cell atac sequencing analysis: from data preprocessing to hypothesis generation. *Computational and structural biotechnology journal*, 18:1429–1439. (Cited on page 9.)

[Ballas and Mandel, 2005] Ballas, N. and Mandel, G. (2005). The many faces of rest oversee epigenetic programming of neuronal genes. *Current opinion in neurobiology*, 15(5):500–506. (Cited on page 95.)

[Bartee et al., 2004] Bartee, E., Mansouri, M., Hovey Nerenberg, B. T., Gouveia, K., and Früh, K. (2004). Downregulation of major histocompatibility complex class i by human ubiquitin ligases related to viral immune evasion proteins. *Journal of virology*, 78(3):1109–1120. (Cited on page 75.)

[Baysoy et al., 2023] Baysoy, A., Bai, Z., Satija, R., and Fan, R. (2023). The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, 24(10):695–713. (Cited on page 109.)

[Beirlant et al., 1997] Beirlant, J., Dudewicz, E. J., Györfi, L., Van der Meulen, E. C., et al. (1997). Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39. (Cited on pages 26 and 161.)

[Bellazzi et al., 2021] Bellazzi, R., Codegoni, A., Gualandi, S., Nicora, G., and Vercesi, E. (2021). The gene mover's distance: Single-cell similarity via optimal transport. *arXiv preprint arXiv:2102.01218*. (Cited on pages 20, 27, 33, 37, 51, 52, 62, 155, and 162.)

[Bellet et al., 2013] Bellet, A., Habrard, A., and Sebban, M. (2013). A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*. (Cited on page 52.)

[Benamou et al., 2016] Benamou, J.-D., Carlier, G., Mérigot, Q., and Oudet, E. (2016). Discretization of functionals involving the monge–ampère operator. *Numerische mathematik*, 134(3):611–636. (Cited on pages 26 and 161.)

[Benmansour et al., 2010] Benmansour, F., Carlier, G., Peyré, G., and Santambrogio, F. (2010). Derivatives with respect to metrics and applications: subgradient marching algorithm. *Numerische Mathematik*, 116(3):357–381. (Cited on page 53.)

[Bergen et al., 2020] Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12):1408–1414. (Cited on pages 11, 15, 89, 146, and 150.)

[Bergen et al., 2021] Bergen, V., Soldatov, R. A., Kharchenko, P. V., and Theis, F. J. (2021). Rna velocity—current challenges and future perspectives. *Molecular systems biology*, 17(8):e10282. (Cited on pages 11, 88, and 146.)

[Bernard et al., 1999] Bernard, K., Cambiaggi, A., Guia, S., Bertucci, F., Granjeaud, S., Tagett, R., N'Guyen, C., Jordan, B. R., and Vivier, E. (1999). Engagement of natural cytotoxicity programs regulates ap-1 expression in the nkl human nk cell line. *The Journal of Immunology*, 162(7):4062–4068. (Cited on page 77.)

[Beyer et al., 1999] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*, pages 217–235. Springer. (Cited on pages 11 and 145.)

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008. (Cited on pages 11, 39, 68, and 145.)

[Bogachev and Ruas, 2007] Bogachev, V. I. and Ruas, M. A. S. (2007). *Measure theory*, volume 1. Springer. (Cited on pages 20 and 155.)

[Boles et al., 2005] Boles, K. S., Barchet, W., Diacovo, T., Cella, M., and Colonna, M. (2005). The tumor suppressor tslc1/necl-2 triggers nk-cell and cd8+ t-cell responses through the cell-surface receptor crtam. *Blood*, 106(3):779–786. (Cited on page 75.)

[Bonneel et al., 2011] Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W. (2011). Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12. (Cited on page 54.)

[Bonnet et al., 2018] Bonnet, F., Molina, A., Roussat, M., Azais, M., Bel-Vialar, S., Gautrais, J., Pituello, F., and Agius, E. (2018). Neurogenic decisions require a cell cycle independent function of the cdc25b phosphatase. *Elife*, 7:e32937. (Cited on page 94.)

[Bouland et al., 2023] Bouland, G. A., Mahfouz, A., and Reinders, M. J. (2023). Consequences and opportunities arising due to sparser single-cell rna-seq datasets. *Genome biology*, 24(1):86. (Cited on page 9.)

[Braccioli et al., 2018] Braccioli, L., Vervoort, S. J., Puma, G., Nijboer, C. H., and Coffer, P. J. (2018). Sox4 inhibits oligodendrocyte differentiation of embryonic neural stem cells in vitro by inducing hes5 expression. *Stem Cell Research*, 33:110–119. (Cited on page 97.)

[Bredikhin et al., 2022] Bredikhin, D., Kats, I., and Stegle, O. (2022). Muon: multimodal omics analysis framework. *Genome Biology*, 23(1):42. (Cited on pages 31, 69, 80, 82, and 166.)

[Brogat-Motte et al., 2022] Brogat-Motte, L., Flamary, R., Brouard, C., Rousu, J., and d'Alché Buc, F. (2022). Learning to predict graphs with fused gromov-wasserstein barycenters. In *International Conference on Machine Learning*, pages 2321–2335. PMLR. (Cited on pages 25 and 160.)

[Brunet et al., 2004] Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169. (Cited on page 77.)

[Buenrostro et al., 2013] Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218. (Cited on pages 7, 8, 142, and 143.)

[Bunne et al., 2019] Bunne, C., Alvarez-Melis, D., Krause, A., and Jegelka, S. (2019). Learning generative models across incomparable spaces. *arXiv preprint arXiv:1905.05461*. (Cited on pages 25, 102, and 160.)

[Bunne et al., 2022] Bunne, C., Papaxanthos, L., Krause, A., and Cuturi, M. (2022). Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR. (Cited on pages 24, 27, 88, 89, 90, 92, 101, 102, 103, 159, 161, and 233.)

[Burger et al., 2009] Burger, M., Carrillo de la Plata, J. A., and Wolfram, M.-T. (2009). A mixed finite element method for nonlinear diffusion equations. (Cited on pages 26 and 160.)

[Burkhardt et al., 2022] Burkhardt, D., Bloom, J., Cannoodt, R., Luecken, M. D., Krishnaswamy, S., Lance, C., Pisco, A. O., and Theis, F. J. (2022). Multimodal single-cell integration across time, individuals, and batches. *NeurIPS Competitions*. (Cited on page 111.)

[Butler et al., 2018] Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420. (Cited on pages 9 and 143.)

[Cantini et al., 2021] Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature communications*, 12(1):124. (Cited on pages 15, 68, and 149.)

[Cao et al., 2018] 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. (Cited on pages 13, 66, and 148.)

[Cao et al., 2017] Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667. (Cited on pages 8 and 142.)

[Cao et al., 2022] Cao, K., Hong, Y., and Wan, L. (2022). Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics*, 38(1):211–219. (Cited on page 33.)

[Cao and Gao, 2022] Cao, Z.-J. and Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466. (Cited on pages 15 and 149.)

[Carlier et al., 2023] Carlier, G., Dupuy, A., Galichon, A., and Sun, Y. (2023). Sista: learning optimal transport costs under sparsity constraints. *Communications on Pure and Applied Mathematics*, 76(9):1659–1677. (Cited on pages 21, 53, and 156.)

[Carrillo et al., 2015] Carrillo, J. A., Chertock, A., and Huang, Y. (2015). A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17(1):233–258. (Cited on pages 26 and 160.)

[Carrillo et al., 2019] Carrillo, J. A., Craig, K., and Patacchini, F. S. (2019). A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58:1–53. (Cited on pages 26 and 161.)

[Castro-Mondragon et al., 2022] Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). Jaspar 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 50(D1):D165–D173. (Cited on page 83.)

[Chalise and Fridley, 2017] Chalise, P. and Fridley, B. L. (2017). Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PloS one*, 12(5):e0176278. (Cited on page 81.)

[Chander et al., 2019] Chander, P., Kennedy, M. J., Winckler, B., and Weick, J. P. (2019). Neuron-specific gene 2 (nsg2) encodes an ampa receptor interacting protein that modulates excitatory neurotransmission. *eneuro*, 6(1). (Cited on page 94.)

[Chen et al., 2022a] Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., et al. (2022a). Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792. (Cited on pages 14, 16, 87, 88, 91, 95, 99, 148, 149, and 151.)

[Chen et al., 2013] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14:1–14. (Cited on page 83.)

[Chen and Guan, 2022] Chen, J. and Guan, Z. (2022). Function of oncogene mycn in adult neurogenesis and oligodendrogenesis. *Molecular Neurobiology*, 59(1):77–92. (Cited on page 95.)

[Chen et al., 2015] Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090. (Cited on page 97.)

[Chen et al., 2019] Chen, S., Lake, B. B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457. (Cited on page 66.)

[Chen et al., 2022b] Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., et al. (2022b). Live-seq enables temporal transcriptomic recording of single cells. *Nature*, 608(7924):733–740. (Cited on page 111.)

[Chen et al., 2021] Chen, W., Zhao, Y., Chen, X., Yang, Z., Xu, X., Bi, Y., Chen, V., Li, J., Choi, H., Ernest, B., et al. (2021). A multicenter study benchmarking single-cell rna sequencing technologies using reference samples. *Nature Biotechnology*, 39(9):1103–1114. (Cited on page 35.)

[Chen et al., 2020] Chen, W.-T., Lu, A., Craessaerts, K., Pavie, B., Frigerio, C. S., Corthout, N., Qian, X., Laláková, J., Kühnemund, M., Voytyuk, I., et al. (2020). Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182(4):976–991. (Cited on pages 16, 87, and 151.)

[Cheng et al., 2022] Cheng, Y., Yin, Y., Zhang, A., Bernstein, A. M., Kawaguchi, R., Gao, K., Potter, K., Gilbert, H.-Y., Ao, Y., Ou, J., et al. (2022). Transcription factor network analysis identifies rest/nrsf as an intrinsic regulator of cns regeneration in mice. *Nature communications*, 13(1):4418. (Cited on page 95.)

[Chizat et al., 2018] Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2018). Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123. (Cited on page 63.)

[Chizat et al., 2020] Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020). Faster wasserstein distance estimation with the sinkhorn divergence. In *Proc. NeurIPS'20*. (Cited on page 60.)

[Clark et al., 2018] Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., et al. (2018). scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature communications*, 9(1):781. (Cited on page 66.)

[Collins et al., 2001] Collins, M., Dasgupta, S., and Schapire, R. E. (2001). A generalization of principal components analysis to the exponential family. *Advances in neural information processing systems*, 14. (Cited on pages 10 and 144.)

[Corces et al., 2016] Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics*, 48(10):1193–1203. (Cited on pages 35, 47, and 48.)

[Courty et al., 2017] Courty, N., Flamary, R., and Ducoffe, M. (2017). Learning wasserstein embeddings. *arXiv preprint arXiv:1710.07457*. (Cited on pages 109 and 111.)

[Cui et al., 2024] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. (2024). scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11. (Cited on page 109.)

[Cusanovich et al., 2015] Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914. (Cited on pages 8 and 143.)

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26. (Cited on pages 19, 25, 34, 37, 52, 59, 60, 79, 102, 154, and 159.)

[Cuturi and Avis, 2014] Cuturi, M. and Avis, D. (2014). Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564. (Cited on pages 20, 51, 52, and 155.)

[Cuturi et al., 2022] Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. (2022). Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*. (Cited on pages 24, 25, 31, 88, 103, 159, 160, and 166.)

[Cuturi and Peyré, 2016] Cuturi, M. and Peyré, G. (2016). A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343. (Cited on pages 25 and 160.)

[Davis and Dhillon, 2008] Davis, J. V. and Dhillon, I. S. (2008). Structured metric learning for high dimensional problems. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 195–203. (Cited on page 52.)

[De Rop et al., 2023] De Rop, F. V., Hulselmans, G., Flerin, C., Soler-Vila, P., Rafels, A., Christiaens, V., González-Blas, C. B., Marchese, D., Caratù, G., Poovathingal, S., et al. (2023). Systematic benchmarking of single-cell atac-sequencing protocols. *Nature biotechnology*, pages 1–11. (Cited on pages 8 and 143.)

[DeepMind et al., 2020] DeepMind, Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Dedieu, A., Fantacci, C., Godwin, J., Jones, C., Hemsley, R., Hennigan, T., Hessel, M., Hou, S., Kapturowski, S., Keck, T., Kemaev, I., King, M., Kunesch, M., Martens, L., Merzic, H., Mikulik, V., Norman, T., Papamakarios, G., Quan, J., Ring, R., Ruiz, F., Sanchez, A., Sartran, L., Schneider, R., Sezener, E., Spencer, S., Srinivasan, S., Stanojević, M., Stokowiec, W., Wang, L., Zhou, G., and Viola, F. (2020). The DeepMind JAX Ecosystem. (Cited on pages 31, 88, 104, and 166.)

[Demetci et al., 2022a] Demetci, P., Santorella, R., Chakravarthy, M., Sandstede, B., and Singh, R. (2022a). Scotv2: Single-cell multiomic alignment with disproportionate cell-type representation. *Journal of Computational Biology*, 29(11):1213–1228. (Cited on pages 28 and 162.)

[Demetci et al., 2022b] Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. (2022b). Scot: single-cell multi-omics alignment with optimal transport. *Journal of computational biology*, 29(1):3–18. (Cited on pages 6, 22, 28, 33, 141, 156, and 162.)

[Doñate et al., 2013] Doñate, C., Ody, C., McKee, T., Ruault-Jungblut, S., Fischer, N., Ropraz, P., Imhof, B. A., and Matthes, T. (2013). Homing of human b cells to lymphoid organs and b-cell lymphoma engraftment are controlled by cell adhesion molecule jam-c. *Cancer research*, 73(2):640–651. (Cited on page 76.)

[Dong et al., 2011] Dong, W., Moses, C., and Li, K. (2011). Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*, pages 577–586. (Cited on page 109.)

[Dries et al., 2021] Dries, R., Zhu, Q., Dong, R., Eng, C.-H. L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F., et al. (2021). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology*, 22:1–31. (Cited on pages 14 and 149.)

[Du et al., 2019] Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., and Zhi, D. (2019). Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20(1):7–15. (Cited on pages 20, 27, 51, 62, 64, 155, and 162.)

[Dukler et al., 2019] Dukler, Y., Li, W., Lin, A., and Montúfar, G. (2019). Wasserstein of wasserstein loss for learning generative models. In *International conference on machine learning*, pages 1716–1725. PMLR. (Cited on page 111.)

[Dupuy and Galichon, 2014] Dupuy, A. and Galichon, A. (2014). Personality traits and the marriage market. *Journal of Political Economy*, 122(6):1271–1319. (Cited on pages 21 and 156.)

[Dupuy et al., 2019] Dupuy, A., Galichon, A., and Sun, Y. (2019). Estimating matching affinity matrices under low-rank constraints. *Information and Inference: A Journal of the IMA*, 8(4):677–689. (Cited on page 53.)

[Duraikannu et al., 2018] Duraikannu, A., Martinez, J. A., Chandrasekhar, A., and Zochodne, D. W. (2018). Expression and manipulation of the apc-β-catenin pathway during peripheral neuron regeneration. *Scientific Reports*, 8(1):13197. (Cited on page 95.)

[Duren et al., 2022] Duren, Z., Chang, F., Naqing, F., Xin, J., Liu, Q., and Wong, W. H. (2022). Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with screg. *Genome biology*, 23(1):114. (Cited on pages 67 and 69.)

[Düsterwald and Yamada, 2024] Düsterwald, K. M. and Yamada, M. (2024). Unsupervised ground metric learning with tree wasserstein distance. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*. (Cited on page 109.)

[Economou et al., 2012] Economou, A. D., Ohazama, A., Porntaveetus, T., Sharpe, P. T., Kondo, S., Basson, M. A., Gritli-Linde, A., Cobourne, M. T., and Green, J. B. (2012). Periodic stripe formation by a turing mechanism operating at growth zones in the mammalian palate. *Nature genetics*, 44(3):348–351. (Cited on page 98.)

[Erhard et al., 2022] Erhard, F., Saliba, A.-E., Lusser, A., Toussaint, C., Hennig, T., Prusty, B. K., Kirschenbaum, D., Abadie, K., Miska, E. A., Friedel, C. C., et al. (2022). Time-resolved single-cell rna-seq using metabolic rna labelling. *Nature Reviews Methods Primers*, 2(1):77. (Cited on page 111.)

[Fang et al., 2021] Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A. K., Zhou, X., Xie, F., et al. (2021). Comprehensive analysis of single cell atac-seq data with snapatac. *Nature communications*, 12(1):1337. (Cited on pages 8, 10, 143, and 144.)

[Fatras et al., 2021] Fatras, K., Zine, Y., Majewski, S., Flamary, R., Gribonval, R., and Courty, N. (2021). Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*. (Cited on pages 103 and 110.)

[Feydy et al., 2019] Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR. (Cited on pages 20, 24, 25, 38, 60, 154, and 159.)

[Flamary et al., 2021] Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8. (Cited on pages 24, 31, 53, 62, 159, 165, and 166.)

[Flamary et al., 2018] Flamary, R., Cuturi, M., Courty, N., and Rakotomamonjy, A. (2018). Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945. (Cited on page 53.)

[Frank et al., 1956] Frank, M., Wolfe, P., et al. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110. (Cited on pages 25 and 160.)

[Frogner et al., 2015] Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015). Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061. (Cited on page 51.)

[Galichon and Salanié, 2020] Galichon, A. and Salanié, B. (2020). Cupid's invisible hand: Social surplus and identification in matching models. *Available at SSRN 1804623*. (Cited on page 53.)

[Gao et al., 2024] Gao, L. L., Bien, J., and Witten, D. (2024). Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342. (Cited on pages 11 and 145.)

[Gaublomme et al., 2015] Gaublomme, J. T., Yosef, N., Lee, Y., Gertner, R. S., Yang, L. V., Wu, C., Pandolfi, P. P., Mak, T., Satija, R., Shalek, A. K., et al. (2015). Single-cell genomics unveils critical regulators of th17 cell pathogenicity. *Cell*, 163(6):1400–1412. (Cited on pages 12 and 147.)

[Gayoso et al., 2022] Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., et al. (2022). A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology*, 40(2):163–166. (Cited on pages 11, 15, 61, 62, 145, and 149.)

[Gayoso et al., 2021] Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., and Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature methods*, 18(3):272–282. (Cited on pages 15, 67, 69, and 149.)

[Genevay et al., 2019] Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR. (Cited on pages 37, 38, 52, and 60.)

[Genevay et al., 2018] Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. (Cited on pages 19, 20, 53, 60, 102, 154, and 155.)

[Genshaft et al., 2016] Genshaft, A. S., Li, S., Gallant, C. J., Darmanis, S., Prakadan, S. M., Ziegler, C. G., Lundberg, M., Fredriksson, S., Hong, J., Regev, A., et al. (2016). Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome biology*, 17:1–15. (Cited on pages 16 and 151.)

[Glass et al., 2020] Glass, D. R., Tsai, A. G., Oliveria, J. P., Hartmann, F. J., Kimmey, S. C., Calderon, A. A., Borges, L., Glass, M. C., Wagar, L. E., Davis, M. M., et al. (2020). An integrated multi-omic single-cell atlas of human b cell identity. *Immunity*, 53(1):217–232. (Cited on pages 75, 76, and 77.)

[Gong et al., 2021] Gong, B., Zhou, Y., and Purdom, E. (2021). Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome biology*, 22:1–21. (Cited on pages 15, 67, 69, 81, and 149.)

[Goodwin et al., 2016] Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics*, 17(6):333–351. (Cited on pages 7 and 142.)

[Gorin et al., 2020] Gorin, G., Svensson, V., and Pachter, L. (2020). Protein velocity and acceleration from single-cell multiomics experiments. *Genome biology*, 21:1–6. (Cited on pages 15 and 150.)

[Gramfort et al., 2015] Gramfort, A., Peyré, G., and Cuturi, M. (2015). Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer. (Cited on page 52.)

[Gretton et al., 2012] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773. (Cited on page 61.)

[Guo et al., 2015a] Guo, H., Cruz-Munoz, M.-E., Wu, N., Robbins, M., and Veillette, A. (2015a). Immune cell inhibition by slamf7 is mediated by a mechanism requiring src kinases, cd45, and ship-1 that is defective in multiple myeloma cells. *Molecular and cellular biology*. (Cited on page 77.)

[Guo et al., 2015b] Guo, M., Wang, H., Potter, S. S., Whitsett, J. A., and Xu, Y. (2015b). Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS computational biology*, 11(11):e1004575. (Cited on pages 33 and 39.)

[Gury-BenAri et al., 2016] Gury-BenAri, M., Thaiss, C. A., Serafini, N., Winter, D. R., Giladi, A., Lara-Astiaso, D., Levy, M., Salame, T. M., Weiner, A., David, E., et al. (2016). The spectrum and regulatory landscape of intestinal innate lymphoid cells are shaped by the microbiome. *Cell*, 166(5):1231–1246. (Cited on pages 12 and 146.)

[Hafemeister and Satija, 2019] Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):296. (Cited on page 35.)

[Haghverdi et al., 2016] Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845–848. (Cited on pages 11, 12, 15, 146, 147, and 150.)

[Han et al., 2014] Han, A., Glanville, J., Hansmann, L., and Davis, M. M. (2014). Linking t-cell receptor sequence to functional phenotype at the single-cell level. *Nature biotechnology*, 32(7):684–692. (Cited on pages 13 and 148.)

[Han et al., 2018] Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., et al. (2018). Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386. (Cited on pages 94, 97, and 107.)

[Hao et al., 2021] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587. (Cited on pages 15, 16, 62, 67, 69, 71, 81, 149, 150, and 184.)

[Hashimoto et al., 2016] Hashimoto, T., Gifford, D., and Jaakkola, T. (2016). Learning population-level diffusions with generative rnns. In *International Conference on Machine Learning*, pages 2417–2426. PMLR. (Cited on pages 6, 12, 17, 24, 27, 28, 52, 88, 89, 90, 92, 101, 102, 103, 141, 146, 151, 159, 161, 163, and 233.)

[Hausrat et al., 2021] Hausrat, T. J., Radwitz, J., Lombino, F. L., Breiden, P., and Kneussel, M. (2021). Alpha-and beta-tubulin isotypes are differentially expressed during brain development. *Developmental neurobiology*, 81(3):333–350. (Cited on page 95.)

[Haviv et al., 2024] Haviv, D., Kunes, R. Z., Dougherty, T., Burdziak, C., Nawy, T., Gilbert, A., and Pe'er, D. (2024). Wasserstein wormhole: Scalable optimal transport distance with transformers. *arXiv preprint arXiv:2404.09411*. (Cited on pages 109 and 111.)

[He and Sun, 2007] He, F. and Sun, Y. E. (2007). Glial cells more than support cells? *The international journal of biochemistry & cell biology*, 39(4):661–665. (Cited on page 95.)

[Heek et al., 2023] Heek, J., Levskaya, A., Oliver, A., Ritter, M., Rondepierre, B., Steiner, A., and van Zee, M. (2023). Flax: A neural network library and ecosystem for JAX. (Cited on page 103.)

[Heitz et al., 2020] Heitz, M., Bonneel, N., Coeurjolly, D., Cuturi, M., and Peyré, G. (2020). Ground metric learning on graphs. *Journal of Mathematical Imaging and Vision*, pages 1–19. (Cited on page 53.)

[Heumos et al., 2023] Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., et al. (2023). Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572. (Cited on pages 8, 9, 10, 11, 13, 15, 143, 144, 145, 147, and 149.)

[Hirvonen et al., 1994] Hirvonen, H., Salonen, R., Sandberg, M., Vuorio, E., Västrik, I., Kotilainen, E., and Kalimo, H. (1994). Differential expression of myc, max and rb1 genes in human gliomas and glioma cell lines. *British Journal of Cancer*, 69(1):16–25. (Cited on page 97.)

[Hivert et al., 2024] Hivert, B., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2024). Post-clustering difference testing: Valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, 193:107916. (Cited on pages 11 and 145.)

[Hooke et al., 1665] Hooke, R., Allestry, J., and Martyn, J. (1665). *Micrographia, or, Some physiological descriptions of minute bodies made by magnifying glasses :with observations and inquiries thereupon*. London, Printed by Jo. Martyn and Ja. Allestry, printers to the Royal Society, 1665. https://www.biodiversitylibrary.org/bibliography/904. (Cited on pages 6 and 141.)

[Hou and Ji, 2022] Hou, W. and Ji, Z. (2022). Decomposing spatial heterogeneity of cell trajectories with paella. *bioRxiv*, pages 2022–09. (Cited on pages 15 and 150.)

[Hoyer, 2004] Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(9). (Cited on page 80.)

[Huang et al., 2016] Huang, G., Quo, C., Kusner, M. J., Sun, Y., Weinberger, K. Q., and Sha, F. (2016). Supervised word mover's distance. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4869–4877. (Cited on page 53.)

[Huang et al., 2007] Huang, S., Guo, Y.-P., May, G., and Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental biology*, 305(2):695–713. (Cited on pages 22, 88, 97, and 157.)

[Huang et al., 2018] Huang, X.-t., Li, X., Qin, P.-z., Zhu, Y., Xu, S.-n., and Chen, J.-p. (2018). Technical advances in single-cell rna sequencing and applications in normal and malignant hematopoiesis. *Frontiers in oncology*, 8:582. (Cited on pages 8 and 143.)

[Hubert and Schultz, 1976] Hubert, L. and Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British journal of mathematical and statistical psychology*, 29(2):190–241. (Cited on page 39.)

[Huguet et al., 2022] Huguet, G., Magruder, D. S., Tong, A., Fasina, O., Kuchroo, M., Wolf, G., and Krishnaswamy, S. (2022). Manifold interpolating optimal-transport flows for trajectory inference. *Advances in neural information processing systems*, 35:29705–29718. (Cited on pages 88 and 111.)

[Huizing et al., 2022] Huizing, G.-J., Peyré, G., and Cantini, L. (2022). Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics*, 38(8):2169–2177. (Cited on pages 52, 62, 67, 69, 78, and 79.)

[Hunter et al., 2021] Hunter, M. V., Moncada, R., Weiss, J. M., Yanai, I., and White, R. M. (2021). Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nature communications*, 12(1):6278. (Cited on page 87.)

[Intlekofer et al., 2005] Intlekofer, A. M., Takemoto, N., Wherry, E. J., Longworth, S. A., Northrup, J. T., Palanivel, V. R., Mullen, A. C., Gasink, C. R., Kaech, S. M., Miller, J. D., et al. (2005). Effector and memory cd8+ t cell fate coupled by t-bet and eomesodermin. *Nature immunology*, 6(12):1236–1244. (Cited on page 75.)

[Iqbal and Zhou, 2023] Iqbal, W. and Zhou, W. (2023). Computational methods for single-cell dna methylome analysis. *Genomics, Proteomics and Bioinformatics*, 21(1):48–66. (Cited on pages 10 and 144.)

[Islam et al., 2011] Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, 21(7):1160–1167. (Cited on pages 8 and 142.)

[Jaitin et al., 2014] Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779. (Cited on pages 11, 12, 145, and 146.)

[Janati et al., 2020] Janati, H., Cuturi, M., and Gramfort, A. (2020). Spatio-temporal alignments: Optimal transport through space and time. In *International Conference on Artificial Intelligence and Statistics*, pages 1695–1704. PMLR. (Cited on page 52.)

[Jhamb et al., 2011] Jhamb, D., Rao, N., Milner, D. J., Song, F., Cameron, J. A., Stocum, D. L., and Palakal, M. J. (2011). Network based transcription factor analysis of regenerating axolotl limbs. *BMC bioinformatics*, 12:1–12. (Cited on page 95.)

[Jin et al., 2020] Jin, S., Zhang, L., and Nie, Q. (2020). scai: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome biology*, 21:1–19. (Cited on pages 67 and 69.)

[Jordan et al., 1998] Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17. (Cited on pages 23, 89, and 158.)

[Jurga et al., 2021] Jurga, A. M., Paleczna, M., Kadluczka, J., and Kuter, K. Z. (2021). Beyond the gfap-astrocyte protein markers in the brain. *Biomolecules*, 11(9):1361. (Cited on page 96.)

[Kaileh and Sen, 2012] Kaileh, M. and Sen, R. (2012). Nf-κb function in b lymphocytes. *Immunological reviews*, 246(1):254–271. (Cited on page 75.)

[Kantorovich, 1942] Kantorovich, L. (1942). On the transfer of masses (in Russian). *Doklady Akademii Nauk*, 37(2):227–229. (Cited on pages 16, 18, 36, 51, 52, 78, 151, and 153.)

[Kawakami et al., 2006] Kawakami, Y., Esteban, C. R., Raya, M., Kawakami, H., Martí, M., Dubova, I., and Belmonte, J. C. I. (2006). Wnt/β-catenin signaling regulates vertebrate limb regeneration. *Genes & development*, 20(23):3232–3237. (Cited on page 95.)

[Kerdoncuff et al., 2021] Kerdoncuff, T., Emonet, R., and Sebban, M. (2021). Metric Learning in Optimal Transport for Domain Adaptation. In *International Joint Conference on Artificial Intelligence*, Kyoto, Japan. (Cited on page 52.)

[Kharchenko, 2021] Kharchenko, P. V. (2021). The triumphs and limitations of computational methods for scrna-seq. *Nature methods*, 18(7):723–732. (Cited on pages 8 and 142.)

[Kim et al., 2020] Kim, H. J., Lin, Y., Geddes, T. A., Yang, J. Y. H., and Yang, P. (2020). Citefuse enables multi-modal analysis of cite-seq data. *Bioinformatics*, 36(14):4137–4143. (Cited on pages 67 and 69.)

[Kim et al., 2019] Kim, T., Chen, I. R., Lin, Y., Wang, A. Y.-Y., Yang, J. Y. H., and Yang, P. (2019). Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in bioinformatics*, 20(6):2316–2326. (Cited on pages 33 and 41.)

[Kinderlehrer and Walkington, 1999] Kinderlehrer, D. and Walkington, N. J. (1999). Approximation of parabolic equations using the wasserstein metric. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(4):837–852. (Cited on pages 26 and 160.)

[Kiselev et al., 2019] Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282. (Cited on pages 11, 33, and 145.)

[Klein et al., 2015] Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201. (Cited on pages 8 and 142.)

[Klein et al., 2023] Klein, D., Palla, G., Lange, M., Klein, M., Piran, Z., Gander, M., Meng-Papaxanthos, L., Sterr, M., Bastidas-Ponce, A., Tarquis-Medina, M., et al. (2023). Mapping cells through time and space with moscot. *bioRxiv*, pages 2023–05. (Cited on pages 15, 22, 28, 88, 104, 105, 110, 150, 157, 162, and 163.)

[Klemm et al., 2019] Klemm, S. L., Shipony, Z., and Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220. (Cited on pages 8 and 143.)

[Kleshchevnikov et al., 2022] Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., et al. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology*, 40(5):661–671. (Cited on page 97.)

[Kloeden et al., 1992] Kloeden, P. E., Platen, E., Kloeden, P. E., and Platen, E. (1992). *Stochastic differential equations*. Springer. (Cited on pages 26 and 161.)

[Kobak and Berens, 2019] Kobak, D. and Berens, P. (2019). The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):5416. (Cited on pages 10 and 145.)

[Korhonen et al., 2017] Korhonen, J. H., Palin, K., Taipale, J., and Ukkonen, E. (2017). Fast motif matching revisited: high-order pwms, snps and indels. *Bioinformatics*, 33(4):514–521. (Cited on page 68.)

[Korsunsky et al., 2019] Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296. (Cited on page 100.)

[Kotliar et al., 2019] Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., and Sabeti, P. C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *Elife*, 8:e43803. (Cited on page 77.)

[Kuhn, 1955] Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97. (Cited on pages 25 and 159.)

[Kulis et al., 2012] Kulis, B. et al. (2012). Metric learning: A survey. *Foundations and trends in machine learning*, 5(4):287–364. (Cited on page 52.)

[Kusner et al., 2015] Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. Q. (2015). From word embeddings to document distances. In *Proc. of the 32nd Intern. Conf. on Machine Learning*, pages 957–966. (Cited on pages 20, 51, 52, 62, and 155.)

[La Manno et al., 2018] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018). Rna velocity of single cells. *Nature*, 560(7719):494–498. (Cited on pages 11, 15, 88, 89, 146, and 150.)

[Lähnemann et al., 2020] 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 biology*, 21:1–35. (Cited on pages 9, 11, 71, and 145.)

[Laidlaw and Cyster, 2021] Laidlaw, B. J. and Cyster, J. G. (2021). Transcriptional regulation of memory b cell differentiation. *Nature Reviews Immunology*, 21(4):209–220. (Cited on page 77.)

[Lance et al., 2022] Lance, C., Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Rautenstrauch, P., Laddach, A., Ubingazhibov, A., Cao, Z.-J., Deng, K., Khan, S., et al. (2022). Multimodal single cell data integration challenge: results and lessons learned. *BioRxiv*, pages 2022–04. (Cited on page 71.)

[Lange et al., 2022] Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H. B., Pe'er, D., and Theis, F. J. (2022). Cellrank for directed single-cell fate mapping. *Nat. Methods*. (Cited on pages 12 and 146.)

[Lanier, 2015] Lanier, L. L. (2015). Nkg2d receptor and its ligands in host defense. *Cancer immunology research*, 3(6):575–582. (Cited on page 75.)

[Lareau et al., 2019] Lareau, C. A., Duarte, F. M., Chew, J. G., Kartha, V. K., Burkett, Z. D., Kohlway, A. S., Pokholok, D., Aryee, M. J., Steemers, F. J., Lebofsky, R., et al. (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology*, 37(8):916–924. (Cited on pages 8 and 143.)

[Larsson et al., 2021] Larsson, L., Frisén, J., and Lundeberg, J. (2021). Spatially resolved transcriptomics adds a new dimension to genomics. *Nature methods*, 18(1):15–18. (Cited on pages 13 and 148.)

[Laug et al., 2018] Laug, D., Glasgow, S. M., and Deneen, B. (2018). A glial blueprint for gliomagenesis. *Nature Reviews Neuroscience*, 19(7):393–403. (Cited on page 95.)

[Lause et al., 2021] Lause, J., Berens, P., and Kobak, D. (2021). Analytic pearson residuals for normalization of single-cell rna-seq umi data. *Genome biology*, 22:1–20. (Cited on pages 10, 46, and 144.)

[Le and Cuturi, 2015] Le, T. and Cuturi, M. (2015). Unsupervised riemannian metric learning for histograms using aitchison transformations. In *International Conference on Machine Learning*, pages 2002–2011. PMLR. (Cited on page 52.)

[Le Roux et al., 2015] Le Roux, J., Weninger, F. J., and Hershey, J. R. (2015). Sparse nmf–half-baked or well done? *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, 11:13–15. (Cited on page 80.)

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791. (Cited on pages 11, 67, 69, and 145.)

[Lee et al., 2020] Lee, J., Hyeon, D. Y., and Hwang, D. (2020). Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine*, 52(9):1428–1442. (Cited on pages 33 and 66.)

[Lemmens and Nussbaum, 2012] Lemmens, B. and Nussbaum, R. (2012). *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press. (Cited on page 54.)

[Leng et al., 2021] Leng, K., Li, E., Eser, R., Piergies, A., Sit, R., Tan, M., Neff, N., Li, S. H., Rodriguez, R. D., Suemoto, C. K., et al. (2021). Molecular characterization of selectively vulnerable neurons in alzheimer's disease. *Nature neuroscience*, 24(2):276–287. (Cited on pages 12 and 147.)

[Li et al., 2023a] Li, C., Virgilio, M. C., Collins, K. L., and Welch, J. D. (2023a). Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nature biotechnology*, 41(3):387–398. (Cited on pages 15 and 150.)

[Li et al., 2017] Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., Kong, S. L., Chua, C., Hon, L. K., Tan, W. S., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics*, 49(5):708–718. (Cited on pages 35, 41, 47, and 48.)

[Li et al., 2019] Li, R., Ye, X., Zhou, H., and Zha, H. (2019). Learning to match via inverse optimal transport. *Journal of machine learning research*, 20. (Cited on page 53.)

[Li et al., 2023b] Li, Y. E., Preissl, S., Miller, M., Johnson, N. D., Wang, Z., Jiao, H., Zhu, C., Wang, Z., Xie, Y., Poirion, O., et al. (2023b). A comparative atlas of single-cell chromatin accessibility in the human brain. *Science*, 382(6667):eadf7044. (Cited on pages 12 and 147.)

[Liero et al., 2015] Liero, M., Mielke, A., and Savaré, G. (2015). Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, pages 1–149. (Cited on page 63.)

[Liu et al., 2022] Liu, C., Li, R., Li, Y., Lin, X., Zhao, K., Liu, Q., Wang, S., Yang, X., Shi, X., Ma, Y., et al. (2022). Spatiotemporal mapping of gene expression landscapes and developmental trajectories during zebrafish embryogenesis. *Developmental Cell*, 57(10):1284–1298. (Cited on pages 16, 87, 88, 91, 99, and 151.)

[Liu et al., 2019] Liu, L., Liu, C., Quintero, A., Wu, L., Yuan, Y., Wang, M., Cheng, M., Leng, L., Xu, L., Dong, G., et al. (2019). Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nature communications*, 10(1):470. (Cited on pages 13, 35, 41, 43, 47, 48, 66, 69, 71, 73, 84, and 148.)

[Liu et al., 2021] Liu, Z., Liu, L., Qi, Y., Li, H., and Pan, S. (2021). Glis family zinc finger 3 promoting cell malignant behaviors and nf-κb signaling in glioma. *Brain Research*, 1770:147623. (Cited on page 96.)

[Loshchilov and Hutter, 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. (Cited on page 104.)

[Lotfollahi et al., 2024] Lotfollahi, M., Hao, Y., Theis, F. J., and Satija, R. (2024). The future of rapid and automated single-cell data analysis using reference mapping. *Cell*, 187(10):2343–2358. (Cited on pages 12 and 147.)

[Lotfollahi et al., 2022] Lotfollahi, M., Litinetskaya, A., and Theis, F. J. (2022). Multigrate: single-cell multi-omic data integration. *BioRxiv*, pages 2022–03. (Cited on pages 67, 69, and 81.)

[Luecken et al., 2021] Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Lance, C., Agrawal, A., Aliee, H., Chen, A. T., Deconinck, L., Detweiler, A. M., Granados, A. A., et al. (2021). A sandbox for prediction and integration of dna, rna, and proteins in single cells. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*. (Cited on pages 71 and 85.)

[Luecken et al., 2022] Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50. (Cited on page 71.)

[Luecken and Theis, 2019] Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746. (Cited on pages 9, 10, 11, 33, 35, 39, 44, 51, 144, and 145.)

[Lukin et al., 2008] Lukin, K., Fields, S., Hartley, J., and Hagman, J. (2008). Early b cell factor: Regulator of b lineage specification and commitment. In *Seminars in immunology*, volume 20, pages 221–227. Elsevier. (Cited on page 75.)

[Lun et al., 2016] Lun, A. T., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5. (Cited on page 35.)

[Luo et al., 2017] Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J. R., Sandoval, J. P., et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351):600–604. (Cited on pages 8, 9, 35, 47, 48, and 143.)

[Luo et al., 2018] Luo, C., Rivkin, A., Zhou, J., Sandoval, J. P., Kurihara, L., Lucero, J., Castanon, R., Nery, J. R., Pinto-Duarte, A., Bui, B., et al. (2018). Robust single-cell dna methylome profiling with snmc-seq2. *Nature communications*, 9(1):3824. (Cited on pages 8 and 143.)

[Ma et al., 2020a] Ma, A., McDermaid, A., Xu, J., Chang, Y., and Ma, Q. (2020a). Integrative methods and practical challenges for single-cell multi-omics. *Trends in biotechnology*, 38(9):1007–1022. (Cited on page 33.)

[Ma et al., 2020b] Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., et al. (2020b). Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116. (Cited on pages 16 and 151.)

[Macosko et al., 2015] Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214. (Cited on pages 8, 33, 142, and 143.)

[Mairal et al., 2009] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009). Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696. (Cited on page 109.)

[Marbach et al., 2016] Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature methods*, 13(4):366–370. (Cited on page 84.)

[Marchetti, 2018] Marchetti, B. (2018). Wnt/β-catenin signaling pathway governs a full program for dopaminergic neuron survival, neurorescue and regeneration in the mptp mouse model of parkinson's disease. *International journal of molecular sciences*, 19(12):3743. (Cited on page 95.)

[Marconato et al., 2024] Marconato, L., Palla, G., Yamauchi, K. A., Virshup, I., Heidari, E., Treis, T., Vierdag, W.-M., Toth, M., Stockhaus, S., Shrestha, R. B., et al. (2024). Spatialdata: an open and universal data framework for spatial omics. *Nature Methods*, pages 1–5. (Cited on pages 14 and 149.)

[Martens et al., 2024] Martens, L. D., Fischer, D. S., Yépez, V. A., Theis, F. J., and Gagneur, J. (2024). Modeling fragment counts improves single-cell atac-seq analysis. *Nature Methods*, 21(1):28–31. (Cited on pages 10 and 144.)

[Marx, 2021] Marx, V. (2021). Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14. (Cited on pages 13 and 148.)

[Massoni-Badosa et al., 2024] Massoni-Badosa, R., Aguilar-Fernández, S., Nieto, J. C., Soler-Vila, P., Elosua-Bayes, M., Marchese, D., Kulis, M., Vilas-Zornoza, A., Bühler, M. M., Rashmi, S., et al. (2024). An atlas of cells in the human tonsil. *Immunity*, 57(2):379–399. (Cited on page 110.)

[McInnes et al., 2018] McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. (Cited on pages 10, 68, 70, 83, and 145.)

[McKenna et al., 2016] McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907. (Cited on page 111.)

[Mémoli, 2011] Mémoli, F. (2011). Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487. (Cited on pages 21, 102, and 156.)

[Mena and Weed, 2019] Mena, G. and Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *arXiv preprint arXiv:1905.11882*. (Cited on page 52.)

[Mereu et al., 2020] Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D. J., Álvarez-Varela, A., Batlle, E., Sagar, n., Gruen, D., Lau, J. K., et al. (2020). Benchmarking single-cell rna-sequencing protocols for cell atlas projects. *Nature biotechnology*, 38(6):747–755. (Cited on pages 8 and 142.)

[Miao et al., 2021] Miao, Z., Humphreys, B. D., McMahon, A. P., and Kim, J. (2021). Multi-omics integration in the age of million single-cell data. *Nature Reviews Nephrology*, 17(11):710–724. (Cited on pages 66 and 67.)

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. (Cited on pages 20 and 51.)

[Mimitou et al., 2019] Mimitou, E. P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., et al. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and crispr perturbations in single cells. *Nature methods*, 16(5):409–412. (Cited on page 66.)

[Miras-Portugal et al., 2016] Miras-Portugal, M. T., Gomez-Villafuertes, R., Gualix, J., Diaz-Hernandez, J. I., Artalejo, A. R., Ortega, F., Delicado, E. G., and Perez-Sen, R. (2016). Nucleotides in neuroregeneration and neuroprotection. *Neuropharmacology*, 104:243–254. (Cited on page 95.)

[Monge, 1781] Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704. (Cited on pages 16, 17, 36, 51, 52, 67, 78, 151, and 152.)

[Moon et al., 2018] Moon, K. R., Stanley III, J. S., Burkhardt, D., van Dijk, D., Wolf, G., and Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46. (Cited on pages 11 and 145.)

[Moon et al., 2019] Moon, K. R., Van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. v. d., Hirn, M. J., Coifman, R. R., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12):1482–1492. (Cited on pages 10 and 145.)

[Mootha et al., 2003] Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–273. (Cited on page 68.)

[Moriel et al., 2021] Moriel, N., Senel, E., Friedman, N., Rajewsky, N., Karaiskos, N., and Nitzan, M. (2021). Novosparc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature protocols*, 16(9):4177–4200. (Cited on pages 28 and 162.)

[Nawy, 2014] Nawy, T. (2014). Single-cell sequencing. *Nature methods*, 11(1):18–18. (Cited on pages 7 and 142.)

[Nemirovski et al., 2009] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609. (Cited on page 181.)

[Niles-Weed and Rigollet, 2019] Niles-Weed, J. and Rigollet, P. (2019). Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*. (Cited on page 53.)

[Noh, 2012] Noh, S. (2012). χ^2 metric learning for nearest neighbor classification and its analysis. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 991–995. IEEE. (Cited on page 52.)

[Nunez and Fischer, 1997] Nunez, J. and Fischer, I. (1997). Microtubule-associated proteins (maps) in the peripheral nervous system during development and regeneration. *Journal of Molecular Neuroscience*, 8:207–222. (Cited on page 94.)

[Ody et al., 2007] Ody, C., Jungblut-Ruault, S., Cossali, D., Barnet, M., Aurrand-Lions, M., Imhof, B., and Matthes, T. (2007). Junctional adhesion molecule c (jam-c) distinguishes cd27+ germinal center b lymphocytes from non-germinal center cells and constitutes a new diagnostic tool for b-cell malignancies. *Leukemia*, 21(6):1285–1293. (Cited on page 76.)

[Ohtsuka et al., 2001] Ohtsuka, T., Sakamoto, M., Guillemot, F., and Kageyama, R. (2001). Roles of the basic helix-loop-helix genes hes1 and hes5 in expansion of neural stem cells of the developing brain. *Journal of Biological Chemistry*, 276(32):30467–30474. (Cited on page 94.)

[Palla et al., 2022] Palla, G., Fischer, D. S., Regev, A., and Theis, F. J. (2022). Spatial components of molecular tissue biology. *Nature Biotechnology*, 40(3):308–318. (Cited on pages 13, 14, 16, 87, 97, 148, and 150.)

[Palla et al., 2021] Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., Rybakov, S., Ibarra, I. L., Holmberg, O., Virshup, I., et al. (2021). Squidpy: a scalable framework for spatial single cell analysis. *BioRxiv*, pages 2021–02. (Cited on pages 14, 90, and 149.)

[Papalexi and Satija, 2018] Papalexi, E. and Satija, R. (2018). Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35–45. (Cited on pages 33 and 66.)

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32. (Cited on pages 31, 80, 165, and 166.)

[Paty and Cuturi, 2019] Paty, F.-P. and Cuturi, M. (2019). Subspace robust wasserstein distances. In *International Conference on Machine Learning*, pages 5072–5081. PMLR. (Cited on page 53.)

[Paty and Cuturi, 2020] Paty, F.-P. and Cuturi, M. (2020). Regularized optimal transport is ground cost adversarial. In *International Conference on Machine Learning*, pages 7532–7542. PMLR. (Cited on page 53.)

[Paul et al., 2015] Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677. (Cited on pages 12 and 146.)

[PE de Souza et al., 2020] PE de Souza, C., Andronescu, M., Masud, T., Kabeer, F., Biele, J., Laks, E., Lai, D., Ye, P., Brimhall, J., Wang, B., et al. (2020). Epiclomal: probabilistic clustering of sparse single-cell dna methylation data. *PLOS Computational Biology*, 16(9):e1008270. (Cited on page 46.)

[Peco et al., 2012] Peco, E., Escude, T., Agius, E., Sabado, V., Medevielle, F., Ducommun, B., and Pituello, F. (2012). The cdc25b phosphatase shortens the g2 phase of neural progenitors and promotes efficient neuron production. *Development*, 139(6):1095–1104. (Cited on page 94.)

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830. (Cited on pages 39, 62, and 83.)

[Peng et al., 2024] Peng, Q., Zhou, P., and Li, T. (2024). stvcr: Reconstructing spatio-temporal dynamics of cell development using optimal transport. *bioRxiv*, pages 2024–06. (Cited on page 88.)

[Peyré, 2015] Peyré, G. (2015). Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351. (Cited on pages 26 and 160.)

[Peyré et al., 2019] Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607. (Cited on pages 6, 16, 17, 18, 24, 25, 33, 37, 52, 89, 141, 151, 152, 153, and 159.)

[Peyré et al., 2016] Peyré, G., Cuturi, M., and Solomon, J. (2016). Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. (Cited on pages 21, 25, 26, 102, 156, and 160.)

[Pham et al., 2023] Pham, D., Tan, X., Balderson, B., Xu, J., Grice, L. F., Yoon, S., Willis, E. F., Tran, M., Lam, P. Y., Raghubar, A., et al. (2023). Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nature communications*, 14(1):7739. (Cited on pages 15 and 150.)

[Picelli et al., 2013] Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098. (Cited on pages 8 and 142.)

[Picelli et al., 2014] Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181. (Cited on pages 8 and 143.)

[Pijuan-Sala et al., 2019] Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C., Ho, D. L. L., et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495. (Cited on pages 12 and 147.)

[Pirron et al., 1990] Pirron, U., Schlunck, T., Prinz, J. C., and Rieber, E. P. (1990). Ige-dependent antigen focusing by human b lymphocytes is mediated by the low-affinity receptor for ige. *European journal of immunology*, 20(7):1547–1551. (Cited on page 75.)

[Ponti et al., 2002] Ponti, C., Gibellini, D., Boin, F., Melloni, E., Manzoli, F. A., Cocco, L., Zauli, G., and Vitale, M. (2002). Role of creb transcription factor in c-fos activation in natural killer cells. *European journal of immunology*, 32(12):3358–3365. (Cited on page 77.)

[Potter, 2018] Potter, S. S. (2018). Single-cell rna sequencing for the study of development, physiology and disease. *Nature Reviews Nephrology*, 14(8):479–492. (Cited on pages 33 and 66.)

[Preissl et al., 2023] Preissl, S., Gaulton, K. J., and Ren, B. (2023). Characterizing cis-regulatory elements using single-cell epigenomics. *Nature Reviews Genetics*, 24(1):21–43. (Cited on pages 7, 8, 142, and 143.)

[Prince et al., 1992] Prince, H. E., York, J., and Jensen, E. R. (1992). Phenotypic comparison of the three populations of human lymphocytes defined by cd45ro and cd45ra expression. *Cellular immunology*, 145(2):254–262. (Cited on page 75.)

[Qian et al., 2016] Qian, W., Hong, B., Cai, D., He, X., Li, X., et al. (2016). Non-negative matrix factorization with sinkhorn distance. In *IJCAI*, pages 1960–1966. (Cited on page 69.)

[Qin et al., 2023] Qin, X., Rodriguez, F. C., Sufi, J., Vlckova, P., Claus, J., and Tape, C. J. (2023). A single-cell perturbation landscape of colonic stem cell polarisation. *bioRxiv*, pages 2023–02. (Cited on page 106.)

[Qiu, 2020] Qiu, P. (2020). Embracing the dropouts in single-cell rna-seq analysis. nat. commun. 11, 1169. (Cited on page 71.)

[Quake, 2021] Quake, S. R. (2021). The cell as a bag of rna. *Trends in Genetics*, 37(12):1064–1068. (Cited on pages 6 and 141.)

[Rai et al., 2020] Rai, V., Quang, D. X., Erdos, M. R., Cusanovich, D. A., Daza, R. M., Narisu, N., Zou, L. S., Didion, J. P., Guan, Y., Shendure, J., et al. (2020). Single-cell atac-seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. *Molecular metabolism*, 32:109–121. (Cited on pages 12 and 147.)

[Rajewsky et al., 2020] Rajewsky, N., Almouzni, G., Gorski, S. A., Aerts, S., Amit, I., Bertero, M. G., Bock, C., Bredenoord, A. L., Cavalli, G., Chiocca, S., et al. (2020). Lifetime and improving european healthcare through cell-based interceptive medicine. *Nature*, 587(7834):377–386. (Cited on pages 33 and 66.)

[Ramsköld et al., 2012] Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., et al. (2012). Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. *Nature biotechnology*, 30(8):777–782. (Cited on pages 8 and 142.)

[Regev et al., 2017] Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The human cell atlas. *elife*, 6:e27041. (Cited on pages 6, 12, 109, 141, and 147.)

[Righelli et al., 2022] Righelli, D., Weber, L. M., Crowell, H. L., Pardo, B., Collado-Torres, L., Ghazanfar, S., Lun, A. T., Hicks, S. C., and Risso, D. (2022). Spatialexperiment: infrastructure for spatially-resolved transcriptomics data in r using bioconductor. *Bioinformatics*, 38(11):3128–3131. (Cited on pages 14 and 149.)

[Roda-Navarro et al., 2000] Roda-Navarro, P., Arce, I., Renedo, M., Montgomery, K., Kucherlapati, R., and Fernández-Ruiz, E. (2000). Human klrf1, a novel member of the killer cell lectin-like receptor gene family: molecular characterization, genomic structure, physical mapping to the nk gene complex and expression analysis. *European Journal of Immunology*, 30(2):568–576. (Cited on page 77.)

[Rodriques et al., 2019] Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467. (Cited on pages 14 and 148.)

[Rolet et al., 2016] Rolet, A., Cuturi, M., and Peyré, G. (2016). Fast dictionary learning with a smoothed Wasserstein loss. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 630–638, Cadiz, Spain. PMLR, PMLR. (Cited on pages 25, 51, 69, 79, 80, and 160.)

[Rosenberg et al., 2018] Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182. (Cited on pages 8 and 142.)

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65. (Cited on page 39.)

[Rubner et al., 2000] Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121. (Cited on page 52.)

[Saelens et al., 2019] Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554. (Cited on pages 11, 87, and 146.)

[Samaran et al., 2024] Samaran, J., Peyre, G., and Cantini, L. (2024). scconfluence: single-cell diagonal integration with regularized inverse optimal transport on weakly connected features. *bioRxiv*, pages 2024–02. (Cited on pages 21, 28, 156, and 162.)

[Santambrogio, 2015] Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94. (Cited on pages 6, 17, 33, 52, 141, and 151.)

[Santambrogio, 2017] Santambrogio, F. (2017). {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154. (Cited on pages 23, 101, and 158.)

[Satija et al., 2015] Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502. (Cited on page 33.)

[Scetbon and Cuturi, 2022] Scetbon, M. and Cuturi, M. (2022). Low-rank optimal transport: Approximation, statistics and debiasing. *Advances in Neural Information Processing Systems*, 35:6802–6814. (Cited on page 110.)

[Schäfer and Frotscher, 2012] Schäfer, M. K. and Frotscher, M. (2012). Role of l1cam for axon sprouting and branching. *Cell and tissue research*, 349(1):39–48. (Cited on page 94.)

[Schiebinger et al., 2019] Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943. (Cited on pages 6, 12, 15, 17, 28, 33, 52, 63, 88, 104, 105, 141, 146, 147, 150, 151, 152, 162, and 163.)

[Schmitz et al., 2018] Schmitz, M. A., Heitz, M., Bonneel, N., Ngole, F., Coeurjolly, D., Cuturi, M., Peyré, G., and Starck, J.-L. (2018). Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678. (Cited on pages 69 and 110.)

[Schölkopf and Smola, 2002] Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press. (Cited on page 61.)

[Schroeder Jr and Cavacini, 2010] Schroeder Jr, H. W. and Cavacini, L. (2010). Structure and function of immunoglobulins. *Journal of allergy and clinical immunology*, 125(2):S41–S52. (Cited on page 75.)

[Schwartzman and Tanay, 2015] Schwartzman, O. and Tanay, A. (2015). Single-cell epigenomics: techniques and emerging applications. *Nature Reviews Genetics*, 16(12):716–726. (Cited on pages 8 and 143.)

[Séjourné et al., 2023] Séjourné, T., Peyré, G., and Vialard, F.-X. (2023). Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471. (Cited on page 110.)

[Setty et al., 2019] Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'Er, D. (2019). Characterization of cell fate probabilities in single-cell data with palantir. *Nature biotechnology*, 37(4):451–460. (Cited on pages 11 and 146.)

[Shah et al., 2021] Shah, K., Al-Haidari, A., Sun, J., and Kazi, J. U. (2021). T cell receptor (tcr) signaling in health and disease. *Signal transduction and targeted therapy*, 6(1):412. (Cited on page 75.)

[Shen et al., 2023] Shen, X., Huang, K., Zuo, L., Ye, Z., Li, Z., Yu, Q., Zou, X., Wei, X., Xu, P., Jin, X., et al. (2023). Inferring cell trajectories of spatial transcriptomics via optimal transport analysis. *bioRxiv*, pages 2023–09. (Cited on pages 15 and 150.)

[Sikkema et al., 2023] Sikkema, L., Ramírez-Suástegui, C., Strobl, D. C., Gillett, T. E., Zappia, L., Madissoon, E., Markov, N. S., Zaragosi, L.-E., Ji, Y., Ansari, M., et al. (2023). An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29(6):1563–1577. (Cited on pages 12 and 147.)

[Singh et al., 2021] Singh, R., Hie, B. L., Narayan, A., and Berger, B. (2021). Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities. *Genome biology*, 22(1):131. (Cited on pages 67 and 69.)

[Sinkhorn, 1964] Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879. (Cited on pages 25 and 159.)

[Smallwood et al., 2014] Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817–820. (Cited on pages 8, 12, 143, and 146.)

[Solomon et al., 2016] Solomon, J., Peyré, G., Kim, V. G., and Sra, S. (2016). Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13. (Cited on pages 21 and 156.)

[Ståhl et al., 2016] Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82. (Cited on pages 14, 97, and 148.)

[Stanojevic et al., 2022] Stanojevic, S., Li, Y., Ristivojevic, A., and Garmire, L. X. (2022). Computational methods for single-cell multi-omics integration and alignment. *Genomics, Proteomics and Bioinformatics*, 20(5):836–849. (Cited on pages 15, 67, 68, and 149.)

[Stegle et al., 2015] Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145. (Cited on pages 33 and 61.)

[Stein-O'Brien et al., 2018] Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., Goff, L. A., Li, Y., Ngom, A., Ochs, M. F., et al. (2018). Enter the matrix: factorization uncovers knowledge from omics. *Trends in Genetics*, 34(10):790–805. (Cited on pages 15, 68, and 149.)

[Stoeckius et al., 2017] Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868. (Cited on pages 13, 14, 66, 67, 147, and 148.)

[Street et al., 2018] Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19:1–16. (Cited on pages 11, 89, and 146.)

[Stuart and Wolfram, 2020] Stuart, A. M. and Wolfram, M.-T. (2020). Inverse optimal transport. *SIAM Journal on Applied Mathematics*, 80(1):599–619. (Cited on page 53.)

[Stuart et al., 2019] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902. (Cited on pages 61, 67, and 71.)

[Stuart and Satija, 2019] Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. *Nature reviews genetics*, 20(5):257–272. (Cited on page 85.)

[Stuart et al., 2021] Stuart, T., Srivastava, A., Madad, S., Lareau, C. A., and Satija, R. (2021). Single-cell chromatin state analysis with signac. *Nature methods*, 18(11):1333–1341. (Cited on page 83.)

[Sturm, 2012] Sturm, K.-T. (2012). The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*. (Cited on pages 21 and 156.)

[Su et al., 1994] Su, B., Bochan, M. R., Hanna, W. L., Froelich, C. J., and Brahmi, Z. (1994). Human granzyme b is essential for dna fragmentation of susceptible target cells. *European journal of immunology*, 24(9):2073–2080. (Cited on page 77.)

[Sun et al., 2020] Sun, H., Zhou, H., Zha, H., and Ye, X. (2020). Learning cost functions for optimal transport. *arXiv preprint arXiv:2002.09650*. (Cited on pages 21 and 155.)

[Svensson et al., 2020] Svensson, V., Gayoso, A., Yosef, N., and Pachter, L. (2020). Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421. (Cited on pages 11, 67, and 145.)

[Svensson et al., 2017] Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., and Teichmann, S. A. (2017). Power analysis of single-cell rna-sequencing experiments. *Nature methods*, 14(4):381–387. (Cited on pages 8 and 142.)

[Svensson et al., 2018] Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell rna-seq in the past decade. 13(4):599–604. (Cited on pages 8, 109, and 142.)

[Swanson et al., 2021] Swanson, E., Lord, C., Reading, J., Heubeck, A. T., Genge, P. C., Thomson, Z., Weiss, M. D., Li, X.-j., Savage, A. K., Green, R. R., et al. (2021). Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using tea-seq. *Elife*, 10:e63632. (Cited on pages 13, 66, 67, 85, and 148.)

[Tanay and Regev, 2017] Tanay, A. and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637):331–338. (Cited on page 61.)

[Tarjan, 1997] Tarjan, R. E. (1997). Dynamic trees as search trees via euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177. (Cited on pages 25 and 159.)

[Taylor et al., 2007] Taylor, M. K., Yeager, K., and Morrison, S. J. (2007). Physiological notch signaling promotes gliogenesis in the developing peripheral and central nervous systems. *Development*, 134(13):2435–2447. (Cited on page 97.)

[Tenenbaum et al., 2000] Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323. (Cited on pages 10 and 145.)

[Terpin et al., 2024] Terpin, A., Lanzetti, N., and Dörfler, F. (2024). Learning diffusion at lightspeed. *arXiv preprint arXiv:2406.12616*. (Cited on pages 27, 97, 110, and 161.)

[Theodoris et al., 2023] Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624. (Cited on page 109.)

[Tian et al., 2023] Tian, W., Zhou, J., Bartlett, A., Zeng, Q., Liu, H., Castanon, R. G., Kenworthy, M., Altshul, J., Valadon, C., Aldridge, A., et al. (2023). Single-cell dna methylation and 3d genome architecture in the human brain. *Science*, 382(6667):eadf5357. (Cited on pages 12 and 147.)

[Tipping and Bishop, 1999] Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622. (Cited on pages 10 and 144.)

[Tirosh et al., 2016a] Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016a). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196. (Cited on page 95.)

[Tirosh et al., 2016b] Tirosh, I., Venteicher, A. S., Hebert, C., Escalante, L. E., Patel, A. P., Yizhak, K., Fisher, J. M., Rodman, C., Mount, C., Filbin, M. G., et al. (2016b). Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, 539(7628):309–313. (Cited on page 230.)

[Tolchin et al., 2020] Tolchin, D., Yeager, J. P., Prasad, P., Dorrani, N., Russi, A. S., Martinez-Agosto, J. A., Haseeb, A., Angelozzi, M., Santen, G., Ruivenkamp, C., et al. (2020). De novo sox6 variants cause a neurodevelopmental syndrome associated with adhd, craniosynostosis, and osteochondromas. *The American Journal of Human Genetics*, 106(6):830–845. (Cited on page 95.)

[Tolstikhin et al., 2017] Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2017). Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*. (Cited on page 52.)

[Tong et al., 2020] Tong, A., Huang, J., Wolf, G., Van Dijk, D., and Krishnaswamy, S. (2020). Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pages 9526–9536. PMLR. (Cited on page 88.)

[Tong et al., 2022] Tong, A., Huguet, G., Shung, D., Natik, A., Kuchroo, M., Lajoie, G., Wolf, G., and Krishnaswamy, S. (2022). Embedding signals on graphs with unbalanced diffusion earth mover's distance. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5647–5651. IEEE. (Cited on pages 27, 52, 109, and 162.)

[Townes et al., 2019] Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20:1–16. (Cited on pages 10 and 144.)

[Traag et al., 2019] Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233. (Cited on pages 11, 39, 68, 83, and 145.)

[Trapnell, 2015] Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome research*, 25(10):1491–1498. (Cited on pages 6, 7, 11, 13, 141, 142, 146, and 148.)

[Trapnell et al., 2014] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386. (Cited on pages 11, 12, 87, 89, 146, and 147.)

[Tritschler et al., 2019] Tritschler, S., Büttner, M., Fischer, D. S., Lange, M., Bergen, V., Lickert, H., and Theis, F. J. (2019). Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, 146(12). (Cited on page 88.)

[Turing, 1990] Turing, A. M. (1990). The chemical basis of morphogenesis. *Bulletin of mathematical biology*, 52:153–197. (Cited on page 98.)

[Van der Maaten and Hinton, 2008] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11). (Cited on pages 10, 68, and 145.)

[Van Dijk et al., 2018] Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729. (Cited on page 106.)

[Vandereyken et al., 2023] Vandereyken, K., Sifrim, A., Thienpont, B., and Voet, T. (2023). Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, 24(8):494–515. (Cited on pages 6, 13, 109, 141, 147, and 148.)

[Vayer et al., 2018a] Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2018a). Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties. *arXiv preprint arXiv:1811.02834*. (Cited on page 102.)

[Vayer et al., 2018b] Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2018b). Optimal transport for structured data with application on graphs. *arXiv preprint arXiv:1805.09114*. (Cited on pages 22, 25, 88, 90, 157, and 160.)

[Velmeshev et al., 2019] Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H., and Kriegstein, A. R. (2019). Single-cell genomics identifies cell type–specific molecular changes in autism. *Science*, 364(6441):685–689. (Cited on pages 12 and 146.)

[Velten et al., 2022] Velten, B., Braunger, J. M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., Zeller, G., and Stegle, O. (2022). Identifying temporal and spatial patterns of variation from multimodal data using mefisto. *Nature methods*, 19(2):179–186. (Cited on pages 15, 111, and 149.)

[Vickovic et al., 2019] Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J. F., et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. *Nature methods*, 16(10):987–990. (Cited on pages 14, 87, and 148.)

[Villani, 2003] Villani, C. (2003). *Topics in C. Transportation*. Graduate studies in Math. AMS. (Cited on pages 17, 52, and 151.)

[Villani et al., 2009] Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer. (Cited on pages 6, 89, and 141.)

[Virshup et al., 2023] Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Berger, B., et al. (2023). The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature biotechnology*, 41(5):604–606. (Cited on pages 31, 88, and 166.)

[Virshup et al., 2021] Virshup, I., Rybakov, S., Theis, F. J., Angerer, P., and Wolf, F. A. (2021). anndata: Annotated data. *BioRxiv*, pages 2021–12. (Cited on pages 9, 31, 143, and 165.)

[Virtanen et al., 2020] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272. (Cited on page 106.)

[Vivier et al., 2008] Vivier, E., Tomasello, E., Baratin, M., Walzer, T., and Ugolini, S. (2008). Functions of natural killer cells. *Nature immunology*, 9(5):503–510. (Cited on page 77.)

[Von Luxburg, 2007] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17:395–416. (Cited on page 39.)

[Wang and Guibas, 2012] Wang, F. and Guibas, L. J. (2012). Supervised earth mover's distance learning and its computer vision applications. In *European Conference on Computer Vision*, pages 442–455. Springer. (Cited on pages 20, 52, and 155.)

[Wang et al., 2020] Wang, X., Sun, Z., Zhang, Y., Xu, Z., Xin, H., Huang, H., Duerr, R. H., Chen, K., Ding, Y., and Chen, W. (2020). Brem-sc: a bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic acids research*, 48(11):5814–5824. (Cited on pages 67 and 69.)

[Wang and Navin, 2015] Wang, Y. and Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Molecular cell*, 58(4):598–609. (Cited on pages 8 and 143.)

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63. (Cited on pages 7 and 142.)

[Weber et al., 2007] Weber, C., Fraemohs, L., and Dejana, E. (2007). The role of junctional adhesion molecules in vascular inflammation. *Nature Reviews Immunology*, 7(6):467–477. (Cited on page 76.)

[Wei et al., 2022] Wei, X., Fu, S., Li, H., Liu, Y., Wang, S., Feng, W., Yang, Y., Liu, X., Zeng, Y.-Y., Cheng, M., et al. (2022). Single-cell stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science*, 377(6610):eabp9444. (Cited on pages 16, 87, 88, 91, 92, 94, 99, 100, 151, and 231.)

[Weiler et al., 2023] Weiler, P., Lange, M., Klein, M., Pe'er, D., and Theis, F. (2023). Unified fate mapping in multiview single-cell data. (Cited on pages 12, 15, 31, 88, 90, 106, 146, 150, and 166.)

[Weinberger et al., 2006] Weinberger, K. Q., Blitzer, J., and Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480. (Cited on page 52.)

[Weinreb et al., 2018] Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M., and Klein, A. M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476. (Cited on pages 97 and 110.)

[Welch et al., 2017] Welch, J. D., Hartemink, A. J., and Prins, J. F. (2017). Matcher: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome biology*, 18:1–19. (Cited on pages 15 and 150.)

[Welch et al., 2019] Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887. (Cited on pages 67 and 69.)

[Westdickenberg and Wilkening, 2010] Westdickenberg, M. and Wilkening, J. (2010). Variational particle schemes for the porous medium equation and for the system of isentropic euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(1):133–166. (Cited on pages 26 and 161.)

[Wolf et al., 2018] Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5. (Cited on pages 9, 31, 39, 61, 62, 69, 80, 82, 88, 90, 109, 143, 165, and 184.)

[Xia et al., 2019] Xia, C., Fan, J., Emanuel, G., Hao, J., and Zhuang, X. (2019). Spatial transcriptome profiling by merfish reveals subcellular rna compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, 116(39):19490–19499. (Cited on pages 14 and 148.)

[Xing et al., 2002] Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. (2002). Distance metric learning with application to clustering with side-information. In *NIPS*, volume 15, page 12. (Cited on page 52.)

[Xiong et al., 2019] Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., Zhang, M., Jiang, T., and Zhang, Q. C. (2019). Scale method for single-cell atac-seq analysis via latent feature extraction. *Nature communications*, 10(1):4576. (Cited on pages 11, 33, 46, and 145.)

[Xu et al., 2018] Xu, J., Luo, L., Deng, C., and Huang, H. (2018). Multi-level metric learning via smoothed wasserstein distance. In *IJCAI*, pages 2919–2925. (Cited on page 52.)

[Yang et al., 2020] Yang, K. D., Damodaran, K., Venkatachalapathy, S., Soylemezoglu, A. C., Shivashankar, G., and Uhler, C. (2020). Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828. (Cited on page 37.)

[Yang et al., 2022] Yang, S.-G., Wang, X.-W., Qian, C., and Zhou, F.-Q. (2022). Reprogramming neurons for regeneration: The fountain of youth. *Progress in neurobiology*, 214:102284. (Cited on page 95.)

[Yang et al., 2015] Yang, W., Xu, L., Chen, X., Zheng, F., and Liu, Y. (2015). Chi-squared distance metric learning for histogram data. *Mathematical Problems in Engineering*, 2015. (Cited on page 52.)

[Yeo et al., 2021] Yeo, G. H. T., Saksena, S. D., and Gifford, D. K. (2021). Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nature communications*, 12(1):3222. (Cited on pages 88, 89, 90, 92, 101, 102, 103, 104, and 105.)

[Yip et al., 2017] Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C., and Wang, J. (2017). Linnorm: improved statistical analysis for single cell rna-seq expression data. *Nucleic acids research*, 45(22):e179–e179. (Cited on page 35.)

[Yuan et al., 2017] Yuan, G.-C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., et al. (2017). Challenges and emerging directions in single-cell analysis. *Genome biology*, 18(1):1–8. (Cited on page 61.)

[Yurochkin et al., 2019] Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., and Solomon, J. M. (2019). Hierarchical optimal transport for document representation. *Advances in neural information processing systems*, 32. (Cited on pages 52, 53, and 111.)

[Zappia et al., 2017] Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174. (Cited on page 34.)

[Zappia and Theis, 2021] Zappia, L. and Theis, F. J. (2021). Over 1000 tools reveal trends in the single-cell rna-seq analysis landscape. *Genome biology*, 22:1–18. (Cited on pages 9 and 143.)

[Zeira et al., 2022] Zeira, R., Land, M., Strzalkowski, A., and Raphael, B. J. (2022). Alignment and integration of spatial transcriptomics data. *Nature Methods*, 19(5):567–575. (Cited on pages 22 and 157.)

[Zen et al., 2014] Zen, G., Ricci, E., and Sebe, N. (2014). Simultaneous ground metric learning and matrix factorization with earth mover's distance. In *2014 22nd International Conference on Pattern Recognition*, pages 3690–3695. IEEE. (Cited on page 53.)

[Zhang et al., 2021] Zhang, J., Le Gras, S., Pouxvielh, K., Faure, F., Fallone, L., Kern, N., Moreews, M., Mathieu, A.-L., Schneider, R., Marliac, Q., et al. (2021). Sequential actions of eomes and t-bet promote stepwise maturation of natural killer cells. *Nature communications*, 12(1):5446. (Cited on page 77.)

[Zhang, 2021] Zhang, S. Y. (2021). A unified framework for non-negative matrix and tensor factorisations with a smoothed wasserstein loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4203. (Cited on page 69.)

[Zhang et al., 2023] Zhang, Y., Zhao, Q., Chen, Q., Xu, L., and Yi, S. (2023). Transcriptional control of peripheral nerve regeneration. *Molecular Neurobiology*, 60(1):329–341. (Cited on page 95.)

[Zheng et al., 2019] Zheng, R., Li, M., Liang, Z., Wu, F.-X., Pan, Y., and Wang, J. (2019). Sinnlrr: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. *Bioinformatics*, 35(19):3642–3650. (Cited on page 39.)

[Zhou et al., 2023] Zhou, R., Yang, G., Zhang, Y., and Wang, Y. (2023). Spatial transcriptomics in development and disease. *Molecular Biomedicine*, 4(1). (Cited on page 97.)

[Zhu et al., 2010] Zhu, J., Zhang, Y.-T., Alber, M. S., and Newman, S. A. (2010). Bare bones pattern formation: a core regulatory network in varying geometries reproduces major features of vertebrate limb development and evolution. *PLoS One*, 5(5):e10892. (Cited on page 98.)

[Ziegenhain et al., 2017] Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643. (Cited on pages 8, 142, and 143.)

[Zuo and Chen, 2021] Zuo, C. and Chen, L. (2021). Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Briefings in Bioinformatics*, 22(4):bbaa287. (Cited on pages 67 and 69.)

Introduction (en français)

Au milieu du XVIIe siècle, Robert Hooke a fait une découverte révolutionnaire en observant un morceau de liège au microscope. Il a observé des structures qu'il a nommées "cellules" et a ainsi marqué le début de la biologie cellulaire. [Hooke et al., 1665] Les cellules ont depuis été identifiées comme les unités structurelles et fonctionnelles fondamentales de la vie, et les biologistes se sont efforcés de cartographier les divers types de cellules qui composent les organismes multicellulaires. En outre, ils ont cherché à comprendre les états cellulaires transitoires pendant le développement, la progression de la maladie et la régénération des tissus [Trapnell, 2015].

Les efforts récents visant à cartographier le paysage de l'identité cellulaire, comme le projet Human Cell Atlas [Regev et al., 2017], sont allés au-delà de la microscopie pour quantifier les états cellulaires au niveau moléculaire. Plus précisément, les technologies de séquençage d'ARN à haut débit à cellule unique (scRNA-seq) ont permis de caractériser l'identité cellulaire grâce à l'expression génique. Compte tenu du succès de ces méthodes, la cellule a même été comparée à un « sac d'ARN ». [Quake, 2021].

Cependant, l'identité cellulaire est influencée par une interaction complexe de diverses couches moléculaires, notamment la protéomique, la transcriptomique et l'épigénomique. En conséquence, deux nouvelles frontières technologiques cruciales sont apparues pour comprendre la cellule de manière plus complète. La première est le séquençage unicellulaire multiomique à haut débit, qui mesure simultanément plusieurs omiques au sein d'une même cellule. La seconde est la transcriptomique spatiale, qui quantifie l'expression des gènes tout en préservant le contexte spatial des cellules. Ensemble, ces avancées relèvent du séquençage multimodal unicellulaire conjoint. [Vandereyken et al., 2023]

L'analyse conjointe de plusieurs modalités de données multimodales unicellulaires appariées, également appelée *intégration multimodale unicellulaire conjointe*, est difficile et nécessite des méthodes de calcul adaptées. En effet, les données unicellulaires sont rares, bruitées et de grande dimension, et les modalités complémentaires ont souvent des propriétés très différentes [Vandereyken et al., 2023]. Cette thèse se concentre sur le développement de méthodes de calcul répondant à ces défis en utilisant le cadre mathématique du transport optimal (OT) [Villani et al., 2009, Santambrogio, 2015]. L'OT est un moyen prometteur de comparer et d'étudier des objets de grande dimension d'une manière géométriquement pertinente, et de récentes avancées informatiques ont fait de l'OT un outil essentiel dans les applications modernes de l'apprentissage automatique [Peyré et al., 2019]. En particulier, l'OT a suscité un intérêt considérable pour l'analyse unicellulaire [Hashimoto et al., 2016, Schiebinger et al., 2019, Demetci et al., 2022b], justifiant les contributions dans cette thèse.

1.1 Analyse de données de séquençage de cellules uniques à haut débit

Le séquençage omique à haut débit de cellules uniques fait référence à un ensemble de technologies capables de quantifier l'identité moléculaire des cellules uniques. Ces technologies analysent diverses couches moléculaires, notamment l'épigénomique et la transcriptomique, au sein de chaque cellule. Elles ont permis la caractérisation précise des types de cellules et des trajectoires de différenciation, améliorant ainsi notre compréhension de la diversité et de la dynamique des cellules dans les systèmes biologiques complexes. [Trapnell, 2015].

Ces dernières années, de nombreuses technologies de séquençage de cellules uniques ont été développées, ainsi que des méthodes informatiques pour analyser le volume toujours croissant de données produites dans les expériences sur cellules uniques. Dans cette section, je donne un aperçu du domaine du séquençage de cellules uniques, y compris les technologies qui ont été développées (Section 1.1.1), les méthodes informatiques qui ont été développées pour l'analyse des données de cellules uniques (Section 1.1.2) et les applications du séquençage de cellules uniques (Section 1.1.3).

1.1.1 Technologies de séquençage de cellules uniques

Les technologies de séquençage de cellules uniques s'appuient sur des plateformes de séquençage *à haut débit* telles qu'Illumina, qui permettent le séquençage parallèle de grandes quantités de brins d'ADN [Goodwin et al., 2016]. Des protocoles tels que scRNA-seq et scATAC-seq utilisent cette technologie pour séquencer des bibliothèques d'ADN constituées à partir des informations moléculaires contenues dans des cellules uniques, permettant une quantification moléculaire précise à la résolution de la cellule [Nawy, 2014, Preissl et al., 2023].

Les protocoles de séquençage de cellules uniques contrastent avec les protocoles de séquençage *bulk* tels que RNA-seq et ATAC-seq, qui construisent des bibliothèques à partir d'un ensemble de cellules [Wang et al., 2009, Buenrostro et al., 2013]. Comme ils moyennent des signaux provenant de groupes hétérogènes de cellules, les analyses *bulk* souffrent du paradoxe de Simpson, qui est un phénomène statistique où les tendances qui apparaissent dans différents groupes de données disparaissent ou s'inversent lorsque ces groupes sont combinés [Trapnell, 2015].

Dans cette section, je couvrirai plus en détail les trois modalités de séquençage de cellules uniques qui sont pertinentes pour le reste de la thèse, à savoir la transcriptomique, l'accessibilité de la chromatine et la méthylation.

Transcriptomique Le *single-cell RNA sequencing* (scRNA-seq) est une technologie à haut débit qui permet de mesurer l'expression des gènes à la résolution d'une cellule unique [Ziegenhain et al., 2017, Svensson et al., 2017, Mereu et al., 2020] dont le nombre de cellules par expérience croit de manière exponentielle [Svensson et al., 2018, Kharchenko, 2021]. Les protocoles de scRNA-seq commencent par dissocier le tissu en une suspension de cellules uniques, qui sont ensuite physiquement isolées les unes des autres à l'aide de technologies "plate-based" [Islam et al., 2011, Ramsköld et al., 2012, Picelli et al., 2013] ou "droplet-based" [Macosko et al., 2015, Klein et al., 2015]. Les protocoles récents introduisent un code-barres moléculaire propre à chaque cellule, qui se fixe aux ARNm et permet de regrouper les cellules avant le séquençage [Klein et al., 2015]. Dans le cas du code-barres combinatoire [Cao et al., 2017, Rosenberg et al., 2018], les cellules n'ont même pas besoin d'être physiquement isolées. Comme dans le bulk RNA-seq, les ARNm capturés sont rétrotranscrits en molécules d'ADNc par la transcriptase inverse [Svensson et al., 2018]. Les ADNc sont ensuite amplifiés en raison de la faible quantité de matériel génétique dans une seule cellule : $\sim 0,1$ picogramme d'ARNm,

comparé au microgramme requis par le RNA-seq [Huang et al., 2018, Wang and Navin, 2015]. Pour garder une trace de la quantité absolue d'ARNm avant l'amplification, des protocoles récents marquent chaque ARNm avec un identifiant moléculaire unique (UMI) [Ziegenhain et al., 2017]. L'ADNc est ensuite séquencé, en lisant soit l'extrémité 3' du transcrit (par exemple [Macosko et al., 2015]) soit le transcrit complet [Picelli et al., 2014]. Le séquençage génère des "reads" qui subissent un pipeline de traitement comme CellRanger, qui génère une *matrice de comptage* $\mathbf{A} \in \mathbb{N}^{m \times n}$ où m est le nombre d'espèces d'ARNm différentes (il existe environ $m = 20000$ gènes codant pour des protéines dans le génome humain) et n est le nombre de codes-barres uniques.

Accessibilité de la chromatine Le *single-cell assay for transposase-accessible chromatin sequencing* (scATAC-seq) est une technologie à haut débit qui permet de mesurer l'accessibilité de la chromatine à une résolution cellulaire en mesurant les sites d'hypersensibilité à la transposase Tn5 [Preissl et al., 2023]. Ces informations sont cruciales pour comprendre le paysage régulateur de la cellule, car l'architecture de la chromatine façonne l'état transcriptionnel de la cellule [Klemm et al., 2019]. Les protocoles scATAC-seq s'appuient sur l'ATAC-seq bulk [Buenrostro et al., 2013] et tirent parti de l'amplification pour atteindre une résolution cellulaire [De Rop et al., 2023]. Comme dans la transcriptomique de cellule unique, les avancées technologiques telles que l'isolement en gouttelettes et l'indexation combinatoire [Lareau et al., 2019, Cusanovich et al., 2015] ont permis une mise à l'échelle vers un grand nombre de cellules. Représenter les données scATAC-seq comme une matrice de comptage est moins simple qu'avec le scRNA-seq, car il n'existe pas de features universelles pour mapper les reads [Heumos et al., 2023]. Plutôt que de manipuler directement les reads, il est courant de diviser le génome en *bins* de taille uniforme (par exemple de 5 kbp) ou en *peaks* de longueur variable et de compter les événements de transposition Tn5 dans ces régions [Fang et al., 2021].

Méthylation Le *single-cell DNA methylation sequencing* est une technologie à haut débit qui permet de mesurer la méthylation de l'ADN à une résolution ceullulaire. La forme la plus courante de méthylation de l'ADN est l'ajout d'un groupe méthyle au 5ème carbone d'un nucléotide cytosine (noté 5mC). Cette modification épigénétique se produit généralement lorsque la cytosine est suivie d'une guanine (un dinucléotide CpG) et est souvent associée à une répression de l'expression des gènes. Elle offre ainsi des informations précieuses sur les mécanismes de régulation sous-jacents à l'identité cellulaire [Schwartzman and Tanay, 2015, Preissl et al., 2023]. La méthylation de l'ADN en cellule unique est généralement mesurée à l'aide de techniques de séquençage du génome entier au bisulfite, qui convertissent les cytosines non méthylées en uracile tout en laissant les cytosines méthylées inchangées [Preissl et al., 2023]. Parmi les exemples de telles technologies, on peut citer scBS-seq [Smallwood et al., 2014], snmC-seq [Luo et al., 2017] et snmC-seq2 [Luo et al., 2018]. Comme pour scATAC-seq, les informations de séquençage sont généralement regroupées dans des bins, et on compte le nombre de CpG méthylés dans chaque bin pour produire une matrice de cellule par bins [Luo et al., 2017].

1.1.2 Analyse computationnelle des données de séquençage de cellules uniques

À mesure que les technologies de séquençage de cellules uniques se généralisent, de nombreuses méthodes ont été développées pour analyser le volume toujours croissant de données issues d'expériences sur des cellules uniques. Par exemple, [Zappia and Theis, 2021] compte plus de 1 000 outils pour l'analyse scRNA-seq, couvrant une diversité de tâches. Heureusement, des normes ont émergé pour assurer l'intercompatibilité de ces méthodes, comme Seurat et Bioconductor en R [Butler et al., 2018, Amezquita et al., 2020] ou Scanpy et AnnData en Python [Wolf et al., 2018, Virshup et al., 2021].

Ci-dessous, je décris les principales tâches computationelles dans l'analyse de cellules uniques abordées dans cette thèse. Un compte rendu plus complet des méthodes d'analyse de cellules uniques peut être trouvé dans [Luecken and Theis, 2019, Heumos et al., 2023].

Prétraitement

Les pipelines d'analyse à cellule unique soumettent généralement les matrices de comptage $\mathbf{A} \in \mathbb{N}^{m \times n}$ (m feautres et n cellules) décrites ci-dessus à des étapes de prétraitement telles que le *contrôle qualité* et la *normalisation* pour tenir compte des limitations des protocoles de séquençage [Luecken and Theis, 2019, Heumos et al., 2023].

Le contrôle qualité consiste à filtrer les cellules non informatives telles que les doublets ou les cellules mortes. Dans le scRNA-seq, le contrôle qualité est généralement basé sur *(i)* le nombre total de transcrits capturés dans une cellule (le "count depth") *(ii)* le nombre de gènes par code-barres et *(iii)* la fraction de comptages provenant de gènes mitochondriaux. [Luecken and Theis, 2019].
En scATAC-seq, le contrôle qualité est généralement basé sur *(i)* le nombre total de fragments par cellule *(ii)* le log du nombre total de fragments et *(iii)* le score d'enrichissement du site de démarrage de la transcription [Heumos et al., 2023]. Dans le séquençage de méthylation en cellule unique, le contrôle qualité est généralement effectué au niveau de la séquence [Iqbal and Zhou, 2023].

Ensuite, les étapes de normalisation visent à réduire le bruit technique et à faciliter l'analyse en aval. Par exemple, dans le scRNA-seq, la capture d'ARNm est un processus stochastique et les différences de count depth sont donc peu informatives en ce qui concerne l'hétérogénéité cellulaire [Luecken and Theis, 2019]. De nombreuses stratégies de normalisation ont été proposées, notamment les résidus de Pearson [Lause et al., 2021], mais une approche simple pourtant largement utilisée consiste simplement à diviser le nombre de cellules par sa count depth, suivie d'une transformation non linéaire telle que $x \mapsto \log(1 + x)$ [Ahlmann-Eltze and Huber, 2023]. Dans l'ATAC-seq, les pics sont souvent binarisés [Fang et al., 2021] pour représenter la chromatine ouverte ou fermée, bien que des travaux récents aient montré un avantage à la modélisation du nombre de fragments [Martens et al., 2024].

La transformation logarithmique avec un offset préserve la sparsité des données mais réduit leur asymétrie, rendant les valeurs normalisées plus proches d'une distribution normale [Luecken and Theis, 2019]. L'analyse en composantes principales (ACP), qui est souvent l'étape suivante de l'analyse à cellule unique, est bien adaptée aux données distribuées normalement. En effet, [Tipping and Bishop, 1999] montre que l'ACP résulte de l'estimation du maximum de vraisemblance des paramètres d'un modèle de variable latente gaussien. Cependant, [Townes et al., 2019] soutiennent que les données de comptage sprase et asymétriques sont mieux modélisées avec des distributions de Poisson ou binomiales négatives, et proposent donc d'utiliser une généralisation de l'ACP probabiliste [Collins et al., 2001]. L'ACP généralisée peut être approximée avec l'ACP habituelle sur les résidus de Pearson (qui sont denses et approximativement distribués normalement) [Lause et al., 2021, Townes et al., 2019].

Réduction de la dimensionnalité

Les données à cellule unique sont généralement en très haute dimension. Par conséquent, de nombreuses méthodes de réduction de dimensionnalité ont été développées pour faciliter la visualisation et les applications en aval, telles que le clustering et l'inférence de trajectoire [Heumos et al., 2023].

Avant la réduction de dimensionnalité, les features sont souvent sous-ensembleées à celles qui varient le plus dans le jeu de données [Luecken and Theis, 2019, Heumos et al., 2023].

La visualisation des données en cellule unique repose pour la grande majorité sur t-SNE [Van der Maaten and Hinton, 2008] ou UMAP [McInnes et al., 2018]. Bien que ces méthodes soient efficaces pour visualiser des clusters discrets de cellules [Kobak and Berens, 2019], d'autres méthodes, telles que PHATE [Moon et al., 2019] et Isomap [Tenenbaum et al., 2000], sont souvent préférées pour visualiser des trajectoires continues comme le développement.

Les applications en aval comme le clustering et l'inférence de trajectoire bénéficient également grandement de la réduction de la dimensionnalité. L'approche la plus utilisée est l'ACP, bien que de nombreux autres outils aient été développés [Heumos et al., 2023]. Les approches de factorisation en matrices non négatives (NMF) présentent un intérêt particulier pour cette thèse, car elles offrent une représentation intuitive des données basée sur un "somme de parties" [Lee and Seung, 1999]. Les intégrations basées sur des réseaux de neurones ont attiré l'attention pour leur pouvoir expressif dans la modélisation de données à cellule unique [Gayoso et al., 2022]. Cependant, malgré des efforts récents pour améliorer leur interprétabilité [Ainsworth et al., 2018, Svensson et al., 2020], les réseaux de neurones manquent souvent d'interprétabilité biologique, par rapport à des modèles linéaires comme le NMF.

Clustering non supervisé

Un objectif fréquent dans l'analyse des données à cellule unique est l'identification des types de cellules et des états cellulaires dans un échantillon [Luecken and Theis, 2019]. Cela est généralement réalisé par le biais d'un clustering non supervisé, où les cellules ayant des profils similaires sont regroupées en clusters, qui sont ensuite annotés en fonction de marqueurs connus [Kiselev et al., 2019, Xiong et al., 2019]. Par conséquent, la qualité du clustering est essentielle pour une découverte biologique précise.

Au début de l'analyse de données en cellule unique, les méthodes utilisées dans les analyses de bulk étaient appliquées. Par exemple, MARS-seq [Jaitin et al., 2014] a utilisé le clustering hiérarchique pour détecter des populations de cellules immunitaires dans la rate. Cependant, en raison de l'augmentation de la taille des jeux de données et de l'amélioration des performances, la plupart des pipelines d'analyse en cellule unique modernes s'appuient sur des algorithmes de clustering graphique (ou de détection de communauté) comme Leiden [Traag et al., 2019] ou Louvain [Blondel et al., 2008].

Le succès du clustering non supervisé dans la révélation de l'hétérogénéité cellulaire dans les données en cellule unique de haute dimension semble défier la malédiction de la dimensionnalité, qui suggère que dans les dimensions élevées, les points de données deviennent uniformément distribués [Beyer et al., 1999]. Mais malgré la dimensionnalité apparemment élevée, les données en cellule uniques s'organisent souvent en variétés de faible dimension qui peuvent être capturées par réduction de dimensionnalité [Luecken and Theis, 2019, Moon et al., 2018].

Un défi général avec l'annotation des clusters est également le phénomène de double-dipping : lors de l'utilisation de tests statistiques pour l'expression différentielle des gènes marqueurs, il faut tenir compte du fait que les deux populations ne sont pas indépendantes puisqu'elles proviennent du clustering. [Lähnemann et al., 2020, Gao et al., 2024, Hivert et al., 2024]

Inférence de trajectoire

L'inférence de trajectoire diffère du clustering en ce qu'elle recherche une caractérisation continue de l'identité cellulaire le long des destins cellulaires, plutôt qu'une caractérisation discrète. [Saelens et al., 2019, Trapnell, 2015]

Méthodes basées sur le pseudotime. Monocle [Trapnell et al., 2014] a initié ce domaine de recherche actif en introduisant le concept de pseudotime, qui ordonne les cellules le long d'un processus biologique, tel que la différenciation cellulaire ou la réponse à un stimulus. Depuis lors, de nombreuses autres méthodes d'inférence de trajectoire basées sur le pseudotime ont été développées, notamment Slingshot [Street et al., 2018], diffusion pseudotime [Haghverdi et al., 2016] et Palantir [Setty et al., 2019].

Méthodes basées sur la vélocité. Un outil fondamentalement différent pour l'inférence de trajectoire est la "RNA velocity", qui prédit l'état futur d'une cellule en fonction du rapport entre l'ARNm épissé et non épissé [La Manno et al., 2018, Bergen et al., 2020]. Cependant, la "RNA velocity" repose sur des modèles cinétiques simples qui peuvent mal interpréter la dynamique cellulaire, par exemple dans le cas de boosts transitoires de la transcription [Bergen et al., 2021].

Méthodes basées sur l'OT. L'utilisation du transport optimal (OT) pour l'inférence de trajectoire dans des jeux de données où plusieurs points temporels ont été profilés est particulièrement intéressante pour cette thèse. Waddington-OT [Schiebinger et al., 2019] utilise l'OT pour calculer les probabilités de transition entre des paires de cellules à des points temporels adjacents, puis utilise ces probabilités pour déduire les trajectoires cellulaires. Une autre approche basée sur l'OT a été mise au point par [Hashimoto et al., 2016], qui a utilisé l'OT pour apprendre un modèle causal de différenciation cellulaire via un réseau neuronal formé pour calculer le potentiel de différenciation des cellules.

Un wrapper pratique autour de différentes méthodes d'inférence de trajectoire est CellRank [Lange et al., 2022, Weiler et al., 2023], qui permet la combinaison de plusieurs méthodes pour améliorer la précision des trajectoires déduites. Cellrank permet également la découverte de gènes importants dans une trajectoire, en régressant leur expression par rapport à la variable pseudotime, en utilisant des techniques telles que les splines ou les modèles additifs généralisés [Lange et al., 2022, Weiler et al., 2023].

1.1.3 Impact biologique de l'analyse de cellules uniques

L'analyse des données de séquençage de cellules uniques a été utilisée pour étudier un large éventail de questions biologiques, notamment *(i)* la caractérisation de la diversité et de la fonction des cellules, et *(ii)* la compréhension des processus dynamiques tels que le développement et la maladie.

Caractérisation de la diversité et de la fonction des cellules Depuis ses débuts, les techniques de séquençage omique en cellule unique ont été utilisées pour sonder l'hétérogénéité cellulaire dans les tissus, en bonne santé et en maladie. Par exemple, la technique de séquençage MARS-seq a été utilisée pour identifier des groupes de cellules immunitaires transcriptionnellement distincts dans la rate [Jaitin et al., 2014], la moelle osseuse [Paul et al., 2015] et l'intestin [Gury-BenAri et al., 2016]. Le protocole scBS-seq a été utilisé pour explorer l'hétérogénéité de 5mC dans des cellules souches embryonnaires [Smallwood et al., 2014]. Par exemple, la plateforme Single Cell 3' de 10X Genomics pour snRNA-seq a été utilisée pour sonder les signatures transcriptomiques de types de cellules spécifiques dans le cerveau, dans le contexte des troubles du spectre autistique [Velmeshev et al., 2019]

et de la maladie d'Alzheimer [Leng et al., 2021]. Le protocole sci-ATAC-seq basé sur l'indexation combinatoire a été utilisé pour définir des signatures régulatrices spécifiques au type de cellule dans les cellules pancréatiques dans le contexte du diabète de type 2 [Rai et al., 2020]. Aujourd'hui, de nombreux efforts sont consacrés à l'assemblage de grands atlas multi-donneurs à partir de données profilées dans différentes expériences. L'initiative phare est le Human Cell Atlas [Regev et al., 2017], qui a déjà fourni un atlas transcriptomique du poumon [Sikkema et al., 2023], un atlas d'accessibilité de la chromatine du cerveau [Li et al., 2023b] et un atlas de méthylation de l'ADN du cerveau [Tian et al., 2023]. Ces grands jeux de données annotées peuvent servir de référence pour annoter des jeux de données plus petits à cellule unique [Lotfollahi et al., 2024].

Comprendre la nature dynamique de l'identité cellulaire Outre la caractérisation de l'hétérogénéité cellulaire, le deuxième objectif de la recherche sur le séquençage de cellules uniques a été de comprendre la dynamique de l'identité cellulaire dans des contextes tels que le développement, le vieillissement et la maladie. À partir des premières expériences sur des cellules uniques, les chercheurs ont décrit des profils continus entre des types de cellules bien caractérisés et ont cherché à comprendre les mécanismes sous-jacents à cette dynamique. L'analyse des profils transcriptionnels des cellules immunitaires Th17 obtenus avec SMART-seq a révélé un continuum d'états cellulaires le long de la différenciation et des gènes impliqués dans l'encéphalomyélite auto-immune expérimentale [Gaublomme et al., 2015]. Un tournant dans l'étude de la dynamique des cellules uniques a été la formalisation du *pseudotime* et de l'inférence de trajectoire [Trapnell et al., 2014], qui a ouvert la voie à des études à grande échelle des processus dynamiques. Par exemple, Diffusion Pseudotime [Haghverdi et al., 2016] et Waddington OT [Schiebinger et al., 2019] ont été utilisés pour analyser les trajectoires de plus de 100 000 cellules à partir de neuf points temporels au cours du développement de la souris [Pijuan-Sala et al., 2019].

1.2 Intégration conjointe de données de séquençage multimodal de cellules uniques

La compréhension des types de cellules, des états cellulaires et de l'évolution des cellules au cours du développement, du vieillissement et de la maladie nécessite une approche holistique basée sur la mesure simultanée de plusieurs couches moléculaires dans la même cellule [Vandereyken et al., 2023]. Ceci est facilité par le séquençage apparié multimodal de cellules uniques, un terme générique désignant le séquençage conjoint de plusieurs omiques et le séquençage omique résolu spatialement [Heumos et al., 2023].

Dans cette section, je couvre les protocoles multimodaux utilisés dans cette thèse (Section 1.2.1), l'analyse intégrative des données multimodales (Section 1.2.2), et les applications du séquençage et de l'intégration multimodale de cellules uniques (Section 1.2.3).

1.2.1 Technologies de séquençage multimodal apparié de cellules uniques

Profilage multiomique de cellules uniques

Les technologies multiomiques appariées à cellule unique telles que CITE-seq [Stoeckius et al., 2017] ou la plateforme commerciale 10x Multiome étendent les technologies monoomiques en permettant la lecture simultanée de plusieurs couches moléculaires des mêmes cellules. Elles complètent généralement la transcriptomique avec des informations protéomiques ou épigénétiques pour donner une caractérisation plus complète de la cellule [Vandereyken et al., 2023].

L'une des approches multiomiques à cellule unique les plus répandues est le profilage conjoint de l'expression génétique et de l'abondance des protéines de surface, en utilisant des protocoles comme REAP-seq [Han et al., 2014] ou CITE-seq [Stoeckius et al., 2017]. Ces technologies s'appuient sur le droplet-based scRNA-seq grâce à l'ajout d'oligonucléotides conjugués à des anticorps. Elles permettent de mesurer jusqu'à des centaines de protéines, mais la nécessité de créer un panel spécifique d'anticorps limite les possibilités de découverte non biaisée au niveau protéomique. L'analyse conjointe de la couche d'ARN (high-plex) et de la couche protéique (low-plex) est donc essentielle [Vandereyken et al., 2023].

L'autre type principal de protocole multiomique à cellule unique est le profilage conjoint de l'expression des gènes et de l'accessibilité de la chromatine en combinant la transcription inverse et la tagmentation. Une gamme de technologies a été développée, comme 10x Multiome, sci-CAR [Cao et al., 2018] ou scCAT-seq [Liu et al., 2019], qui diffèrent en termes de stratégies d'isolement et de code-barres, et de stratégies de séparation des fragments d'ARNm et d'ADN [Vandereyken et al., 2023]. Les technologies récentes comme TEA-seq permettent le profilage conjoint des protéines de surface en plus de l'expression des gènes et de l'accessibilité de la chromatine [Swanson et al., 2021].

Profilage omique de cellules uniques résolu spatialement

Le séquençage d'une cellule unique nécessite généralement la dissociation du tissu (voir Section 1.1.1). Cependant, la dissociation supprime les cellules de leur contexte, obscurcissant les schémas spatiaux et les interactions entre cellules [Trapnell, 2015, Vandereyken et al., 2023, Larsson et al., 2021]. Les technologies de *transcriptomique spatiale* permettent de profiler les cellules sans dissociation, comblant ainsi le fossé entre les aspects structurels et fonctionnels de la cellule [Palla et al., 2022]. Pour cette raison, la transcriptomique spatiale a été nommée méthode de l'année 2020 par Nature Methods [Marx, 2021].

Les technologies de transcriptomique spatiale peuvent être séparées en deux catégories: *basées sur l'optique* et *basées sur les points* [Palla et al., 2022].

- Les techniques basées sur l'optique telles que MERFISH [Xia et al., 2019] utilisent des sondes fluorescentes pour un ensemble prédéfini de gènes et détectent chaque ARNm marqué ainsi que ses coordonnées spatiales à une résolution subcellulaire. Cependant, les techniques basées sur l'optique ne permettent pas de profiler l'ensemble du transcriptome, ce qui limite son potentiel de découverte non biaisée [Palla et al., 2022].
- Cette thèse se concentre sur les méthodes basées sur des "spots" comme Stereo-seq [Chen et al., 2022a], qui permettent un profilage complet du transcriptome à l'aide du code-barres spatial [Palla et al., 2022]. Les méthodes basées sur des spots génèrent une matrice de comptage $\mathbf{A} \in \mathbb{N}^{m \times n}$ où m est le nombre de gènes et n est le nombre de codes-barres spatiaux. De plus, chacun des n codes-barres correspond à une position spatiale $(x, y) \in \mathbb{R}^2$.

Les premières méthodes basées sur des spots, telles que [Ståhl et al., 2016] (qui a ensuite été acquise par 10X Genomics), avaient une résolution spatiale supérieure à la taille cellulaire typique, c'est-à-dire un nombre n de codes-barres inférieur au nombre de cellules. La résolution subcellulaire a depuis été obtenue par des technologies telles que Slide-seq [Rodriques et al., 2019], HDST [Vickovic et al., 2019] et Stereo-seq [Chen et al., 2022a], qui offrent un nombre n de codes-barres supérieur au nombre de cellules.

Cette haute résolution se fait au prix d'un faible "count depth" par bille, ce qui est généralement

résolu en regroupant les codes-barres dans des bins de la taille d'une cellule ou en se basant sur la segmentation cellulaire des images histologiques [Chen et al., 2022a]. On obtient alors une matrice de comptage $\mathbf{A} \in \mathbb{N}^{m \times n}$ où m est le nombre de gènes et n le nombre de cellules, comme dans la Section 1.1.1.

1.2.2 Aspects computationels de l'intégration conjointe de données multimodales appariées de cellules uniques

La popularité croissante du profilage multimodal de cellules uniques nécessite des méthodes sur-mesure pour l'analyse de données, différentes des méthodes développées pour l'analyse omique unimodale à cellule unique. L'analyse conjointe de plusieurs modalités appariées est également connue sous le nom d'intégration multimodale à cellule unique. Une multitude de méthodes ont été proposées pour combler ce besoin. Comme dans l'analyse unimodale à cellule unique, des normes ont émergé pour harmoniser l'application de ces méthodes, telles que Giotto [Dries et al., 2021] et SpatialExperiment [Righelli et al., 2022] en R ou Squidpy [Palla et al., 2021] et SpatialData [Marconato et al., 2024] en Python.

Les premières étapes du pipeline d'analyse telles que le contrôle qualité et la normalisation sont généralement appliquées séparément pour chaque modalité [Heumos et al., 2023]. Par conséquent, dans ce qui suit, je me concentre sur les méthodes de réduction de dimensionnalité et d'inférence de trajectoire adaptées aux données multimodales appariées en cellule unique.

Réduction dimensionnelle conjointe

Les techniques de réduction de dimensionnalité conjointe permettent d'intégrer des cellules dans un espace de faible dimension informé par plusieurs modalités. Les méthodes de réduction de dimensionnalité conjointe les plus performantes pour l'intégration multi-omique à cellule unique peuvent être largement classées en modèles *linéaires* tels que la factorisation matricielle [Cantini et al., 2021] et modèles *non linéaires* tels que les réseaux de neurones [Gayoso et al., 2022] ou les méthodes basées sur des graphes [Hao et al., 2021].

D'une part, les méthodes linéaires telles que MOFA+ [Argelaguet et al., 2020] et MEFISTO [Velten et al., 2022] bénéficient d'une forte interprétabilité puisque leurs intégrations de faible dimension peuvent être directement liées aux données d'origine via leurs "loadings", qui représentent des features agrégées telles que les "métagènes" ou les "métapeaks" [Stein-O'Brien et al., 2018]. Cependant, ils sont limités par leur linéarité, qui peut ne pas capturer les relations complexes entre les différentes modalités [Stanojevic et al., 2022].

D'un autre côté, les méthodes d'apprentissage profond telles que Cobolt [Gong et al., 2021], MultiVI [Ashuach et al., 2023], totalVI [Gayoso et al., 2021], GLUE [Cao and Gao, 2022] ou les modèles basés sur des graphes tels que Weighted Nearest Neighbours [Hao et al., 2021] sont mieux équipés pour gérer les non-linéarités des données multimodales unicellulaires. Mais contrairement aux modèles linéaires, ils ne bénéficient pas d'une interprétabilité naturelle au niveau des features.

Trouver un équilibre entre l'interprétabilité des méthodes linéaires et l'expressivité des méthodes non linéaires est donc crucial pour tirer le meilleur parti des données multimodales unicellulaires, et constitue un objectif du Chapitre 4.

Inférence de trajectoire

Comme dans l'analyse unimodale à cellule unique, les méthodes d'inférence de trajectoire basées sur le pseudotime [Welch et al., 2017, Hou and Ji, 2022, Pham et al., 2023], la vélocité [Li et al., 2023a] et l'OT [Shen et al., 2023, Klein et al., 2023] sont utilisées pour attribuer des cellules à des trajectoires continues de destin cellulaire [Weiler et al., 2023].

Méthodes basées sur le pseudotime. Les méthodes de pseudotime basées sur la similarité cellule-cellule comme la diffusion Pseudotime [Haghverdi et al., 2016] sont généralement basées uniquement sur la couche d'expression génétique, bien que des travaux récents aient cherché à exploiter les informations spatiales conjointes [Pham et al., 2023, Hou and Ji, 2022] ou les informations épigénomiques [Welch et al., 2017].

Méthodes basées sur la vélocité. De même, le modèle original de la vélocité ARN ne prend en compte que les informations sur l'expression des gènes [La Manno et al., 2018, Bergen et al., 2020] mais a récemment été étendu pour inclure les informations épigénétiques [Li et al., 2023a] et les informations sur les protéines [Gorin et al., 2020].

Méthodes basées sur l'OT. Moscot [Klein et al., 2023] et spaTrack [Shen et al., 2023] ont étendu l'OT de Waddington [Schiebinger et al., 2019] à la transcriptomique spatiale en utilisant respectivement une extension quadratique de l'OT et un coût personnalisé informé spatialement pour l'OT linéaire. Bien que les méthodes basées sur l'OT puissent en principe être appliquées à la réduction de la dimensionnalité conjointe des données multiomiques, de telles expériences n'ont pas encore été réalisées.

L'inférence des trajectoires de destin cellulaire à partir de la transcriptomique spatiale est essentielle pour découvrir les mécanismes sous-jacents à l'organisation spatiale dans les processus biologiques dynamiques tels que le développement [Palla et al., 2022]. Le Chapitre 5 propose une nouvelle méthode basée sur les flots de gradients de Wasserstein et l'OT quadratique (voir la Section 1.3.1), qui permet d'apprendre un modèle causal de différenciation à partir de coupes de transcriptomique spatiale profilées à plusieurs moments.

1.2.3 Impact biologique de l'intégration conjointe de données multimodales de cellules uniques

Les technologies de cellules uniques multimodales appariées offrent un aperçu précieux des contextes régulateurs et spatiaux de l'identité cellulaire. Ci-dessous, je décris quelques succès de l'analyse intégrative de données multimodales pour étendre les connaissances recueillies à partir de l'omique unimodale.

Caractérisation de la diversité et de la fonction des cellules Les cellules ont traditionnellement été annotées sur la base de protéines de surface à l'aide de cytométrie de flux. Par exemple, en immunologie, CD4 est un marqueur des cellules T auxiliaires. Le profilage conjoint de la transcriptomique et des protéines de surface a ainsi offert la possibilité de combler le fossé entre l'annotation traditionnelle du type de cellule basée sur les protéines de surface et la découverte non biaisée de signatures transcriptomiques. Par exemple, [Hao et al., 2021] exploite le profilage conjoint des marqueurs immunitaires et de l'expression des gènes pour identifier une sous-population de lymphocytes T mémoire CD8+ dans le sang périphérique avec des caractéristiques de cellules mémoires résidentes dans les tissus, pointant vers une population intermédiaire jusqu'alors inconnue. La même étude a révélé que le profilage conjoint de la transcriptomique et de l'accessibilité de la chromatine

était plus pertinent pour séparer les cellules T CD8+ naïves et CD4+ naïves que le profilage conjoint de la transcriptomique et des protéines de surface, soulignant l'importance des profils épigénétiques dans la découverte de l'hétérogénéité cellulaire. Enfin, la transcriptomique résolue spatialement a découvert des modèles spatiaux d'expression génétique qui seraient masqués par la dissociation dans l'ARN-seq unicellulaire. Par exemple, [Chen et al., 2020] étudie la niche cellulaire entourant la plaque amyloïde dans la maladie d'Alzheimer, identifiant ainsi des réponses transcriptionnelles aux stades précoce et tardif de la maladie.

Comprendre la nature dynamique de l'identité cellulaire Le profilage conjoint de l'accessibilité de la chromatine et de l'expression génétique peut révéler la dynamique régulatrice sous-jacente à la différenciation. Par exemple, SHARE-seq a été utilisé pour identifier le destin des cellules dans la différenciation des follicules pileux [Ma et al., 2020b]. Cette étude a observé que les changements dans l'accessibilité de la chromatine précédaient les changements dans l'expression des gènes au cours de la différenciation et a souligné des facteurs de transcription régulant la différenciation au sein des différentes lignées. [Genshaft et al., 2016] a utilisé des protéines et des ARN profilés conjointement pour quantifier la réponse au traitement à plusieurs moments dans une lignée cellulaire de cancer du sein, mettant en évidence les gènes qui présentent des signaux transcriptomiques et protéomiques différents. Une perspective passionnante pour l'étude de la dynamique de l'identité cellulaire est l'utilisation de la transcriptomique spatiale pour construire des atlas spatiotemporels étendus, tels que [Chen et al., 2022a, Liu et al., 2022] dans le développement de la souris et du poisson zèbre et [Wei et al., 2022] dans la régénération de l'axolotl. Le Chapitre 5 étudie ces jeux de données plus en détail.

1.3 Transport optimal

À la fin du XVIIIe siècle, Gaspard Monge a formulé le problème du transport d'un de terre d'un endroit à un autre de manière à minimiser le coût total du transport [Monge, 1781]. Deux siècles plus tard, une formulation exploitable du transport optimal (OT) sous forme de programme linéaire a valu à Leonid Kantorovich un prix Nobel d'économie [Kantorovich, 1942]. Aujourd'hui, les développements dans le calcul efficace de l'OT en ont fait un outil puissant dans l'apprentissage automatique, les statistiques et le traitement du signal [Peyré et al., 2019].

Des travaux pionniers tels que [Schiebinger et al., 2019] et [Hashimoto et al., 2016] ont lancé un domaine de recherche actif à l'interface de l'OT et de la biologie en cellule unique, ce qui motive l'approche de cette thèse pour l'intégration multimodale en cellule unique. Dans cette section, je donne un bref aperçu du cadre mathématique de l'OT 1.3.1, de ses aspects computationels 1.3.2 et des principales applications de l'OT à la biologie en cellule unique 1.3.3.

1.3.1 Contexte mathématique sur le transport optimal

Les paragraphes suivants décrivent la théorie sous-jacente aux contributions de cette thèse, y compris le problème original de Monge-Kantorovich, le problème quadratique de Gromov-Wasserstein et les flots de gradients de Wasserstein. Les monographies [Villani, 2003, Santambrogio, 2015] fournissent un compte rendu plus complet de la théorie de l'OT.

L'objectif commun de toutes les formulations est de faire correspondre une mesure source $\mu \in \mathcal{M}(\mathcal{X})$ à une mesure cible $\nu \in \mathcal{M}(\mathcal{Y})$ tout en minimisant une certaine notion de coût. Ici, $\mathcal{M}(\mathcal{X})$ désigne l'espace des mesures de Radon sur un espace métrique $\mathcal{X}$. En général, on considère les mesures de

probabilité $\mu \in \mathcal{P}(\mathcal{X})$, ce qui signifie que μ est positif et que $\int_{\mathcal{X}} \mathrm{d}\mu = 1$, bien qu'il existe des extensions aux mesures positives de masse arbitraire [Peyré et al., 2019].

Toutes les applications numériques de cette thèse impliquent des distributions de probabilité discrètes sur $\mathcal{X} = \mathbb{R}^d$. Dans ce cadre discret, notons $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ avec $\mathbf{x}_i \in \mathbb{R}^d$ et $\mathbf{a} \in \Sigma_n \overset{\text{def.}}{=} \{\mathbf{u} \in \mathbb{R}^n : \forall i \in [1, n], u_i \geq 0 \text{ et } \sum_i u_i = 1\}$.

Différents chapitres de cette thèse utilisent le cadre mathématique de l'OT pour représenter deux contextes biologiques radicalement différents.

- Dans le Chapitre 2, le Chapitre 3 et le Chapitre 4, **la distribution $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ représente une cellule**. Les vecteurs $\mathbf{x}_i \in \mathbb{R}^d$ correspondent à des gènes (ou d'autres features moléculaires), représentés dans un espace de dimension d. L'histogramme $\mathbf{a} \in \Sigma_m$, où m est le nombre de gènes, est proportionnel aux données de comptage normalisées.
- Dans le Chapitre 5, **la distribution $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ représente une population de cellules**. Les cellules $\mathbf{x}_i \in \mathbb{R}^d$ sont représentées dans un espace de dimension d, et correspondent à des poids a_i. L'histogramme $\mathbf{a} \in \Sigma_n$, avec n le nombre des cellules, peut être uniforme ou proportionnel à un taux de croissance estimé [Schiebinger et al., 2019].

La plupart des définitions de cette section sont données dans le cas discret mais s'étendent naturellement aux distributions de probabilité continues, comme décrit dans [Peyré et al., 2019]. Les algorithmes présentés dans la Section 1.3.2 pour résoudre les différents problèmes d'OT se concentrent cependant sur le cadre discret.

Problème d'appariement de Monge

La formulation originale de l'OT de Gaspard Monge [Monge, 1781] est un problème de matching, où un ouvrier cherche à déplacer des tas de sable d'une configuration $\mu \in \mathcal{P}(\mathcal{X})$ à une autre configuration $\nu \in \mathcal{P}(\mathcal{X})$ avec le moins d'effort.

Dans un cadre discret où μ et ν ont le même nombre de points et des masses uniformes, c'est-à-dire

$$\mu = \frac{1}{n} \sum_i \delta_{\mathbf{x}_i} \text{ with } \mathbf{x}_i \in \mathcal{X}, \quad \nu = \frac{1}{n} \sum_j \delta_{\mathbf{y}_j} \text{ with } \mathbf{y}_j \in \mathcal{X}, \tag{6.1}$$

L'OT de Monge revient au problème d'affectation

$$\min_{\sigma \text{ permutation}} \frac{1}{n} \sum_i c(\mathbf{x}_i, \mathbf{y}_{\sigma(i)}). \tag{6.2}$$

Dans l'Equation 1.2 la fonction $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ est la *métrique au sol* qui encode le coût du transport, généralement une distance euclidienne au carré. La fonction bijective $\sigma : [1, \dots, n] \to [1, \dots, n]$ *transporte* le point $\mathbf{x}_i$ au point $\mathbf{y}_{\sigma(i)}$.

Dans un contexte comme le Chapitre 5, où μ et ν sont deux populations de cellules, l'Equation 1.2 compare μ et ν en transportant les cellules $\mathbf{x}_i$ vers les cellules $\mathbf{y}_{\sigma(i)}$ d'une manière qui minimise une divergence entre leurs profils moléculaires. Cependant, le problème d'affectation défini dans l'Equation 1.2 ne peut pas comparer μ et ν s'ils ont des nombres de cellules différents.

Bien que ce problème d'affectation puisse être étendu à des mesures de probabilité discrètes plus générales (nombres de points différents et poids non uniformes), il n'admet pas toujours une solution dans ce cas [Peyré et al., 2019]. Une formulation plus exploitable du problème d'OT est le programme linéaire de Kantorovich, décrit dans la section suivante.

Programme linéaire de Kantorovich

La formulation du problème d'OT de Kantorovich est un programme linéaire [Kantorovich, 1942]. Il peut être vu comme une relaxation probabiliste du problème déterministe de Monge dans le sens où il permet la *séparation de masse*. Considérons des mesures de probabilité discrètes plus générales μ et ν, éventuellement avec des nombres de points différents et des masses non uniformes :

$$\mu = \sum_i a_i \delta_{\mathbf{x}_i} \text{ with } \mathbf{a} \in \Sigma_n \text{ and } \mathbf{x}_i \in \mathcal{X}, \quad \nu = \sum_j b_j \delta_{\mathbf{y}_j} \text{ with } \mathbf{b} \in \Sigma_m \text{ and } \mathbf{y}_j \in \mathcal{X}. \tag{6.3}$$

Le transfert de masse entre les mesures de probabilité discrètes μ et ν (comme défini dans l'Equation 1.3) est ainsi codé par un couplage $\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \{\mathbf{P} \in \mathbb{R}_+^{n \times m} \text{ tel que } \sum_j P_{i,j} = a_i \text{ et } \sum_i P_{i,j} = b_j\}$.

Le programme linéaire de Kantorovich s'écrit alors

$$\mathcal{W}_c(\mu, \nu) \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j} c(\mathbf{x}_i, \mathbf{y}_j) P_{i,j}. \tag{6.4}$$

Il s'agit d'un problème d'optimisation convexe qui peut être résolu en utilisant la boîte à outils de la programmation linéaire, comme expliqué dans la Section 1.3.2. Contrairement à la formulation de Monge, ce problème d'optimisation admet toujours une solution. Cependant, une solution optimale n'est pas nécessairement unique [Peyré et al., 2019].

Une caractéristique importante de l'Equation 1.4 est qu'elle définit non seulement une correspondance $\mathbf{P}$ entre μ et ν, mais aussi une distance $\mathbb{W}_p(\mu, \nu)$ entre les deux mesures de probabilité. En effet, pour le coût $c(\mathbf{x}, \mathbf{y}) = d(\mathbf{x}, \mathbf{y})^p$ avec d une distance sur $\mathcal{X}$ (typiquement la distance euclidienne), la quantité $\mathbb{W}_p(\mu, \nu) \overset{\text{def.}}{=} \mathcal{W}_c(\mu, \nu)^{1/p}$ est une distance sur $\mathcal{P}(\mathcal{X})$ appelée distance p-*Wasserstein*.

Le problème de minimisation convexe de Kantorovitch admet une formulation duale comme problème de maximisation concave, qui se lit comme suit :

$$\mathcal{W}_c(\mu, \nu) = \max_{\substack{\mathbf{f} \in \mathbb{R}^n,\, \mathbf{g} \in \mathbb{R}^m \\ \text{s.t. } f_i + g_j \leq c(\mathbf{x}_i, \mathbf{x}_j)}} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle \tag{6.5}$$

où $\langle \mathbf{f}, \mathbf{a} \rangle \overset{\text{def.}}{=} \sum_i f_i a_i$ est le produit scalaire usuel dans $\mathbb{R}^n$.

Le Chapitre 2, le Chapitre 3 et le Chapitre 4 proposent de voir une cellule sous forme d'histogramme $\mathbf{a} \in \Sigma_m$ sur un ensemble fixé de m features moléculaires $\mathbf{x}_i$ (par exemple des gènes). Comparer deux cellules $\mathbf{a}$ et $\mathbf{b}$ étant donnée une matrice de coût par paire $\mathbf{C} \in \mathbb{R}^{m \times m}$ entre les features $\mathbf{x}_i$, c'est-à-dire $C_{i,j} = c(\mathbf{x}_i, \mathbf{x}_j)$, une notation pratique est

$$\mathrm{W}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \mathcal{W}_c(\sum_i a_i \delta_{\mathbf{x}_i}, \sum_j b_j \delta_{\mathbf{y}_j}). \tag{6.6}$$

Régularisation entropique

La plupart des succès récents de l'OT dans les problèmes d'apprentissage automatique de grande dimension reposent sur la régularisation entropique de l'Equation 1.4, introduite comme suit dans [Cuturi, 2013].

$$\mathcal{W}_c^\varepsilon(\mu,\nu) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathcal{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} c(\mathbf{x}_i,\mathbf{y}_j)P_{i,j} - \varepsilon \mathrm{E}(\mathbf{P}), \tag{6.7}$$

où $\mathrm{E} : \mathbf{P} \mapsto -\sum_{i,j} P_{i,j}(\log P_{i,j} - 1)$ est l'entropie de Shannon. E est 1-fortement concave sur $\mathcal{U}(\mathbf{a},\mathbf{b})$, ce qui rend le problème d'optimisation ε-fortement convexe. Par conséquent, il possède un minimiseur unique $\mathbf{P}^\star$, contrairement aux formulations de Monge et Kantorovich.

Comme dans le problème de Kantorovich, la formulation régularisée par l'entropie admet une formulation duale, définie comme suit.

$$\mathcal{W}_c^\varepsilon(\mu,\nu) \overset{\text{def.}}{=} \max_{\mathbf{f}\in\mathbb{R}^n,\ \mathbf{g}\in\mathbb{R}^m} \langle \mathbf{f},\mathbf{a}\rangle + \langle \mathbf{g},\mathbf{b}\rangle - \varepsilon\langle e^{\mathbf{f}/\varepsilon}, \mathbf{K}e^{\mathbf{g}/\varepsilon}\rangle, \tag{6.8}$$

où $\mathbf{K} \overset{\text{def.}}{=} [e^{-c(\mathbf{x}_i,\mathbf{x}_j)/\varepsilon}]_{i,j}$ est le noyau de Gibbs associé à la fonction de coût c. [Cuturi, 2013] propose un algorithme rapide et simple pour calculer $\mathcal{W}_c^\varepsilon(\mu,\nu)$: l'algorithme de Sinkhorn, décrit plus en détail dans la Section 1.3.2, qui coïncide avec une bloc coordinate gradient ascent sur les potentiels duaux $\mathbf{f},\mathbf{g}$.

Enfin, le théorème de l'enveloppe de Danskin prouve que $\mathcal{W}_c^\varepsilon(\mu,\nu)$ est différentiable par rapport aux histogrammes $\mathbf{a},\mathbf{b}$ et aux positions $\mathbf{x}_i,\mathbf{y}_j$, ce qui est une propriété souhaitable pour l'apprentissage automatique. Comme précédemment, étant donné une matrice de coûts par paire $\mathbf{C}$ entre les vecteurs $\mathbf{x}_i$, nous notons $\mathrm{W}_\mathbf{C}^\varepsilon(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \mathcal{W}_c^\varepsilon(\mu,\nu)$.

Le Chapitre 4 utilise la fonctionnelle convexe $\mathbf{a} \mapsto \mathrm{W}^\varepsilon_{mathbfC}(\mathbf{a},\mathbf{a}_\text{ref})$ comme objectif pour approcher une cellule $\mathbf{a}_\text{ref} \in \Sigma_m$ avec la factorisation $\mathbf{a} = \mathbf{H}\mathbf{w} \in \Sigma_m$. Ici, $\mathbf{H} \in \mathbb{R}_+^{m\times d}$ est un dictionnaire dont les colonnes $\mathbf{h}_j \in \Sigma_m$ sont des histogrammes sur les m gènes (ou autres features moléculaires), et l'embedding $\mathbf{w} \in \Sigma_d$ est un histogramme sur les d facteurs latents.

Divergences de Sinkhorn

Alors que les distances p-Wasserstein définissent une métrique entre les mesures de probabilité, la régularisation entropique définie dans l'Equation 1.7 n'est pas une métrique. Plus important encore pour les problèmes numériques pratiques, la quantité $\mathcal{W}_c^\varepsilon(\mu,\mu) > 0$ pour $\varepsilon > 0$. Pour remédier à ce biais, [Genevay et al., 2018] a proposé d'utiliser la divergence de Sinkhorn, une version débiaisée de $\mathcal{W}_c^\varepsilon$ définie comme

$$\overline{\mathcal{W}_c^\varepsilon}(\mu,\nu) \overset{\text{def.}}{=} \mathcal{W}_c^\varepsilon(\mu,\nu) - \tfrac{1}{2}\left(\mathcal{W}_c^\varepsilon(\mu,\mu) + \mathcal{W}_c^\varepsilon(\nu,\nu)\right). \tag{6.9}$$

[Feydy et al., 2019] donne un compte rendu approfondi des propriétés des divergences de Sinkhorn, y compris le fait qu'elles interpolent entre les distances de Wasserstein et les Maximum Mean Discrepancies (MMD).

De manière analogue au paragraphe précédent, nous notons $\overline{\mathrm{W}_\mathbf{C}^\varepsilon}(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \mathrm{W}_\mathbf{C}^\varepsilon(\mathbf{a},\mathbf{b}) - \frac{1}{2}\left(\mathrm{W}_\mathbf{C}^\varepsilon(\mathbf{a},\mathbf{a}) + \mathrm{W}_\mathbf{C}^\varepsilon(\mathbf{b},\mathbf{b})\right)$. Le Chapitre 2 propose d'utiliser $\overline{\mathrm{W}_\mathbf{C}^\varepsilon}$ comme notion de similarité entre les cellules $\mathbf{a},\mathbf{b}$ pour améliorer l'analyse en aval, et le Chapitre 3 introduit une nouvelle façon d'apprendre un coût par paire $\mathbf{C}$ entre les features $\mathbf{x}_i,\mathbf{x}_j$.

Apprentissage des métriques au sol

Un aspect crucial de l'OT est la métrique de coût c, qui encode les relations entre les features. Bien qu'une distance euclidienne au carré entre les features soit suffisante dans la plupart des cas, les problèmes pratiques d'apprentissage automatique peuvent nécessiter une métrique sur mesure.

- *Feature engineering.* Une première stratégie consiste à considérer un coût au sol euclidien sur un embedding bien choisi des données. Par exemple, la Word Mover Distance [Kusner et al., 2015] entre les histogrammes de type bag-of-word utilise des distances euclidiennes sur des embeddings Word2vec précalculés, c'est-à-dire

$$c(\mathbf{x}_i, \mathbf{y}_j) = \|\mathrm{W2V}(\mathbf{x}_i) - \mathrm{W2V}(\mathbf{y}_j)\|_2$$

  où la fonction W2V attribue un embedding $\mathrm{W2V}(\mathbf{x}_i) \in \mathbb{R}^{d'}$ à un mot encodé en one-hot vector $\mathbf{x}_i \in \mathbb{R}^d$. La Gene Mover Distance [Bellazzi et al., 2021] a adapté cette idée à la transcriptomique en cellule unique en utilisant des embeddings Gene2vec [Du et al., 2019]. Parallèlement, les travaux présentés dans le Chapitre 2 on proposé de matérialiser les features moléculaires (par exemple les gènes) sous forme de lignes de la matrice de comptage et de calculer directement les distances entre ces vecteurs.

- *Apprentissage supervisé de métrique au sol.* Dans un cadre supervisé, on peut tirer parti de méthodes telles que [Cuturi and Avis, 2014, Wang and Guibas, 2012] qui optimisent une fonction F sur le cône des matrices de distance en se basant sur des informations par paires $\omega_{k,l}$ entre les histogrammes $\mathbf{a}_k, \mathbf{b}_l$. Par exemple, [Cuturi and Avis, 2014] considère un problème de classification d'images et définit $\omega_{k,l} = 1$ si les histogrammes $\mathbf{a}_k$ et $\mathbf{b}_l$ ont le même label, et $\omega_{k,l} = -1$ sinon. L'objectif F est alors défini comme

$$F(\mathbf{C}) = \sum_{k,l} \omega_{k,l} \mathrm{W}_{\mathbf{C}}(\mathbf{a}_k, \mathbf{b}_l).$$

- *Adversarial optimization.* L'apprentissage de métrique au sol peut également être effectué via l'adversarial optimization. De la même manière que les réseaux antagonistes génératifs (GAN), [Genevay et al., 2018] entraîne un modèle génératif f_θ pour mapper une distribution uniforme μ sur une distribution cible ν (par exemple des images), tout en apprenant une métrique c_φ capable de faire la distinction entre les données générées et les données réelles :

$$\min_\theta \max_\varphi \overline{\mathcal{W}^\varepsilon_{c_\varphi}}(f_{\theta\#}\mu, \nu).$$

  Dans le cas discret $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$, la mesure pushworward $f_{\theta\#}\mu$ est simplement l'application ponctuelle de f_θ aux points $\mathbf{x}_i$. Plus formellement, pour $B \subset \mathcal{X}$, $(f_{\theta\#}\mu)(B) = \mu(f_\theta^{-1}(B))$ et vérifie $\int g \, \mathrm{d}\,(f_{\theta\#}\mu) = \int g \circ f_\theta \, \mathrm{d}\mu$, pour g mesurable et $g \circ f_\theta$ intégrable par rapport à μ [Bogachev and Ruas, 2007].

- *Inverse OT.* Enfin, un autre problème d'optimisation bi-niveau appelé OT inverse vise à apprendre une métrique c qui minimise une divergence $\mathcal{D}$ entre un couplage observé $\hat{\mathbf{P}}$ et le couplage $\mathbf{P}_c$ induit par c (c'est-à-dire le couplage minimisant l'Equation 1.7), avec un terme de régularisation $\mathcal{R}(c)$ [Sun et al., 2020], c'est-à-dire

$$\min_c \mathcal{D}(\hat{\mathbf{P}}, \mathbf{P}_c) + \mathcal{R}(c).$$

Par exemple, [Dupuy and Galichon, 2014] considère les relations entre les personnes comme un couplage observé $\hat{\mathbf{P}}$ et apprend une métrique c basée sur des features telles que le niveau d'éducation ou les traits de personnalité. On peut choisir $\mathcal{D}$ et $\mathcal{R}$ tels que ce problème d'optimisation bi-niveau soit convexe [Carlier et al., 2023]. [Samaran et al., 2024] voit l'alignement des données single-cell non appariées comme un problème d'OT inverse, où le coût c est la distance euclidienne sur l'espace latent d'autoencodeurs.

Le Chapitre 3 introduit une nouvelle approche pour apprendre une matrice de coût $\mathbf{C}$ qui ne nécessite *(i)* aucun feature engineering *(ii)* aucune supervision *(iii)* aucune optimisation adversariale et *(iv)* aucun matching observé. La méthode proposée est appliquée à un ensemble de données à cellule unique pour apprendre un coût $\mathbf{C}$ codant la géométrie de l'espace des gènes. Comme il n'existe pas de notion simple de coût entre les gènes, l'apprentissage de $\mathbf{C}$ présente un intérêt biologique et définit une distance $\mathrm{W}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ entre les cellules $\mathbf{a}, \mathbf{b} \in \Sigma_m$ (qui sont des histogrammes sur l'espace des gènes).

Gromov-Wasserstein

De nombreux problèmes numériques tels que la reconnaissance de formes impliquent le matching de points à une isométrie près (par exemple des rotations ou des translations). Dans sa formulation originale, l'OT ne peut pas gérer le matching d'une manière invariante aux isométries. La distance de Gromov-Wasserstein telle qu'introduite dans [Mémoli, 2011] est une extension quadratique de l'Equation 1.4 qui permet une telle invariance.

Elle permet de faire correspondre des mesures de probabilité définies sur différents espaces métriques $\mathcal{X}$ et $\mathcal{Y}$ en faisant correspondre des distances par paires plutôt que des points. Dans le cas discret,

$$\mu = \sum_i a_i \delta_{\mathbf{x}_i} \text{ with } \mathbf{a} \in \Sigma_n \text{ and } \mathbf{x}_i \in \mathcal{X}, \quad \nu = \sum_j b_j \delta_{\mathbf{y}_j} \text{ with } \mathbf{b} \in \Sigma_m \text{ and } \mathbf{y}_j \in \mathcal{Y}. \tag{6.10}$$

Le problème de Gromov-Wasserstein tel qu'introduit par [Mémoli, 2011] s'écrit

$$\mathcal{GW}(\mu, \nu) \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j,i',j'} \ell\left(c_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_{i'}), c_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_{j'})\right) P_{i,j} P_{i',j'} \tag{6.11}$$

où usuellement, $\ell(s, s') \overset{\text{def.}}{=} |s - s'|^p$ et $c_{\mathcal{X}} = d_{\mathcal{X}}^q, c_{\mathcal{Y}} = d_{\mathcal{Y}}^q$ sont des puissances de fonctions de distance définies sur $\mathcal{X}$ et $\mathcal{Y}$ respectivement. Dans ce cas, $\mathcal{GW}(\mu, \nu)^{1/p}$ définit une distance entre les espaces métriques $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ et $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ [Sturm, 2012].

Contrairement au programme linéaire de Kantorovich (Equation 1.4), ce problème est intrinsèquement non convexe et donc plus difficile à résoudre. Pour introduire de la régularité dans l'Equation 1.11, [Solomon et al., 2016, Peyré et al., 2016] introduit un terme de régularisation entropique analogue à l'Equation 1.7 :

$$\mathcal{GW}^{\varepsilon}(\mu, \nu) \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j,i',j'} \ell\left(c_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_k), c_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_l)\right) P_{i,j} P_{i',j'} - \varepsilon \mathrm{E}(\mathbf{P}). \tag{6.12}$$

Pour $\varepsilon > 0$, [Solomon et al., 2016, Peyré et al., 2016] propose un algorithme efficace pour calculer un minimum local approximatif, qui est basé sur une linéarisation itérative du problème (voir la Section 1.3.2).

Comme proposé dans [Demetci et al., 2022b], μ pourrait représenter une distribution de cellules $\mathbf{x}_i$

définie sur l'espace d'expression des gènes $\mathcal{X}$ et μ une distribution de cellules $\mathbf{y}_j$ définie sur l'espace d'accessibilité de la chromatine $\mathcal{Y}$. Alternativement, μ et ν peuvent représenter deux coupes de tissu en transcriptomique spatiale, c'est-à-dire des distributions de cellules dans l'espace physique [Zeira et al., 2022]. Dans ce cas $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$, mais comme les deux coupes ne sont pas alignées, μ et ν ne peuvent être comparés qu'à une isométrie près.

Fused Gromov-Wasserstein

Dans de nombreux contextes d'apprentissage automatique, on peut vouloir matcher des nuages points où certaines features sont comparables et d'autres ne le sont pas. C'est le cas par exemple en transcriptomique spatiale, où l'on peut comparer des cellules directement dans l'espace d'expression des gènes, ou à une isométrie près dans l'espace physique [Klein et al., 2023].

On peut formaliser cela comme $\mu \in \mathcal{P}(\mathcal{Z} \times \mathcal{X})$ et $\nu \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})$. Dans le cas discret,

$$\begin{aligned} \mu &= \sum_i a_i \delta_{(\mathbf{x}_i, \mathbf{r}_i)} \text{ with } \mathbf{a} \in \Sigma_n, \mathbf{x}_i \in \mathcal{Z}, \text{ and } \mathbf{r}_i \in \mathcal{X} \\ \nu &= \sum_j b_j \delta_{(\mathbf{y}_j, \mathbf{s}_j)} \text{ with } \mathbf{b} \in \Sigma_m, \mathbf{y}_j \in \mathcal{Z}, \text{ and } \mathbf{s}_j \in \mathcal{Y} \end{aligned} \tag{6.13}$$

La Fused Gromov-Wasserstein distance [Vayer et al., 2018b] interpole entre un terme linéaire de Wasserstein et un terme quadratique de Gromov-Wasserstein :

$$\mathcal{FGW}_\alpha^\varepsilon(\mu, \nu) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \alpha \sum_{i,j,i',j'} \ell\left(c_\mathcal{X}(\mathbf{r}_i, \mathbf{r}_k), c_\mathcal{Y}(\mathbf{s}_j, \mathbf{s}_l)\right) P_{i,j} P_{i',j'} + (1-\alpha) \sum_{i,j} c_\mathcal{Z}\left(\mathbf{x}_i, \mathbf{y}_j\right) P_{i,j} - \varepsilon \mathrm{E}(\mathbf{P}) \tag{6.14}$$

où $\alpha \in [0, 1]$ paramètrise le compromis entre les deux termes. Dans le cas de la transcriptomique spatiale mentionné ci-dessus, α interpole entre une divergence basée sur l'expression génétique ($\alpha = 0$) et une divergence basée sur la morphologie ($\alpha = 1$).

De la même manière que la distance de Gromov-Wasserstein, la distance Fused Gromov-Wasserstein n'est pas convexe mais peut être approximée efficacement pour $\varepsilon > 0$ (voir la Section 1.3.2). Le Chapitre 5 utilise ainsi $\mathcal{FGW}_\alpha^\varepsilon$ pour comparer une expérience de transcriptomique spatiale de référence μ avec une distribution prédite $\rho(\theta)$, où θ sont les paramètres d'un réseau de neurones.

Flots de gradients de Wasserstein

En biologie, l'étude des processus dynamiques implique souvent de minimiser une certaine notion d'énergie au cours du temps. Par exemple, Waddington a comparé la différenciation cellulaire à une bille roulant dans un paysage vers plusieurs états attracteurs possibles, qui représentent des types de cellules différenciées [Allen, 2015]. [Huang et al., 2007] montrent qu'un tel potentiel naît naturellement de réseaux simples de régulation des gènes.

On peut formaliser cette idée d'une cellule $\mathbf{x} \in \mathbb{R}^d$ minimisant une fonctionnelle $J : \mathbb{R}^d \to \mathbb{R}$ comme un *flot de gradient euclidien*, c'est-à-dire la contrepartie continue de la descente de gradient. Un flot de gradient euclidien $t \mapsto \mathbf{x}_t \in \mathbb{R}^d$ pour la fonctionnelle J vérifie l'équation aux dérivées partielles (EDP) suivante :

$$\frac{\partial \mathbf{x}_t}{\partial t} = -\nabla J(\mathbf{x}_t). \tag{6.15}$$

Une discrétisation "Euler explicite" de cette EDP est simplement une descente de gradient : $\mathbf{x}_{t+\tau} = \mathbf{x}_t - \tau \nabla J(\mathbf{x}_t)$. Une discrétisation "Euler implicite" est donnée par $\mathbf{x}_{t+\tau} = \arg\min_{\mathbf{x}} \tau J(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2$, qui est communément appelée descente de gradient proximale [Santambrogio, 2017].

Pour étudier la dynamique de populations au lieu de la dynamique cellulaire, on peut recourir aux *flots de gradient de Wasserstein*, qui généralisent les flots de gradient euclidiens à l'espace $(\mathcal{P}(\mathcal{X}), \mathbb{W}_2)$ des mesures de probabilité [Ambrosio et al., 2005, Santambrogio, 2017].

Soit une fonctionnelle $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$, les flots de gradient de Wasserstein $t \mapsto \mu_t \in \mathcal{P}(\mathbb{R}^d)$ vérifient l'EDP

$$\frac{\partial \mu_t}{\partial t} = \operatorname{div}(\mu_t \nabla_{\mathrm{W}} \mathcal{F}(\mu_t)), \tag{6.16}$$

où le gradient de Wasserstein $\nabla_{\mathrm{W}} \mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d$ est un champ de gradient défini comme $\nabla_{\mathrm{W}} \mathcal{F}(\mu) \overset{\text{def.}}{=} \nabla\left(\delta \mathcal{F}(\mu)\right)$, et $\delta \mathcal{F}(\mu) : \ RR^d \to \mathbb{R}$ est la première variation de $\mathcal{F}$ évaluée à μ, qui satisfait

$$\mathcal{F}(\mu + \eta\nu) = \mathcal{F}(\mu) + \eta \int_{\mathbb{R}^d} \delta \mathcal{F}(\mu)(\mathbf{x}) \mathrm{d}\nu(\mathbf{x}) + o(\eta). \tag{6.17}$$

Comme expliqué plus en détail dans [Santambrogio, 2017], les schémas de discrétisation d'Euler explicite et implicite pour l'Equation 1.16 sont donnés par

$$\mu_{t+\tau} = (\mathrm{Id} - \tau \nabla_{\mathrm{W}} \mathcal{F})_{\#} \mu_t \quad \text{and} \quad \mu_{t+\tau} = \arg\min_{\mu} \tau \mathcal{F}(\mu) + \tfrac{1}{2}\mathbb{W}_2^2(\mu, \mu_t). \tag{6.18}$$

Dans le cas d'une énergie potentielle $\mathcal{F}(\mu) = \int_{\mathbb{R}^d} J(\mathbf{x}) \mathrm{d}\mu(\mathbf{x})$, la première variation $\delta \mathcal{F}(\mu) = J$ ne dépend pas de μ. L'Equation 1.16 devient alors une EDP d'advection

$$\frac{\partial \mu_t}{\partial t} = \operatorname{div}(\mu_t \nabla J). \tag{6.19}$$

Dans le cas discret, le schéma explicite $\mu_{t+\tau} = (\mathrm{Id} - \tau \nabla J)_{\#} \mu_t$ correspond à la descente de gradient euclidienne (Equation 1.15) appliquée indépendamment en chaque point $\mathbf{x}_i$.

Jordan, Kinderlehrer et Otto [Jordan et al., 1998] montrent que dans le cas de la néguentropie de Gibbs-Boltzmann, $\mathcal{F}(\mu) = \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}x}(\mathbf{x})\right) \mathrm{d}\mu(\mathbf{x})$ et $\mathcal{F}(\mu) = +\infty$ si μ n'a pas de densité, l'Equation 1.16 devient l'équation de la chaleur

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t. \tag{6.20}$$

En combinant ceci avec un term d'advection, par exemple

$$\mathcal{F}(\mu) = \int_{\mathbb{R}^d} \left[J(\mathbf{x}) + \beta^{-1} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}x}(\mathbf{x})\right) \right] \mathrm{d}\mu(\mathbf{x}), \tag{6.21}$$

[Jordan et al., 1998] obtiennent l'équation de Fokker-Planck:

$$\frac{\partial \mu_t}{\partial t} = \operatorname{div}(\mu_t \nabla J) + \beta^{-1} \Delta \mu_t. \tag{6.22}$$

Dans le contexte de la transcriptomique spatiale, le Chapitre 5 vise à former un réseau de neurones J_θ tel que les distributions observées μ_{t_k} pour $k = 0, \dots, K$ puissent être modélisées par un flot de gradient pour l'énergie $\mathcal{F}(\mu) = \int_{\mathbb{R}^d} J_\theta(\mathbf{x}) \mathrm{d}\mu(\mathbf{x})$, discrétisée avec un schéma explicite.

Un cadre similaire a été proposé pour les données d'expression génique par [Hashimoto et al., 2016], qui utilise un schéma d'Euler explicite, et [Bunne et al., 2022] qui utilise un schéma d'Euler implicite. Cependant, [Hashimoto et al., 2016, Bunne et al., 2022] utilise une divergence de Sinkhorn pour évaluer les profils d'expression prédits, tandis que le Chapitre 5 propose un objectif de Fused Gromov-Wasserstein débiaisé pour inclure les coordonnées spatiales.

1.3.2 Contexte computationel sur le transport optimal

Ci-dessous, je décris les principes clés sous-jacents au calcul efficace des matchings d'OT pour les mesures de probabilité discrètes, qui permettent les applications de l'OT proposées dans cette thèse. Pour un compte rendu complet des aspects computationels de l'OT, voir [Peyré et al., 2019].

Des implémentations efficaces des algorithmes présentés ici sont présentes dans les bibliothèques Python suivantes : POT [Flamary et al., 2021], GeomLoss [Feydy et al., 2019] et OTT [Cuturi et al., 2022]. POT vise à fournir une boîte à outils complète pour l'OT, y compris des solveurs pour les cas régularisés et non régularisés. Geomloss se concentre sur le calcul rapide des divergences de Sinkhorn et de leurs gradients pour les applications d'apprentissage automatique à grande échelle. OTT est le plus récent ajout au paysage Python des solveurs d'OT. Il utilise un backend JAX pour calculer les problèmes d'OT régularisés en mettant l'accent sur l'évolutivité et la différentiabilité.

Transport optimal linéaire

Faire correspondre n points à n autres points avec la formulation de Monge est en général intrinsèquement difficile, car ce problème n'est pas convexe et nécessite de rechercher parmi $n!$ correspondances possibles [Peyré et al., 2019].

La première formulation exploitable de l'OT est donc le programme linéaire de Kantorovich, qui est un problème d'optimisation convexe qui peut être résolu efficacement en $\mathcal{O}(n^3)$ en utilisant l'algorithme hongrois lorsque les marginales sont à la fois égales et uniformes [Kuhn, 1955], ou $\mathcal{O}(n^3 \log(n)^2)$ en utilisant l'algorithme du network simplex [Tarjan, 1997].

La formulation régularisée par entropie donne lieu à un calcul efficace du plan de transport en utilisant l'algorithme de Sinkhorn [Sinkhorn, 1964], détaillé dans l'Algorithme 1. Les itérations de Sinkhorn ont une complexité de $\mathcal{O}(n^2)$, et correspondent à une bloc coordinate gradient ascent sur les potentiels duaux de l'Equation 1.8. Un avantage majeur de l'algorithme de Sinkhorn est qu'il est facilement parallélisable et bénéficie de l'accélération GPU [Cuturi, 2013].

L'utilisation de l'OT régularisé par l'entropie comme objectif dans les paramètres d'apprentissage automatique nécessite la différenciation efficace de $\mathcal{W}_c^\varepsilon$ par rapport aux histogrammes $\mathbf{a}, \mathbf{b}$ ou aux positions $\mathbf{x}_i, \mathbf{y}_j$. Une application naïve de Pytorch ou de la différenciation automatique de JAX déroulerait les itérations de Sinkhorn, devenant ainsi coûteuse pour un grand nombre d'itérations [Feydy et al., 2019]. Cependant, comme décrit dans la proposition 4.6. de [Peyré et al., 2019], on peut bénéficier du théorème de l'enveloppe de Danskin pour obtenir l'expression simple

$$\nabla \mathcal{W}_c^\varepsilon(\mu, \nu) = \nabla \mathcal{E}_{\mathbf{f}^\star, \mathbf{g}^\star} \text{ where } \mathcal{E}_{\mathbf{f}^\star, \mathbf{g}^\star} \stackrel{\text{def.}}{=} \langle \mathbf{f}^\star, \mathbf{a} \rangle + \langle \mathbf{g}^\star, \mathbf{b} \rangle - \varepsilon \langle e^{\mathbf{f}^\star/\varepsilon}, \mathbf{K} e^{\mathbf{g}^\star/\varepsilon} \rangle, \tag{6.23}$$

étant donné $(\mathbf{f}^\star, \mathbf{g}^\star)$ une solution de l'Equation 1.8 telle que la somme des coordonnées soit nulle. Ce calcul efficace du gradient est implémenté comme option par défaut dans GeomLoss [Feydy et al., 2019] et OTT [Cuturi et al., 2022].

Enfin, dans certains cas, la minimisation d'un objectif d'OT régularisé par l'entropie ne nécessite même pas de calculer les itérations de Sinkhorn. Par exemple, [Cuturi and Peyré, 2016] fournit une expression en forme close du dual de Fenchel-Legendre de la fonction $\mathbf{a} \mapsto \mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b})$ et de son gradient. [Rolet et al., 2016] exploite cette expression pour fournir un algorithme rapide d'apprentissage de dictionnaire sur des histogrammes avec un objectif d'OT régularisé par l'entropie.

Transport optimal quadratique

Les problèmes non régularisés d'optimisation quadratique de Gromov-Wasserstein et de Fused Gromov-Wasserstein sont non convexes, mais une solution locale peut être obtenue en utilisant l'algorithme de Frank-Wolfe [Frank et al., 1956], qui nécessite de résoudre un problème d'OT linéaire à chaque itération [Vayer et al., 2018b].

La régularisation entropique introduite dans [Peyré et al., 2016] ne rend pas le problème convexe, mais elle donne lieu à un algorithme efficace pour calculer un minimum local approximatif pour $\varepsilon > 0$. Pour une taille de pas bien choisie, [Peyré et al., 2016] montrent que la mirror descent pour la divergence de Kullback-Leibler coïncide avec un schéma de linéarisation itératif, où chaque étape nécessite de résoudre l'algorithme de Sinkhorn. Ceci est détaillé dans l'Algorithme 2.

Comme dans le cas linéaire, la différentiation à travers l'Algorithme 2 ne nécessite pas de dérouler les itérations. Un gradient peut être pris pour $\mathbf{P}$ et $\tilde{\mathbf{C}}$ [Brogat-Motte et al., 2022, Bunne et al., 2019] fixes, ce qui est le comportement par défaut dans OTT [Cuturi et al., 2022].

Flots de gradients de Wasserstein

Les applications numériques des flots de gradient de Wasserstein (présentés dans l'Equation 1.16) nécessitent la discrétisation du flot avec un schéma explicite ou implicite (voir l'Equation 1.18).

Considérons d'abord un cadre eulérien, où l'on cherche des histogrammes $\mathbf{a}_t \in \Sigma_n$ discrétisant un flot de gradient de Wasserstein pour une fonctionnelle $F : \Sigma_n \to \mathbb{R}$. Le schéma implicite s'écrit

$$\mathbf{a}_{t+\tau} = \arg\min_{\mathbf{a} \in \Sigma_n} \tau F(\mathbf{a}) + \tfrac{1}{2}\mathrm{W}_2^2(\mathbf{a}_t, \mathbf{a}). \tag{6.24}$$

L'Equation 1.24 peut être résolue efficacement dans le cas unidimensionnel [Kinderlehrer and Walkington, 1999]. En dimension supérieure, il existe des schémas d'éléments finis et de volumes finis [Burger et al., 2009, Carrillo et al., 2015]. De plus, [Peyré, 2015] propose une méthode approchée s'appuyant sur la régularisation entropique pour calculer un flot de gradient de Wasserstein discrétisé pour F convexe,

$$\mathbf{a}_{t+\tau} = \arg\min_{\mathbf{a} \in \Sigma_n} \tau F(\mathbf{a}) + \mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}_t, \mathbf{a}), \tag{6.25}$$

qui peut être résolu en utilisant un algorithme de Sinkhorn généralisé.

Un cadre lagrangien correspond à opérer sur les positions $\mathbf{x}_t^{(i)} \in \mathbb{R}^d$ plutôt que sur l'histogramme $\mathbf{a}_t \in \Sigma_n$. Lorsque la fonctionnelle $\mathcal{F}$ est finie pour les mesures de probabilités discrètes, le schéma d'Euler explicite revient à une descente de gradient euclidienne $\mathbf{X}_{t+\tau} = \mathbf{X}_t - \tau \nabla F(\mathbf{X}_t)$ où $F(\mathbf{X}) \overset{\text{def.}}{=} \mathcal{F}(\sum_i a_i \delta_{\mathbf{x}_i})$. Le gradient ∇F peut être obtenu par différentiation automatique.

Cependant, la néguentropie de Gibbs-Boltzmann n'est pas finie pour les mesures discrètes. La discrétisation de l'équation de Fokker-Planck (l'Equation 1.22) est obtenue en utilisant l'estimation de

densité [Beirlant et al., 1997] ou par l'ajout d'un terme de bruit avec un schéma d'Euler-Maruyama

$$\mathbf{x}_{t+\tau}^{(i)} = \mathbf{x}_t^{(i)} - \tau \nabla J(\mathbf{x}_t^{(i)}) + \sqrt{2\tau\sigma^2}\mathbf{z}_t \tag{6.26}$$

où $\mathbf{z}_t$ est un tirage indépendant d'une distribution normale multivariée [Kloeden et al., 1992].

L'Equation 1.26 est un schéma de discrétisation stochastique, où les points $\mathbf{x}_i$ n'interagissent pas. Les schémas déterministes impliquant des interactions entre points nécessitent plus de calculs, mais permettent de modéliser une gamme plus large de fonctionnelles d'entropie (et donc des EDP non linéaires sur la densité). L'approche la plus simple consiste à convoluer la mesure empirique avec une fonction de régularisation (c'est-à-dire en remplaçant les Diracs par des mixtures de gaussiennes) ce qui est généralement appelé la *blob method* [Carrillo et al., 2019].

Des solveurs implicites existent également pour le cadre langargien [Benamou et al., 2016, Westdickenberg and Wilkening, 2010]. Récemment, [Bunne et al., 2022] a proposé de reparamétrer le pas implicite en utilisant un Input Convex Neural Network (ICNN) f_ξ :

$$\mu_{t+\tau} = \arg\min_{\xi} \tau \mathcal{F}\left((\nabla f_\xi)_\# \mu_t\right) + \tfrac{1}{2}\int \|\mathbf{x} - \nabla f_\xi(\mathbf{x})\|^2 \mathrm{d}\mu_t(\mathbf{x}). \tag{6.27}$$

Le Chapitre 5 s'intéresse au problème de l'apprentissage d'une fonctionnelle $J_\theta : \mathbb{R}^d \to \mathbb{R}$ telle que les mesures discrètes observées μ_{t_k} puissent être modélisées avec un flot de gradient de Wasserstein pour la fonctionnelle $\mathcal{F} : \mu \mapsto \int J_\theta(\mathbf{x})\mathrm{d}\mu(\mathbf{x})$. Un tel cadre a été introduit à l'origine par [Hashimoto et al., 2016], qui considère une discrétisation d'Euler explicite et s'appuie sur la différenciation automatique pour calculer le gradient $\nabla_{\mathbf{x}} J_\theta(\mathbf{x})$. [Bunne et al., 2022] considère le schéma de discrétisation d'Euler implicite basé sur ICNN décrit ci-dessus, qui conduit à un problème d'optimisation bi-niveau coûteux. Un preprint récent propose une approche moins coûteuse utilisant des conditions d'optimalité qui éliminent la nécessité de résoudre un problème d'optimisation à bi-niveau [Terpin et al., 2024].

1.3.3 Applications du transport optimal en biologie des cellules uniques

Les avancées récentes dans la mise à l'échelle de l'OT pour les applications d'apprentissage automatique ont ouvert la possibilité d'analyser des données en cellule unique de grande dimension dans différentes modalités. Les paragraphes ci-dessous présentent quelques applications diverses de l'OT en génomique single-cell, soulignant la polyvalence de ce cadre pour modéliser des problèmes biologiques et laissant entrevoir les nouvelles opportunités explorées dans cette thèse.

OT comme similarité cellule-cellule

Le Chapitre 2, le Chapitre 3 et le Chapitre 4 explorent l'idée d'exploiter l'OT comme une distance entre des profils de cellules uniques. Dans ce cadre, une cellule est représentée par une distribution de probabilité discrète $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$. Les vecteurs $\mathbf{x}_i \in \mathbb{R}^d$ correspondent à des gènes (ou à d'autres features moléculaires), représentés dans un espace à d dimensions. L'histogramme $\mathbf{a} \in \Sigma_m$, où m est le nombre de gènes, est proportionnel aux données de comptage normalisées.

Le Chapitre 2 propose de matérialiser les features moléculaires $\mathbf{x}_i$ (par exemple les gènes) comme les lignes de la matrice de comptage $\mathbf{A}$, et de calculer des distances entre ces vecteurs pour obtenir une matrice de coût par paire $\mathbf{C} \in \mathbb{R}_+^{m\times m}$ entre les gènes. La divergence de Sinkhorn $\overline{\mathcal{W}}_c^\varepsilon(\mu,\nu)$ fournit alors une notion de distance entre deux cellules μ et ν, qui est utilisée pour les tâches de clustering.

Parallèlement, [Bellazzi et al., 2021] a proposé la Gene Mover Distance qui matérialise les gènes $\mathbf{x}_i$ comme des embeddings Gene2vec [Du et al., 2019], mais n'étend pas l'approche à d'autres omiques. [Tong et al., 2022] répète une expérience sur les données scRNA-seq du Chapitre 2 avec une approximation rapide de l'OT pour accélérer le calcul des distances par paires entre les cellules.

Le Chapitre 3 introduit une nouvelle approche d'apprentissage de métrique pour apprendre conjointement une matrice de coût $\mathbf{C}$ entre les features et une matrice de coût $\mathbf{D}$ entre les cellules, qui retrouvent la structure de type cellulaire dans un jeu de données de cellules mononucléaires du sang périphérique.

Le Chapitre 4 propose un modèle de réduction de dimensionnalité conjoint pour les données multi-omiques à cellule unique appariées basé sur la factorisation en matrices non négatives Wasserstein (Wasserstein NMF), qui utilise la distance d'OT régularisée par l'entropie $\mathrm{W}_{\mathbf{C}}^{\varepsilon}$ comme perte de reconstruction.

Alignement entre modalités basé sur l'OT

Une application notable de l'OT pour les données à cellule unique, bien que hors cadre pour cette thèse, est l'alignement des modalités *non appariées*. Cela correspond à considérer deux populations de cellules $\mu \in \mathcal{P}(\mathcal{X})$ et $\nu \in \mathcal{P}(\mathcal{Y})$ et à rechercher un couplage entre les cellules.

Par exemple, SCOT [Demetci et al., 2022b, Demetci et al., 2022a] utilise Gromov-Wasserstein pour aligner des données à cellule unique profilées pour différentes omiques. De même, NovoSpaRc [Moriel et al., 2021] utilise Fused Gromov-Wasserstein pour aligner la transcriptomique singel-cell sur la transcriptomique spatiale low-plex ou des coordonnées spatiales arbitraires.

Enfin, [Samaran et al., 2024] présente l'alignement de données à cellule unique non appariées comme un problème d'OT inverse, où une fonction de coût $c : (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y} \mapsto \|f_\theta(\mathbf{x}) - g_\varphi(\mathbf{y})\|^2$ est apprise à partir des données en entraînant des réseaux de neurones f_θ et g_φ.

Reconstruction des dynamiques de trajectoire à partir de données à travers le temps

L'application qui a suscité l'intérêt de la communauté single-cell pour l'OT est la modélisation de la dynamique de population à partir d'instantanés de cellules individuelles au fil du temps. Dans ce cadre, on considère des distributions discrètes $\mu_t = \sum_i a_t^{(i)} \delta_{\mathbf{x}_t^{(i)}}$ où $\mathbf{x}_t^{(i)} \in \mathbb{R}^d$ est le profil omique de la cellule i à l'instant t. L'histogramme $\mathbf{a}_t$ représente les poids des cellules, qui peuvent être uniformes ou proportionnels à un taux de croissance estimé [Schiebinger et al., 2019]. Il est important de noter que le séquençage impose la lyse des cellules, de sorte qu'aucune correspondance *a priori* n'existe entre deux points temporels adjacents μ_t et $\mu_{t+\tau}$.

[Schiebinger et al., 2019] adopte une approche statique pour reconstruire la dynamique cellulaire à partir d'instantanés en calculant les correspondances entre les points temporels adjacents μ_t et $\mu_{t+\tau}$ en utilisant l'OT régularisé par l'entropie. Dans le contexte de la transcriptomique spatiale, [Klein et al., 2023] étend l'approche de [Schiebinger et al., 2019] à la transcriptomique spatiale en considérant les instantanés $\mu_t = \sum_i a_t^{(i)} \delta_{(\mathbf{x}_t^{(i)}, \mathbf{r}_t^{(i)})}$ où $\mathbf{x}_t^{(i)} \in \mathbb{R}^d$ est toujours le profil omique d'une cellule, et $\mathbf{r}_t^{(i)} \in \mathbb{R}^2$ est sa position dans l'espace. Au lieu d'un couplage OT linéaire, [Klein et al., 2023] fait

correspondre deux points temporels adjacents en utilisant la fonction de Fused Gromov-Wasserstein régularisée par l'entropie.

Au lieu de cela, [Hashimoto et al., 2016] adopte une approche dynamique en apprenant une fonction de potentiel J_θ telle que le flot de gradient de Wasserstein pour ce potentiel (voir l'Equation 1.19) corresponde aux observations μ_t, en utilisant un objectif de divergence de Sinkhorn (voir l'Equation 1.9). Dans ce contexte, le Chapitre 5 comble une lacune critique : en fournissant un modèle causal dynamique des cellules à travers l'espace et le temps basé sur un objectif de Fused Gromov-Wasserstein, il est à [Hashimoto et al., 2016] ce que [Klein et al., 2023] est à [Schiebinger et al., 2019].

1.4 Contributions de cette thèse

Cette thèse apporte de nouvelles contributions à l'intégration multimodale unicellulaire conjointe. Le Chapitre 2, le Chapitre 3 et le Chapitre 4 se concentrent sur le clustering et la réduction de dimensionnalité pour les données multi-omiques unicellulaires appariées, tandis que le Chapitre 5 se concentre sur l'inférence de trajectoire en transcriptomique spatiale.

Optimal Transport improves cell-cell similarity inference in single-cell omics data (Chapitre 2)

Les similitudes par paires entre les cellules sont à la base de la plupart des analyses en aval des données omiques à cellule unique, comme le clustering et la visualisation. Les pipelines classique d'analyse à cellule unique utilisent des distances bin-bin comme la distance euclidienne pour comparer les cellules. Cependant, les distances bin-bin ignorent les interactions entre les features. Le Chapitre 2 propose de comparer les profils omiques à cellule unique à l'aide de l'OT, qui est robuste à la sparsité et intègre une compréhension géométrique de la structure complexe des features moléculaires.

Le Chapitre 2 formalise la cellule comme une mesure de probabilité discrète $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$. Les vecteurs $\mathbf{x}_i \in \mathbb{R}^m$ correspondent aux gènes (ou à d'autres features moléculaires), et l'histogramme $\mathbf{a} \in \Sigma_m$, où m est le nombre de gènes, est proportionnel aux données de comptage normalisées. Nous proposons de choisir $\mathbf{x}_i$ comme lignes de la matrice de comptage normalisée, et de calculer une matrice de coût par paires $\mathbf{C}$ entre les gènes. Nous proposons alors d'utiliser la divergence de Sinkhorn $\overline{\mathrm{W}}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ comme distance entre deux histogrammes $\mathbf{a}, \mathbf{b} \in \Sigma_m$.

Nous montrons que la divergence Sinkhorn entre les cellules reflète mieux la structure du type cellulaire que les distances traditionnellement utilisées dans l'analyse unicellulaire. Nous évaluons l'approche sur des données scRNA-seq simulées et sur des jeux de données omiques unicellulaires réels couvrant l'expression des gènes, l'accessibilité de la chromatine et la méthylation de l'ADN. Nous avons ensuite intégré notre approche dans un pipeline de clustering classique utilisant l'algorithme de Leiden et montrons une amélioration des performances de clustering sur tous les jeux de données.

Unsupervised ground metric learning with Wasserstein Singular Sectors (Chapitre 3)

Le chapitre précédent introduit une nouvelle façon de comparer les cellules, étant donné une matrice de coût ad hoc $\mathbf{C}$ qui encode la géométrie de l'espace des features. Cependant, l'absence de choix

simple du coût au sol motive l'apprentissage d'une telle métrique à partir des données. Le Chapitre 3 propose une méthode non supervisée pour apprendre la métrique au sol **C**.

Comme dans le chapitre précédent, le Chapitre 3 formalise les cellules sous forme d'histogrammes $\mathbf{a}_k \in \Sigma_m$ sur l'espace des gènes, et cherche un coût $\mathbf{C} \in \mathbb{R}_+^{m \times m}$ entre les gènes. De plus, le Chapitre 3 transpose le problème et formalise les gènes sous forme d'histogrammes $\mathbf{b}_i \in \Sigma_n$ sur l'espace des cellules. Nous fournissons un algorithme qui calcule une matrice de distance de Wasserstein par paires **C** entre les gènes et **D** entre les cellules telles que, à changement d'échelle près, $C_{i,j} = \mathrm{W}_{\mathbf{D}}(\mathbf{b}_i, \mathbf{b}_j)$ et $D_{k,l} = \mathrm{W}_{\mathbf{C}}(\mathbf{a}_k, \mathbf{a}_l)$. Nous fournissons des résultats sur l'existence et l'unicité de ces Vecteurs Singuliers Wasserstein. De plus, nous améliorons l'évolutivité de la méthode proposée en introduisant une régularisation entropique et une méthode stochastique des puissances itérées.

Sur un jeu de données synthétiques constitué d'histogrammes translatés, nous montrons que la métrique apprise intègre des propriétés géométriques et que l'algorithme proposé converge linéairement. Sur un jeu de données scRNA-seq de sang périphérique, nous montrons que la métrique **D** entre les cellules capture mieux la structure du type cellulaire du jeu de données que d'autres métriques. De plus, la métrique **C** entre les gènes capture les similitudes entre les marqueurs connus des types de cellules immunitaires, ce qui en fait un résultat biologiquement interprétable.

Paired single-cell multi-omics data integration with Mowgli (Chapitre 4)

L'analyse de données multiomiques unicellulaires appariées nécessite des méthodes de réduction de dimensionnalité conjointes adaptées. Les méthodes basées sur la factorisation matricielle sont populaires en raison de leur interprétabilité, mais peuvent manquer des relations non linéaires prises en compte par les méthodes d'apprentissage profond. Il est donc crucial d'améliorer les méthodes basées sur la factorisation matricielle pour trouver un équilibre entre interprétabilité et performances. Le Chapitre 4 propose une méthode de réduction de dimensionnalité conjointe pour les données multiomiques unicellulaires appariées basée sur la factorisation conjointe en matrices non négatives (NMF conjointe) avec une perte de reconstruction basée sur l'OT régularisé par l'entropie.

Pour chaque couche omique p (par exemple, l'expression des gènes), le Chapitre 4 considère le profil moléculaire d'une cellule comme un histogramme $\mathbf{a}^{(p)} \in \Sigma_{m_p}$ sur l'espace des features (par exemple les gènes), qui est équipé d'une matrice de coût $\mathbf{C}^{(p)}$. Nous proposons Mowgli, une méthode de NMF conjointe avec une perte de reconstruction basée sur l'OT régularisé par l'entropie, qui cherche à décomposer l'histogramme $\mathbf{a}^{(p)} \in \Sigma_{m_p}$ en un produit $\mathbf{H}^{(p)}\mathbf{w}$. La matrice $\mathbf{H}^{(p)} \in \mathbb{R}_+^{m_p \times k}$ est un dictionnaire spécifique à l'omique (p) dont les colonnes $\mathbf{h}_j^{(p)} \in \Sigma_{m_p}$ sont des histogrammes sur l'espace des gènes, des pics ou des protéines. L'embedding $\mathbf{w} \in \Sigma_k$ est partagé entre les omiques et représente l'embedding conjoint de la cellule dans un espace latent de k dimensions. Cet embedding est interprétable puisque chaque dimension peut facilement être mappée à une liste pondérée de gènes (ou d'autres features).

Dans des cas contrôlés difficiles dérivés de données de lignées cellulaires, Mowgli surpasse l'état de l'art en termes de qualité d'embedding et de performances de clustering. Dans des jeux de données plus complexes et hétérogènes couvrant plusieurs technologies multiomiques, les performances de Mowgli sont compétitives avec l'état de l'art tout en offrant des embeddings plus interprétables. Les embeddings hautement interprétable de Mowgli permettent la caractérisation approfondie de plusieurs groupes précis de lymphocytes T CD8+ dans un jeu de données TEA-seq, mesuré conjointement pour trois omiques.

Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein (Chapitre 5)

L'inférence de trajectoires cellulaires à partir de la transcriptomique spatiale est essentielle pour découvrir les mécanismes sous-jacents à l'organisation spatiale dans les processus biologiques dynamiques tels que le développement. Un cadre prometteur pour analyser les données de séquençage au fil du temps est l'apprentissage de flots de gradients de Wasserstein. Ce cadre fournit un modèle causal de différenciation construit autour d'un réseau de neurones représentant le potentiel de différenciation. Cependant, les méthodes existantes d'apprentissage de flots de gradients sont mal équipées pour gérer les informations spatiales et ne peuvent pas analyser les données transcriptomiques résolues spatialement. Le Chapitre 5 propose une nouvelle méthode pour exploiter les coordonnées spatiales qui utilise un objectif de Fused Gromov-Wasserstein pour apprendre un potentiel informé spatialement.

Le Chapitre 5 déplace l'attention des cellules individuelles vers les populations de cellules. Une coupe de transcriptomique spatiale au temps t est représentée par une distribution de probabilité discrète $\mu_t = \sum_i a_i \delta_{(\mathbf{x}_i, \mathbf{r}_i)}$. Les vecteurs $\mathbf{x}_i \in \mathbb{R}^d$ représentent le profil d'expression génique de la i-ième cellule, tandis que les vecteurs $\mathbf{r}_i \in \mathbb{R}^2$ représentent les coordonnées spatiales de la cellule. Les poids des cellules individuelles sont codés dans un histogramme $\mathbf{a} \in \Sigma_n$, avec n le nombre de cellules, qui peut être uniforme ou proportionnel à un taux de croissance estimé. Nous proposons STORIES, un modèle basé sur les flots de gradients de Wasserstein. STORIES apprend un potentiel de différenciation hautement interprétable J_θ paramétré par un réseau de neurones. En effet, la fonction de potentiel est analogue au pseudotime car elle ordonne les cellules selon leur stade de différenciation. De plus, le vecteur $-\nabla J_\theta$ est analogue à la vélocité car il permet de prédire l'état transcriptomique futur d'une cellule. Contrairement aux méthodes existantes, STORIES apprend J_θ en utilisant un objectif Fused Gromov-Wasserstein, qui permet de comparer les coordonnées spatiales de différentes coupes d'une manière invariante aux isométries.

Nous avons évalué notre méthode sur trois grands atlas spatiotemporels Stereo-seq couvrant le développement et la régénération. Dans ces atlas, nous démontrons que le matching Fused Gromov-Wasserstein est plus cohérent biologiquement que le matching linéaire utilisé dans les méthodes existantes basées sur un potentiel. Nous avons également effectué une analyse approfondie de la régénération des neurones de l'axolotl et de la gliogenèse de la souris. Dans ces contextes biologiques, STORIES a récupéré les trajectoires et les tendances attendues de l'expression des gènes pour des marqueurs connus. En outre, STORIES a découvert des gènes candidats et de potentiels régulateurs transcriptionnels qui pourraient motiver des recherches biologiques plus poussées.

Contributions logicielles

Les développements méthodologiques de cette thèse ont été distribués sous forme de packages Python open source:

- **Le package** `otscomics`, développé pour le Chapitre 2, permet de calculer des matrices de divergences de Sinkhorn par paires entre les cellules de données omiques à cellule unique. `otscomics` utilise POT [Flamary et al., 2021] pour l'algorithme de Sinkhorn et Pytorch [Paszke et al., 2019] pour les calculs GPU. `otscomics` groupe les calculs pour bénéficier de la parallélisation dans le solveur entropique de POT. Enfin, la documentation de `otscomics` montre comment utiliser l'outil avec un objet AnnData [Virshup et al., 2021] et les fonctions de clustering de Scanpy [Wolf et al., 2018].

`pip install otscomics` https://github.com/cantinilab/ot-scomics

- **Le package `wsingular`**, développé pour le Chapitre 3, permet de calculer des Vecteurs Singuliers Wasserstein, des Vecteurs Singuliers Sinkhorn et leurs variantes stochastiques (voir le Chapitre 3). `wsingular` utilise POT [Flamary et al., 2021] pour son solveur linéaire et son solveur entropique, et Pytorch [Paszke et al., 2019] pour les calculs GPU. De plus, `wsingular` fournit une fonction pour vérifier l'unicité en pratique des vecteurs singuliers en utilisant le critère du Theorème 3.2.5.

`pip install wsingular` https://github.com/CSDUlm/wsingular

- **Le package `mowgli`**, développé pour le Chapitre 4, permet d'effectuer une réduction de dimensionnalité conjointe sur des données multiomiques unicellulaires appariées en utilisant un modèle conjoint de factorisation en matrices non négatives (NMF) avec une perte de reconstruction basée sur l'OT (voir le Chapitre 4). `mowgli` est implémenté en Pytorch [Paszke et al., 2019] pour les calculs GPU et pour son optimiseur L-BFGS. De plus, `mowgli` fournit des fonctions pratiques pour visualiser les résultats et effectuer des analyses en aval. Enfin, `mowgli` prend les objets MuData [Bredikhin et al., 2022] en entrée et est parfaitement intégré dans l'écosystème Scverse [Virshup et al., 2023].

`pip install mowgli` https://github.com/cantinilab/mowgli

- **Le package `stories-jax`**, développé pour le Chapitre 5, permet d'effectuer une inférence de trajectoire sur des coupes de transcriptomique spatiale profilées à plusieurs moments. Contrairement aux packages précédents, `stories-jax` s'appuie sur JAX [DeepMind et al., 2020] pour l'accélération GPU et OTT [Cuturi et al., 2022] pour un calcul rapide des distances de Fused Gromov-Wasserstein. L'implémentation modulaire de `stories-jax` permet de choisir une discrétisation d'Euler explicite ou une discrétisation d'Euler implicite basée sur ICNN, ainsi que le nombre d'étapes dans la discrétisation et la présence de teacher-forcing (voir le Chapitre 5). Comme auparavant, `stories-jax` est parfaitement intégré dans l'écosystème Scverse et prend les objets AnnData comme entrée, ce qui facilite l'interface de `stories-jax` avec les outils Python standard pour l'inférence de trajectoire, tels que CellRank [Weiler et al., 2023].

`pip install stories-jax` https://github.com/cantinilab/stories

Appendix A

Optimal transport improves cell type identification in single-cell multi-omics

Supplementary material "Optimal Transport improves cell-cell similarity inference in single-cell omics data"

Index:

1. Supplementary Text 1. Splatter parameters for the generation of simulated scRNA-seq data.
2. Supplementary Text 2. Analysis of the transport plans between cells with similar activated transcriptional programs
3. Supplementary Text 3. Analysis of dropouts in the Liu scRNA cell lines dataset, for entropic regularized Optimal Transport and for classical Optimal Transport.
4. Supplementary Text 4. Comparing Optimal Transport against baseline metrics based on spectral clustering
5. Supplementary Text 5. Comparing a typical single-cell Louvain clustering workflow against its counterpart based on Optimal Transport
6. Supplementary Figures 1-5

Supplementary Text 1. Splatter parameters for the generation of simulated scRNA-seq data.

Parameters were estimated from the (Liu *et al.*, 2019) cell lines scRNA data through the function `splatEstimate` from R package Splatter (Zappia *et al.*, 2017).

The following parameters were refined manually:

- `nGenes` (number of genes in simulated dataset): 5000 for all simulations
- `batchCells` (number of cells in simulated dataset): 500, 1000 or 10000 depending on the simulation.
- `group.prob` (probability of cells belonging to a particular group): (⅓, ⅓, ⅓) everywhere except (.75, .20, .05) for the uneven groups dataset.
- `de.prob` (probability of a genes being differentially expressed across groups): 0.7 for all datasets, except 0.4 for the overlapping clusters dataset.

Counts are log-normalized in the same way as real scRNA-seq data.

Supplementary Text 2. Analysis of the transport plans between cells with similar activated transcriptional programs

In order to motivate the use of Optimal Transport for single-cell omics data we analyzed the biological information encoded in the transport plans P (also called coupling) between cells from the same cell type in the Li Tumor (Li *et al.*, 2017) dataset. As a ground cost C, we considered the cosine distance matrix between genes computed on the full dataset.

Analysis of the transport plans for T cells

We choose 100 random pairs of T cells from the dataset and compute the transport plan P associated with each pair. In order to find communities of genes between which transport is strongest, we first define a network of genes by considering the average of the computed transport plans as an adjacency matrix. In order to denoise this network, we restrict it to the 2,000 most expressed genes within these pairs of cells. Finally, we cluster the network by running the Louvain (Blondel *et al.*, 2008) clustering algorithm with resolution parameter 1 (Scanpy default value).

This operation yields 8 gene communities. Among those, only 3 account for at least 1% of the weights in the adjacency matrix. For those three communities we report characteristic biological pathways obtained through an unordered query to the g:Profiler API (Raudvere *et al.*, 2019). The sources considered in this query are "Gene Ontology molecular function (GO:MF)", "Gene Ontology cellular component (GO:CC)", and "Gene Ontology biological process (GO:BP)". We also report the hypergeometric p-value after correction for multiple testing.

Cluster 1 appears to be associated with intracellular structure: intracellular anatomical structure (GO:CC, p=9.997e-66), intracellular organelle (GO:CC, p=3.265e-56), organelle organization (GO:BP, p=3.881e-31).
Cluster 2 appears to be associated with translation: RNA binding (GO:MF, p=6.017e-104), mRNA metabolic process (GO:BP, p=1.014e-93), cytoplasmic translation (GO:BP, p=4.755e-103).
Cluster 3 appears to be T cell specific: T cell activation (GO:BP, p=4.409e-25), T cell differentiation (GO:BP, p=4.317e-16), regulation of T cell activation (GO:BP, p=5.776e-15).

These results suggest that Optimal Transport takes advantage of biological processes when computing similarities between T cells.

Analysis of the transport plans for B cells

We choose 100 random pairs of B cells from the dataset and compute the transport plan associated with each pair. Gene communities are defined using Louvain clustering with resolution 1, in the same way as for the T cells.
This operation yields 9 gene communities. Among those, only 3 account for at least 1% of the weights in the adjacency matrix. For those we report characteristic biological pathways and p-values obtained in the same way as for T cells.

Cluster 1 appears to be associated with intracellular structure: intracellular anatomical structure (GO:CC, p=6.958e-67), intracellular organelle (GO:CC, p=1.868e-56), intracellular transport (GO:BP, p=1.446e-29).
Cluster 2 appears to be associated with translation: RNA binding (GO:MF, p=3.897e-100), mRNA metabolic process (GO:BP, p=4.769e-90), cytoplasmic translation (GO:BP, p=1.557e-98).
Cluster 3 appears to be B cell specific: B cell activation (GO:BP, p=2.196e-12), innate immune response (GO:BP, p=7.889e-07), regulation of B cell activation (GO:BP, p=6.738e-07).

These results suggest that Optimal Transport takes advantage of biological pathways when computing similarities between B cells.

Supplementary Text 3. Analysis of dropouts in the Liu scRNA cell lines dataset, for entropic regularized Optimal Transport and for classical Optimal Transport.

In order to compare the robustness of entropic regularized optimal transport and classical optimal transport, we tested their performances on data with increasing amounts of dropouts. Dropouts refer to values in the gene expression matrix being randomly replaced with zeros. This type of noise is known to be one of the main challenges of scRNA-seq analysis (Lähnemann *et al.*, 2020).

We considered the Liu scRNA dataset. In order to make possible the computation of classical OT we selected only the 500 most varying genes. Both regularized OT and classical OT distance matrices were computed for five different settings: no added noise (41% of zeros), 59% of zeros, 71% of zeros, 80% of zeros and 86% of zeros.

The median computation time of the distance matrices for regularized OT was 0.07s while the median computation time for classical OT was 809.04s (see Supplementary Table 3), highlighting first of all a big computational advantage to the entropic regularization of OT.

We then computed the silhouette scores associated to each distance matrix and found that the performance of regularized OT was less sensitive to noise than classical OT. Indeed, the silhouette score of classical OT dropped by 57% between the case with no noise and the case with 86% of zeros, while the silhouette score of regularized OT dropped only by 21% between these two cases (Supplementary Table 3). This experiment thus motivates the use of entropic regularized OT for high dimensional data with many zero values.

Supplementary Text 4. Comparing Optimal Transport against baseline metrics based on spectral clustering

To test whether the improvement provided by OT in cell clustering depends on the clustering algorithm, we repeated the same tests as in Results section 3.2 with spectral clustering. Also in this case, the number of clusters was optimized based on the Silhouette score (Methods). The numbers of resulting clusters are listed in Supplementary Table 5, while the corresponding values of ARI and NMI are provided in Supplementary Figure 2.

In simulated data, Pearson correlation and OT metrics always detected the correct number of clusters (Supplementary Table 5) and achieved perfect clustering according to ARI and NMI (Supplementary Figure 2A).

For scRNA-seq data, both Pearson correlation and OT correctly predicted the clusters present in Liu scRNA. The two CRC datasets contain seven cell types, corresponding to our ground-truth labels. For Li Tumor, both Pearson and OT predicted the presence of three clusters. In contrast, for Li NM, Pearson correlation over-clustered the data and inferred 13 clusters, while OT suggested the presence of six clusters. Therefore, in colorectal cancer, a better approximation of the true clustering present in the data has been obtained with hierarchical clustering. Finally, for Li cell lines, both OT and Pearson correctly identified seven clusters. Concerning the NMI and ARI scores (Supplementary Figure 2B), OT outperformed Pearson correlation for two of the four datasets, while comparable performances are observed in Liu scRNA and Li NM data.

Concerning other single-cell omics, Pearson correlation and OT correctly identified three clusters for Liu scATAC data. Regarding Leukemia scATAC data, involving six cell lines, Pearson over-clustered the data, predicting 16 clusters, while OT predicted seven clusters. Finally, in methylation data both OT and Pearson underestimated the real number of clusters. Regarding ARI and NMI (Supplementary Figure 2C), for Liu ATAC data, the results are optimal with both metrics. In all other datasets, performances are slightly changed compared to those observed with hierarchical clustering (Figure 3C). In mouse scMethylation and Leukemia scATAC data, OT showed better performances than Pearson correlation, while for human scMethylation data, Pearson correlation performed better.

Overall, the clustering conclusions derived with hierarchical clustering are confirmed with spectral clustering. Indeed, there is a good agreement between spectral clustering and hierarchical clustering for simulated data and scRNA-seq data. Of note, we obtained similar results for fixed numbers of clusters (Supplementary Table 5).

Supplementary Text 5. Comparing a typical single-cell Louvain clustering workflow against its counterpart based on Optimal Transport

To test whether the improvement provided by OT in cell clustering depends on the clustering algorithm, we repeated the same tests as in Results section 3.3 with Louvain clustering. The resolution parameter for the Louvain clustering is optimized based on the Silhouette score (Methods). The numbers of resulting clusters are listed in Supplementary Table 5. Of note, the results are not affected by the choice of the resolution parameter (Supplementary Figure 5).

In simulated data, as observed for Leiden clustering, the data tend to be over-clustred. This happens more frequently in the typical clustering workflow compared to its OT counterpart (Supplementary Table 5). In addition, OT brings an improvement over the typical clustering workflow according to ARI and NMI in all cases (Supplementary Figure 4A). In particular, in three out of five datasets OT achieved perfect performances.

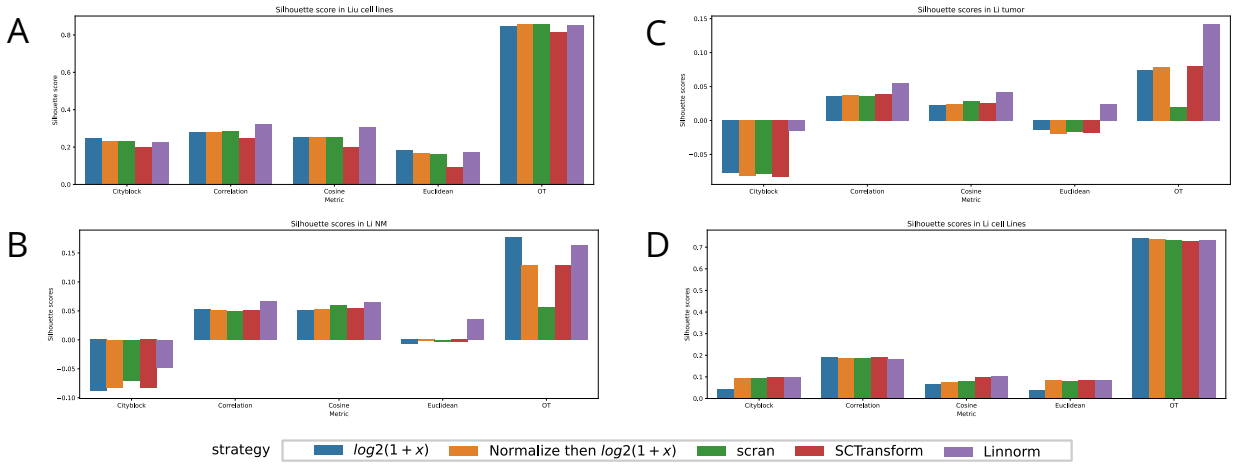
For scRNA-seq data, both the typical clustering workflow and OT correctly predicted the clusters present in Liu scRNA and under-clustered the two CRC datasets (Li tumor and Li NM). Finally, for Li cell lines, OT suggested the presence of nine clusters, while the typical clustering workflow correctly identified seven clusters. Concerning the NMI and ARI scores, OT outperformed the typical clustering workflow in Liu scRNA, Li Tumor and Li cell lines, while the typical clustering workflow performed better in Li NM (Supplementary Figure 4B).

Concerning other single-cell omics, OT correctly identified three clusters for Liu scATAC data, while the typical clustering workflow found seven clusters. Regarding Leukemia scATAC data, involving six cell lines, both OT and the typical clustering workflow under-clustered the data. In methylation data, the typical clustering workflow predicts a number of clusters slightly closer to the correct one. Regarding ARI and NMI (Supplementary Figure 4C), OT outperforms the typical clustering workflow in Liu scATAC, scMethylation human and scMethylation mouse and obtains perfect performances in Liu ATAC data. Finally on leukemia it performs comparably to the typical clustering workflow.

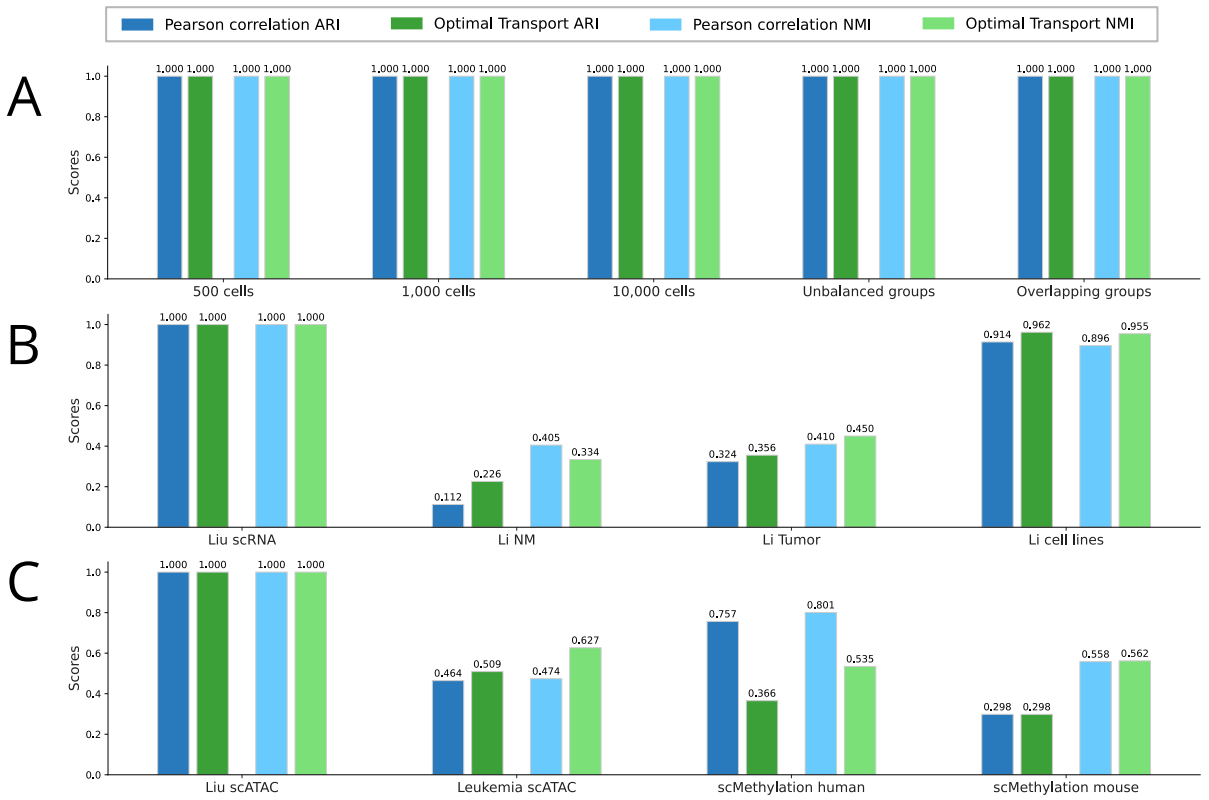
Overall, the clustering conclusions derived with Leiden clustering are confirmed with Louvain clustering.

References

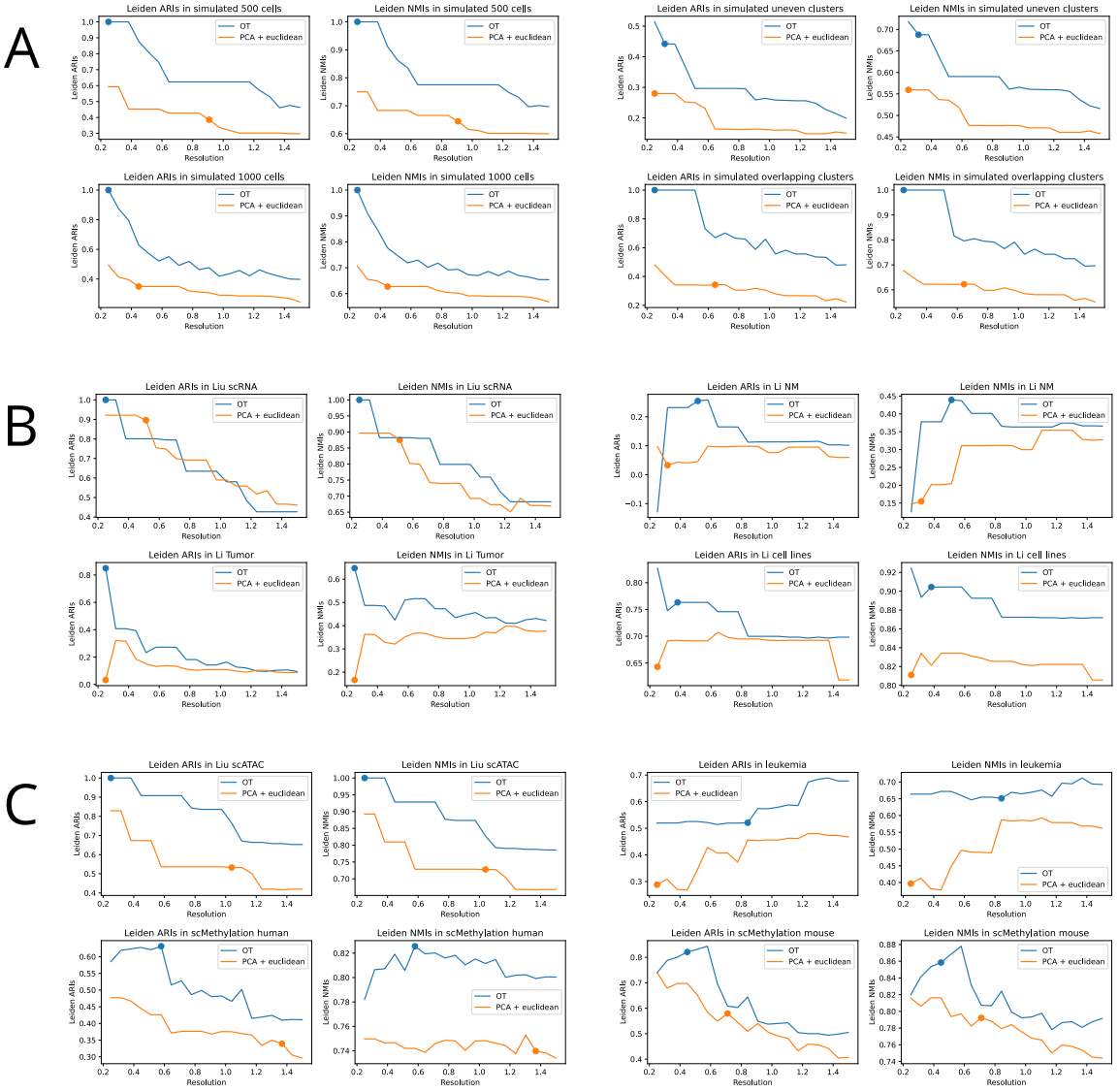
Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, **2008**, P10008.
Lähnemann,D. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome biology*, **21**, 1–35.
Li,H. *et al.* (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet*, **49**, 708–718.
Liu,L. *et al.* (2019) Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat Commun*, **10**, 470.
Raudvere,U. *et al.* (2019) g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, **47**, W191–W198.
Zappia,L. *et al.* (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol*, **18**, 174.



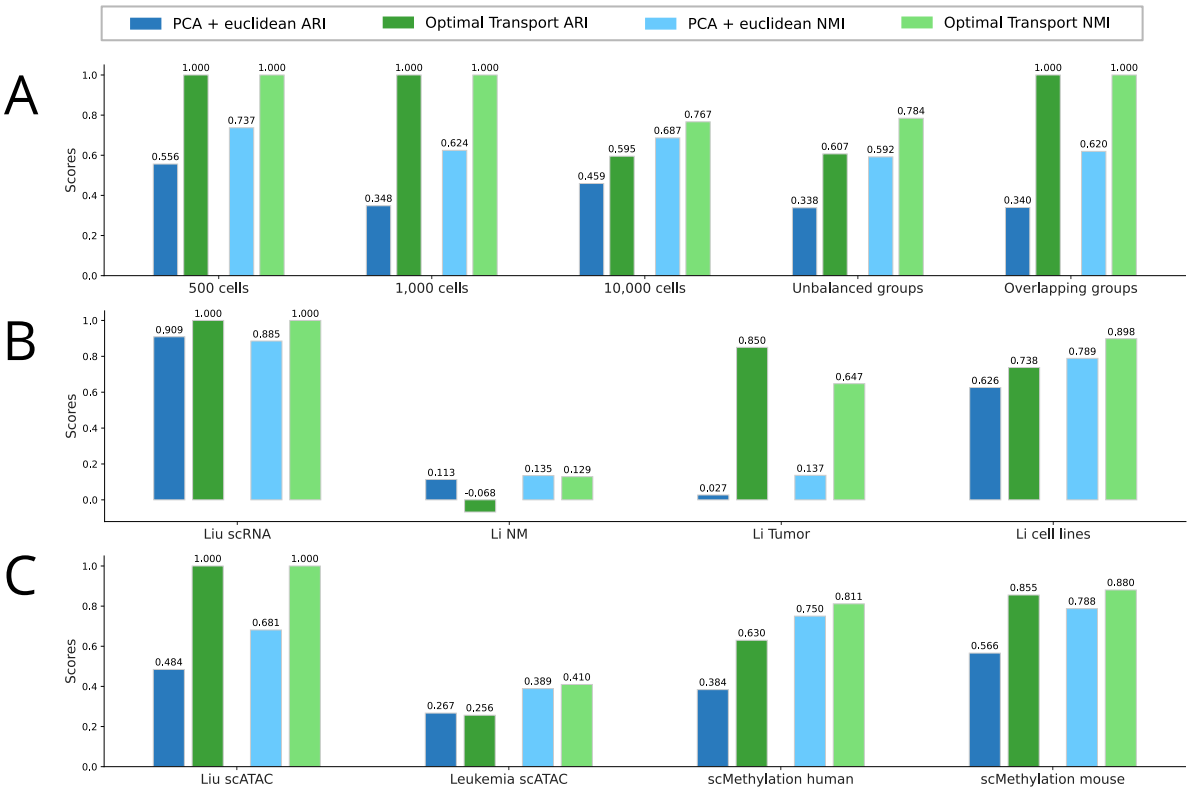
Supplementary Figure 1. Evaluation of the influence of scRNA-seq normalization strategy for baseline distances and for Optimal Transport (OT) on cell-cell similarity inference. The strategies considered are log transformation, per-cell normalization then log transformation, Scran, SCTransform and Linnorm. Barplots for silhouette score are reported for (A) Liu scRNA (B) Li Tumor (C) Li NM (D) Li cell lines.



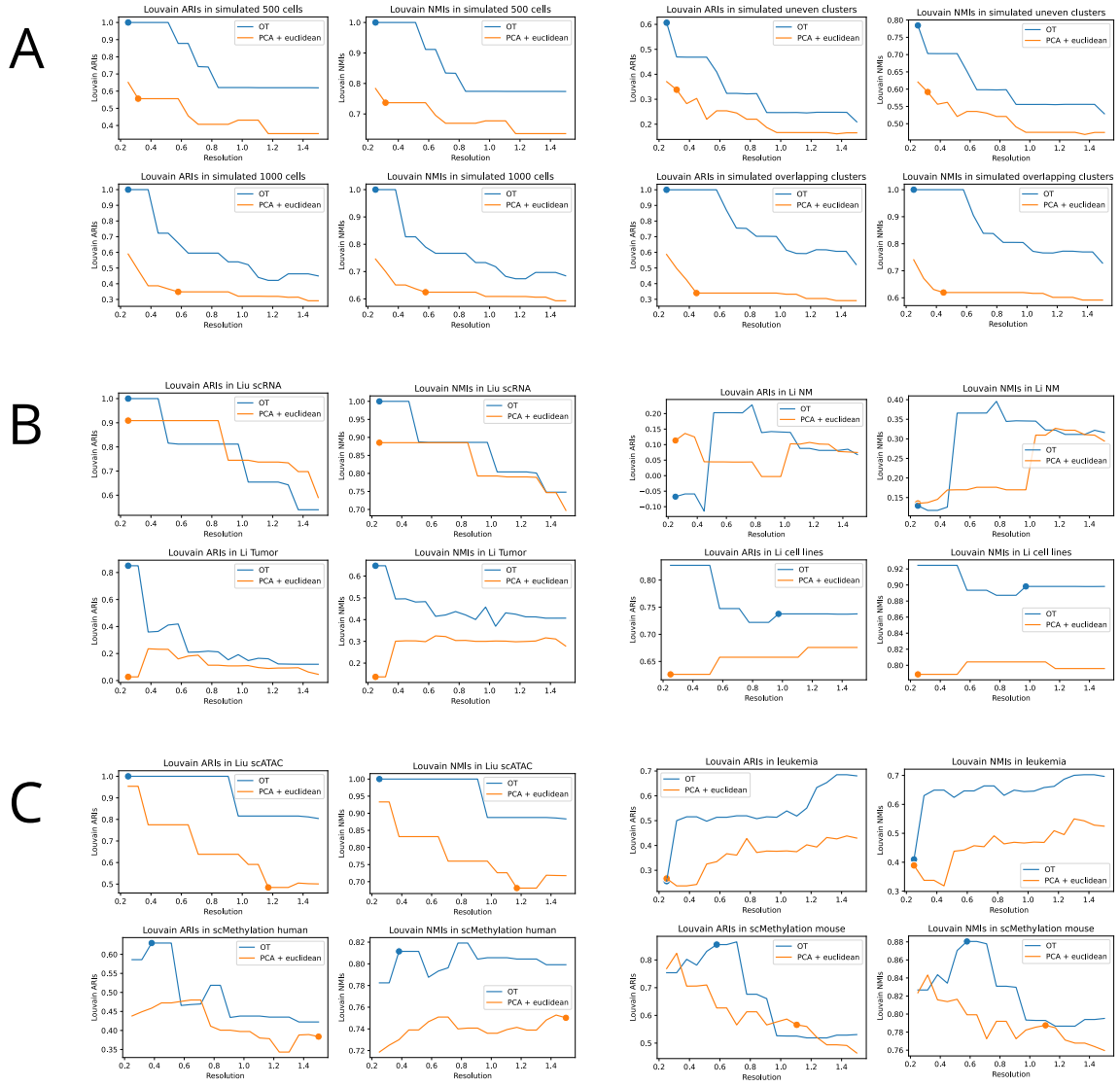
Supplementary Figure 2. Comparison of Optimal Transport (OT) against Pearson correlation in spectral clustering. Barplots for Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) are reported for (A) simulated scRNA-seq data composed of 500 cells, 1,000 and 10,000 cells, with unbalanced groups and overlapping clusters (B) four scRNA-seq datasets (C) two single-cell DNA methylation and two scATAC-seq datasets.



Supplementary Figure 3. Comparison across resolution parameters of a typical single-cell clustering algorithm relying on Leiden clustering after Euclidean distance on PCA components, against Leiden clustering after Optimal Transport (OT) distance. Line plots for Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) are reported for (A) simulated scRNA-seq data composed of 500 cells and 1,000 cells, with unbalanced groups and overlapping clusters (B) four scRNA-seq datasets (C) two single-cell DNA methylation and two scATAC-seq datasets. The dots on each line plot represent the value automatically selected using the Silhouette score.



Supplementary Figure 4. Comparison of a typical single-cell clustering algorithm relying on Louvain clustering after Euclidean distance on PCA components, against Louvain clustering after Optimal Transport (OT) distance. Barplots for Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) are reported for (A) simulated scRNA-seq data composed of 500 cells, 1,000 and 10,000 cells, with unbalanced groups and overlapping clusters (B) four scRNA-seq datasets (C) two single-cell DNA methylation and two scATAC-seq datasets.



Supplementary Figure 5. Comparison across resolution parameters of a typical single-cell clustering algorithm relying on Louvain clustering after Euclidean distance on PCA components, against Louvain clustering after Optimal Transport (OT) distance. Line plots for Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) are reported for (A) simulated scRNA-seq data composed of 500 cells and 1,000 cells, with unbalanced groups and overlapping clusters (B) four scRNA-seq datasets (C) two single-cell DNA methylation and two scATAC-seq datasets. The dots on each line plot represent the value automatically selected using the Silhouette score.

Supplementary Table 1. Silhouette score and C index for baseline metrics (euclidean, cityblock, cosine similarity, Pearson correlation), with and without normalization.

Score	Optional normalization	Metric	Simulated 500 cells	Simulated 1k cells	Simulated uneven	Simulated overlapping	Liu scRNA	Li NM	Li Tumor	Li cell lines	Liu scATAC	Leukemia scATAC	scMethylation human	scMethylation mouse	Liu scRNA (Scran)	Li NM (Scran)	Li Tumor (Scran)	Li cell lines (Scran)	Liu scRNA (SCTransform)	Li NM (SCTransform)	Li Tumor (SCTransform)	Li cell lines (SCTransform)	Liu scRNA (Linnorm)	Li NM (Linnorm)	Li Tumor (Linnorm)	Li cell lines (Linnorm)	Liu scRNA (norm then log)	Li NM (norm then log)	Li Tumor (norm then log)	Li cell lines (norm then log)
Silhouette score	None	cityblock	0.105	0.110	0.110	0.062	0.246	-0.088	-0.077	0.043	0.044	-0.218	-0.082	0.021	0.231	-0.070	-0.079	0.093	0.202	-0.083	-0.083	0.098	0.227	-0.048	-0.016	0.101	0.234	-0.082	-0.082	0.094
		correlation	0.292	0.291	0.291	0.189	0.278	0.053	0.035	0.191	0.198	0.018	0.132	0.100	0.285	0.049	0.036	0.186	0.250	0.052	0.039	0.191	0.326	0.067	0.056	0.182	0.282	0.051	0.037	0.186
		cosine	0.253	0.258	0.253	0.159	0.253	0.051	0.023	0.068	0.243	0.015	0.096	0.087	0.255	0.059	0.029	0.079	0.197	0.054	0.025	0.102	0.308	0.064	0.042	0.106	0.253	0.054	0.024	0.075
		euclidean	0.109	0.112	0.112	0.066	0.182	-0.007	-0.013	0.039	0.047	-0.130	-0.056	0.023	0.162	-0.003	-0.017	0.080	0.094	-0.004	-0.019	0.086	0.175	0.036	0.024	0.086	0.166	-0.002	-0.020	0.085
	Normalized	cityblock	0.137	0.143	0.139	0.083	0.194	0.032	-0.003	0.032	0.134	0.012	0.055	0.053	0.197	0.042	0.000	0.039	0.173	0.037	-0.002	0.055	0.208	0.035	0.012	0.058	0.195	0.036	-0.004	0.037
		correlation	0.292	0.291	0.291	0.189	0.278	0.053	0.035	0.191	0.198	0.018	0.132	0.100	0.285	0.049	0.036	0.186	0.250	0.052	0.039	0.191	0.326	0.067	0.056	0.182	0.282	0.051	0.037	0.186
		cosine	0.253	0.258	0.253	0.159	0.253	0.051	0.023	0.068	0.243	0.015	0.096	0.087	0.255	0.059	0.029	0.079	0.197	0.054	0.025	0.102	0.308	0.064	0.042	0.106	0.253	0.054	0.024	0.075
		euclidean	0.131	0.136	0.133	0.080	0.117	0.020	-0.057	-0.028	0.120	-0.054	0.051	0.047	0.121	0.052	-0.040	-0.013	0.085	0.035	-0.056	0.013	0.155	0.026	-0.050	0.006	0.119	0.034	-0.058	-0.018
	Normalized then scaled	cityblock	0.127	0.131	0.131	0.075	0.191	0.040	0.009	0.044	0.126	-0.020	0.046	0.049	0.190	0.037	0.006	0.046	0.166	0.038	0.009	0.055	0.194	0.045	0.031	0.060	0.190	0.038	0.007	0.045
		correlation	0.248	0.248	0.250	0.157	0.249	0.030	0.018	0.185	0.164	0.011	0.106	0.086	0.258	0.029	0.019	0.178	0.222	0.031	0.020	0.172	0.276	0.032	0.027	0.168	0.256	0.030	0.019	0.179
		cosine	0.212	0.218	0.217	0.130	0.223	0.038	0.022	0.066	0.191	0.011	0.071	0.073	0.224	0.045	0.026	0.074	0.173	0.041	0.022	0.083	0.235	0.038	0.033	0.085	0.223	0.041	0.022	0.071
		euclidean	0.110	0.114	0.114	0.065	0.106	0.044	-0.017	-0.011	0.090	-0.031	0.037	0.039	0.106	0.050	-0.014	-0.008	0.072	0.045	-0.021	0.009	0.127	0.049	0.008	0.002	0.105	0.045	-0.021	-0.010
C index	None	cityblock	0.278	0.278	0.368	0.329	0.061	0.513	0.571	0.179	0.349	0.424	0.166	0.181	0.031	0.602	0.615	0.128	0.038	0.528	0.579	0.120	0.060	0.429	0.460	0.124	0.032	0.526	0.579	0.128
		correlation	0.018	0.016	0.014	0.100	0.013	0.137	0.182	0.092	0.023	0.302	0.078	0.127	0.008	0.149	0.186	0.097	0.013	0.140	0.171	0.082	0.015	0.109	0.161	0.127	0.010	0.143	0.180	0.098
		cosine	0.252	0.252	0.285	0.336	0.125	0.176	0.237	0.198	0.098	0.368	0.138	0.205	0.118	0.166	0.224	0.188	0.157	0.168	0.220	0.187	0.078	0.131	0.198	0.175	0.120	0.171	0.228	0.189
		euclidean	0.253	0.253	0.335	0.313	0.075	0.422	0.498	0.169	0.348	0.457	0.173	0.199	0.048	0.470	0.538	0.161	0.139	0.396	0.483	0.139	0.092	0.187	0.231	0.166	0.038	0.393	0.484	0.153
	Normalized	cityblock	0.303	0.305	0.341	0.372	0.139	0.229	0.287	0.227	0.147	0.408	0.152	0.217	0.130	0.209	0.274	0.207	0.144	0.217	0.270	0.212	0.106	0.196	0.253	0.187	0.131	0.220	0.279	0.208
		correlation	0.018	0.016	0.014	0.100	0.013	0.137	0.182	0.092	0.023	0.302	0.078	0.127	0.008	0.149	0.186	0.097	0.013	0.140	0.171	0.082	0.015	0.109	0.161	0.127	0.010	0.143	0.180	0.098
		cosine	0.252	0.252	0.285	0.336	0.125	0.176	0.237	0.198	0.098	0.368	0.138	0.205	0.118	0.166	0.224	0.188	0.157	0.168	0.220	0.187	0.078	0.131	0.198	0.175	0.120	0.171	0.228	0.189
		euclidean	0.308	0.311	0.353	0.371	0.186	0.289	0.347	0.206	0.163	0.424	0.160	0.227	0.183	0.238	0.306	0.201	0.254	0.264	0.323	0.213	0.142	0.279	0.342	0.189	0.185	0.267	0.327	0.201
	Normalized then scaled	cityblock	0.311	0.313	0.347	0.378	0.132	0.178	0.278	0.231	0.159	0.480	0.164	0.227	0.127	0.200	0.290	0.210	0.153	0.188	0.273	0.216	0.110	0.148	0.210	0.192	0.127	0.190	0.283	0.211
		correlation	0.006	0.003	0.003	0.070	0.008	0.168	0.206	0.070	0.006	0.236	0.122	0.169	0.006	0.173	0.215	0.079	0.021	0.168	0.207	0.078	0.018	0.146	0.140	0.099	0.006	0.170	0.208	0.079
		cosine	0.279	0.279	0.314	0.356	0.127	0.183	0.236	0.206	0.140	0.341	0.154	0.223	0.124	0.164	0.218	0.193	0.183	0.169	0.220	0.199	0.102	0.157	0.203	0.187	0.124	0.173	0.226	0.195
		euclidean	0.327	0.329	0.369	0.385	0.192	0.235	0.318	0.219	0.207	0.462	0.182	0.244	0.191	0.225	0.304	0.211	0.279	0.229	0.304	0.226	0.165	0.208	0.289	0.207	0.191	0.232	0.311	0.212

Supplementary Table 2. Silhouette score and C index for Optimal Transport distance for different ground costs (euclidean, cityblock, cosine similarity, Pearson correlation) and regularization parameters (epsilon=1000, 5, 1, .5, .1, .05, .01)

Score	Ground cost	Epsilon	Simulated 500 cells	Simulated 1k cells	Simulated uneven	Simulated overlapping	Liu scRNA	Li NM	Li Tumor	Li cell lines	Liu scATAC	Leukemia scATAC	scMethylation human	scMethylation mouse	Liu scRNA (Scran)	Li NM (Scran)	Li Tumor (Scran)	Li cell lines (Scran)	Liu scRNA (SCTransform)	Li NM (SCTransform)	Li Tumor (SCTransform)	Li cell lines (SCTransform)	Liu scRNA (Linnorm)	Li NM (Linnorm)	Li Tumor (Linnorm)	Li cell lines (Linnorm)	Liu scRNA (norm then log)	Li NM (norm then log)	Li Tumor (norm then log)	Li cell lines (norm then log)
Silhouette score	cityblock	0.01	0.171	0.178	0.161	0.096	0.315	0.043	0.017	0.146	0.272	0.011	0.094	0.068	0.307	0.046	0.017	0.135	0.255	0.045	0.018	0.146	0.318	0.049	0.033	0.133	0.308	0.044	0.016	0.134
		0.05	0.244	0.248	0.230	0.124	0.398	0.053	0.035	0.282	0.577	-0.002	0.015	0.067	0.394	0.053	0.014	0.248	0.345	0.053	0.035	0.261	0.417	0.058	0.046	0.224	0.393	0.052	0.034	0.248
		0.1	0.600	0.596	0.467	0.354	0.677	0.081	0.043	0.556	0.831	-0.032	0.027	0.133	0.674	0.082	0.044	0.523	0.640	0.080	0.040	0.518	0.698	0.087	0.063	0.500	0.673	0.080	0.041	0.522
		0.5	0.846	0.870	0.745	0.727	0.828	0.095	-0.031	0.666	0.859	0.043	0.304	0.447	0.847	0.037	-0.039	0.707	0.800	0.074	-0.029	0.682	0.831	0.118	0.115	0.717	0.844	0.072	-0.029	0.702
		1	0.850	0.874	0.750	0.734	0.825	0.092	-0.038	0.644	0.857	0.050	0.314	0.457	0.843	0.032	-0.047	0.693	0.795	0.071	-0.036	0.672	0.826	0.116	0.113	0.706	0.840	0.069	-0.036	0.688
		5	0.852	0.876	0.753	0.738	0.822	0.090	-0.042	0.627	0.855	0.052	0.318	0.462	0.840	0.030	-0.052	0.682	0.791	0.068	-0.041	0.664	0.822	0.114	0.111	0.697	0.837	0.067	-0.040	0.676
		1000	0.853	0.877	0.753	0.739	0.821	0.089	-0.043	0.623	0.854	0.053	0.319	0.463	0.840	0.029	-0.053	0.679	0.790	0.068	-0.042	0.661	0.821	0.114	0.110	0.695	0.836	0.066	-0.041	0.673
	correlation	0.01	0.215	0.219	0.180	0.115	0.296	0.034	0.009	0.108	0.254	0.011	0.081	0.065	0.274	0.043	0.007	0.079	0.212	0.038	0.006	0.098	0.298	0.038	0.014	0.092	0.276	0.037	0.005	0.081
		0.05	0.398	0.387	0.324	0.217	0.567	0.038	0.020	0.306	0.563	-0.004	0.115	0.112	0.474	0.052	0.012	0.173	0.336	0.044	0.013	0.220	0.474	0.044	0.012	0.152	0.493	0.043	0.011	0.184
		0.1	0.838	0.845	0.771	0.704	0.807	0.082	0.011	0.647	0.786	-0.033	0.256	0.292	0.799	0.059	-0.014	0.573	0.752	0.073	0.003	0.603	0.748	0.078	0.002	0.554	0.801	0.071	-0.001	0.585
		0.5	0.915	0.937	0.884	0.873	0.793	0.146	0.070	0.678	0.797	0.042	0.340	0.486	0.829	0.022	0.011	0.656	0.801	0.092	0.081	0.681	0.796	0.117	0.095	0.660	0.826	0.091	0.077	0.666
		1	0.917	0.939	0.887	0.877	0.789	0.148	0.073	0.675	0.796	0.042	0.342	0.493	0.829	0.022	0.013	0.657	0.802	0.093	0.085	0.682	0.801	0.119	0.099	0.661	0.825	0.093	0.081	0.666
		5	0.918	0.940	0.889	0.880	0.785	0.150	0.074	0.673	0.794	0.043	0.344	0.497	0.829	0.022	0.014	0.657	0.803	0.094	0.087	0.682	0.804	0.120	0.101	0.661	0.825	0.094	0.083	0.666
		1000	0.919	0.940	0.889	0.880	0.785	0.150	0.074	0.672	0.794	0.043	0.344	0.498	0.829	0.022	0.014	0.657	0.803	0.095	0.088	0.682	0.805	0.120	0.102	0.661	0.825	0.094	0.084	0.666
	cosine	0.01	0.177	0.188	0.173	0.098	0.320	0.045	0.026	0.185	0.314	0.011	0.083	0.071	0.321	0.046	0.023	0.158	0.255	0.046	0.026	0.172	0.333	0.050	0.034	0.148	0.322	0.045	0.024	0.159
		0.05	0.424	0.411	0.346	0.199	0.455	0.058	0.050	0.360	0.750	-0.003	0.009	0.057	0.442	0.055	0.043	0.295	0.371	0.056	0.048	0.308	0.461	0.060	0.048	0.251	0.444	0.056	0.047	0.297
		0.1	0.808	0.813	0.701	0.607	0.769	0.098	0.073	0.668	0.853	-0.032	0.058	0.138	0.761	0.087	0.062	0.618	0.701	0.091	0.069	0.611	0.768	0.094	0.068	0.571	0.762	0.091	0.068	0.621
		0.5	0.896	0.921	0.853	0.827	0.846	0.176	0.074	0.742	0.845	0.041	0.335	0.470	0.858	0.056	0.020	0.735	0.818	0.128	0.081	0.727	0.852	0.164	0.143	0.732	0.858	0.128	0.079	0.739
		1	0.898	0.922	0.855	0.831	0.843	0.175	0.070	0.734	0.842	0.042	0.341	0.479	0.856	0.053	0.015	0.728	0.816	0.126	0.077	0.722	0.851	0.163	0.142	0.726	0.855	0.127	0.076	0.732
		5	0.899	0.923	0.856	0.833	0.842	0.174	0.066	0.729	0.840	0.042	0.344	0.484	0.855	0.051	0.012	0.724	0.814	0.125	0.075	0.718	0.850	0.161	0.141	0.721	0.854	0.125	0.073	0.728
		1000	0.899	0.924	0.857	0.833	0.841	0.174	0.065	0.728	0.840	0.042	0.344	0.484	0.855	0.050	0.011	0.723	0.814	0.124	0.074	0.717	0.850	0.161	0.140	0.720	0.853	0.125	0.073	0.727
	euclidean	0.01	0.167	0.175	0.172	0.095	0.294	0.071	0.046	0.173	0.289	-0.010	0.071	-0.049	0.289	0.059	0.042	0.152	0.254	0.066	0.047	0.164	0.305	0.091	0.095	0.160	0.291	0.067	0.044	0.153
		0.05	0.610	0.601	0.487	0.288	0.453	0.177	0.131	0.433	0.721	-0.059	0.018	0.049	0.477	0.125	0.117	0.372	0.486	0.152	0.126	0.382	0.521	0.210	0.241	0.465	0.481	0.167	0.125	0.375
		0.1	0.829	0.849	0.732	0.651	0.753	0.233	0.147	0.671	0.839	-0.011	0.143	0.202	0.780	0.136	0.130	0.649	0.753	0.195	0.143	0.638	0.783	0.213	0.206	0.686	0.780	0.187	0.145	0.652
		0.5	0.866	0.889	0.794	0.762	0.825	0.204	0.057	0.680	0.847	0.029	0.317	0.437	0.845	0.081	0.017	0.704	0.804	0.158	0.063	0.689	0.825	0.212	0.195	0.702	0.843	0.157	0.062	0.704
		1	0.867	0.891	0.795	0.765	0.824	0.202	0.053	0.673	0.846	0.031	0.321	0.448	0.844	0.078	0.012	0.700	0.802	0.156	0.059	0.686	0.824	0.210	0.193	0.699	0.842	0.155	0.058	0.700
		5	0.868	0.892	0.797	0.767	0.823	0.200	0.050	0.668	0.845	0.031	0.322	0.450	0.843	0.077	0.009	0.696	0.801	0.154	0.056	0.684	0.824	0.209	0.191	0.697	0.841	0.154	0.055	0.696
		1000	0.868	0.892	0.797	0.767	0.823	0.200	0.049	0.667	0.845	0.032	0.323	0.451	0.843	0.076	0.009	0.695	0.801	0.154	0.055	0.683	0.823	0.209	0.191	0.697	0.841	0.153	0.055	0.695
C score	cityblock	0.01	0.241	0.242	0.249	0.348	0.029	0.168	0.221	0.136	0.034	0.373	0.083	0.190	0.028	0.163	0.220	0.134	0.039	0.162	0.211	0.121	0.022	0.140	0.175	0.127	0.029	0.166	0.218	0.135
		0.05	0.138	0.143	0.123	0.288	0.005	0.146	0.178	0.054	0.001	0.394	0.231	0.158	0.004	0.146	0.182	0.063	0.013	0.142	0.165	0.053	0.005	0.132	0.163	0.076	0.004	0.146	0.177	0.063
		0.1	0.017	0.013	0.024	0.107	0.003	0.162	0.206	0.027	0.000	0.411	0.174	0.123	0.001	0.157	0.203	0.027	0.002	0.156	0.195	0.019	0.006	0.148	0.193	0.030	0.001	0.159	0.203	0.027
		0.5	0.006	0.004	0.018	0.040	0.006	0.218	0.322	0.034	0.000	0.264	0.022	0.021	0.001	0.258	0.311	0.023	0.002	0.233	0.321	0.019	0.014	0.157	0.185	0.023	0.002	0.234	0.312	0.024
		1	0.005	0.004	0.018	0.039	0.007	0.226	0.336	0.037	0.000	0.235	0.022	0.019	0.002	0.266	0.324	0.026	0.002	0.242	0.335	0.020	0.015	0.162	0.191	0.025	0.002	0.242	0.325	0.026
		5	0.005	0.004	0.018	0.038	0.008	0.232	0.345	0.040	0.000	0.216	0.022	0.018	0.003	0.272	0.333	0.028	0.003	0.248	0.344	0.022	0.017	0.166	0.196	0.027	0.003	0.248	0.334	0.028
		1000	0.005	0.004	0.018	0.038	0.008	0.234	0.347	0.040	0.000	0.212	0.022	0.018	0.003	0.273	0.335	0.028	0.003	0.249	0.346	0.022	0.017	0.167	0.197	0.028	0.003	0.249	0.336	0.029
	correlation	0.01	0.208	0.211	0.229	0.335	0.067	0.209	0.248	0.170	0.046	0.364	0.100	0.196	0.081	0.197	0.252	0.177	0.093	0.202	0.242	0.172	0.061	0.193	0.244	0.162	0.080	0.206	0.251	0.177
		0.05	0.114	0.124	0.113	0.269	0.019	0.212	0.241	0.078	0.008	0.385	0.048	0.168	0.054	0.197	0.249	0.122	0.066	0.204	0.237	0.109	0.052	0.202	0.257	0.133	0.048	0.209	0.247	0.119
		0.1	0.001	0.001	0.002	0.017	0.002	0.181	0.226	0.027	0.000	0.398	0.023	0.045	0.009	0.201	0.254	0.043	0.002	0.187	0.232	0.021	0.010	0.192	0.261	0.049	0.006	0.191	0.241	0.041
		0.5	0.000	0.000	0.001	0.001	0.004	0.151	0.172	0.025	0.000	0.224	0.021	0.018	0.007	0.226	0.235	0.033	0.002	0.184	0.182	0.014	0.022	0.160	0.186	0.034	0.005	0.186	0.190	0.030
		1	0.000	0.000	0.001	0.001	0.004	0.151	0.172	0.025	0.000	0.206	0.021	0.018	0.007	0.228	0.235	0.032	0.002	0.184	0.181	0.014	0.021	0.160	0.184	0.034	0.005	0.187	0.189	0.030
		5	0.000	0.000	0.001	0.001	0.005	0.152	0.172	0.025	0.000	0.198	0.021	0.017	0.007	0.229	0.235	0.032	0.002	0.185	0.180	0.014	0.020	0.159	0.182	0.034	0.005	0.187	0.188	0.030
		1000	0.000	0.000	0.001	0.001	0.005	0.152	0.172	0.025	0.000	0.196	0.021	0.017	0.007	0.229	0.235	0.032	0.002	0.185	0.180	0.014	0.020	0.159	0.182	0.034	0.005	0.187	0.188	0.030
	cosine	0.01	0.208	0.204	0.191	0.332	0.011	0.150	0.178	0.094	0.018	0.358	0.081	0.190	0.014	0.151	0.192	0.111	0.030	0.148	0.172	0.092	0.010	0.139	0.169	0.112	0.014	0.152	0.183	0.110
		0.05	0.022	0.021	0.018	0.155	0.001	0.128	0.142	0.030	0.000	0.377	0.237	0.156	0.002	0.133	0.160	0.046	0.008	0.127	0.134	0.034	0.003	0.128	0.158	0.063	0.002	0.131	0.149	0.045
		0.1	0.001	0.001	0.003	0.015	0.000	0.134	0.159	0.014	0.000	0.396	0.100	0.122	0.000	0.144	0.178	0.018	0.000	0.137	0.156	0.010	0.004	0.141	0.185	0.022	0.000	0.140	0.166	0.018
		0.5	0.000	0.000	0.002	0.005	0.000	0.133	0.170	0.016	0.000	0.229	0.019	0.022	0.000	0.202	0.208	0.018	0.000	0.157	0.173	0.010	0.007	0.125	0.150	0.017	0.000	0.158	0.174	0.017
		1	0.000	0.000	0.002	0.005	0.001	0.138	0.178	0.016	0.000	0.210	0.019	0.020	0.000	0.207	0.216	0.019	0.000	0.161	0.180	0.010	0.008	0.126	0.153	0.019	0.000	0.162	0.181	0.018
		5	0.000	0.000	0.002	0.005	0.001	0.141	0.183	0.017	0.000	0.202	0.019	0.019	0.000	0.211	0.222	0.020	0.001	0.165	0.186	0.010	0.008	0.128	0.155	0.020	0.000	0.166	0.186	0.018
		1000	0.000	0.000	0.002	0.005	0.001	0.142	0.185	0.017	0.000	0.200	0.019	0.019	0.000	0.211	0.223	0.020	0.001	0.166	0.187	0.011	0.008	0.128	0.156	0.020	0.000	0.166	0.187	0.019
	euclidean	0.01	0.236	0.235	0.224	0.345	0.018	0.094	0.111	0.105	0.026	0.355	0.111	0.166	0.017	0.109	0.134	0.116	0.029	0.100	0.109	0.097	0.010	0.094	0.092	0.102	0.017	0.101	0.122	0.115
		0.05	0.009	0.007	0.011	0.107	0.002	0.077	0.091	0.036	0.000	0.364	0.222	0.140	0.004	0.098	0.113	0.042	0.006	0.086	0.089	0.032	0.003	0.119	0.120	0.043	0.003	0.088	0.102	0.042
		0.1	0.004	0.002	0.009	0.032	0.002	0.080	0.113	0.025	0.000	0.305	0.057	0.099	0.001	0.113	0.131	0.022	0.001	0.090	0.113	0.015	0.008	0.109	0.123	0.021	0.000	0.099	0.119	0.021
		0.5	0.003	0.002	0.010	0.024	0.004	0.128	0.210	0.028	0.000	0.245	0.022	0.033	0.001	0.179	0.223	0.024	0.001	0.142	0.206	0.016	0.014	0.101	0.127	0.025	0.001	0.143	0.201	0.023
		1	0.003	0.002	0.010	0.024	0.004	0.132	0.217	0.029	0.000	0.241	0.022	0.024	0.001	0.184	0.229	0.025	0.001	0.145	0.213	0.016	0.014	0.101	0.129	0.025	0.001	0.146	0.207	0.024
		5	0.003	0.002	0.010	0.023	0.004	0.135	0.221	0.030	0.000	0.239	0.022	0.023	0.001	0.187	0.234	0.025	0.001	0.148	0.218	0.016	0.014	0.102	0.130	0.026	0.001	0.148	0.212	0.024
		1000	0.003	0.002	0.010	0.023	0.004	0.136	0.223	0.030	0.000	0.238	0.022	0.023	0.001	0.187	0.235	0.025	0.001	0.149	0.219	0.017	0.014	0.102	0.131	0.026	0.001	0.149	0.213	0.025

Supplementary Table 3. Comparison of computation time and silhouette score for entropic regularized OT and classical OT, for varying levels of dropout noise on the Liu scRNA dataset (restricted to 500 genes).

Proportion of zeros	Regularized OT execution time on GPU (s)	Classical OT execution time on CPU (s)	Regularized OT silhouette score	Classical OT silhouette score
0.41	3.92	889.23	0.58	0.76
0.59	0.07	828.32	0.57	0.70
0.71	0.07	792.10	0.54	0.57
0.80	0.07	800.69	0.50	0.44
0.86	0.07	809.04	0.46	0.33

Supplementary Table 4. Run times for the computation of OT distance matrices with epsilon=.5 on the different datasets. Computed using a Tesla V100-SXM2-16GB GPU on a cloud provider.

Dataset	Simulated 500 cells	Simulated 1k cells	Simulated uneven groups	Simulated overlapping groups	Liu scRNA	Li tumor	Li NM	Li cell lines	Liu scATAC	scMethylation mouse	scMethylation human	Leukemia scATAC
Run time	5s	22s	22s	21s	4s	11s	6s	25s	4s	15mn16s	9mn57s	7s
Number of batches	50	50	50	50	50	50	50	50	50	200	200	50

Supplementary Table 5. Comparison of clustering performances for Optimal Transport distance and baseline distances.

For hierarchical clustering and spectral clustering, OT is compared to Pearson correlation. The results obtained by optimizing the number of clusters are here compared to those obtained fixing it to the number of ground truth labels. For hierarchical clustering the results with complete linkage are compared to the results with average linkage. For Leiden and Louvain clustering, OT is compared to the euclidean distance on PCA components, according to the typical single-cell clustering workflow.

				Simulated 500 cells	Simulated 1k cells	Simulated uneven	Simulated overlapping	Liu scRNA	Li NM	Li Tumor	Li cell lines	Liu scATAC	Leukemia scATAC	scMethylation human	scMethylation mouse
Hierarchical clustering	Pearson	Fixed number of clusters (complete linkage)	number of clusters	3	3	3	3	3	7	7	7	3	6	21	16
			ARI	1.00	1.00	1.00	1.00	1.00	0.39	0.26	0.48	0.87	0.26	0.68	0.36
			NMI	1.00	1.00	1.00	1.00	1.00	0.49	0.42	0.68	0.87	0.29	0.74	0.65
			AMI	1.00	1.00	1.00	1.00	1.00	0.46	0.40	0.68	0.87	0.28	0.73	0.64
		Silhouette derived number of clusters (complete linkage)	number of clusters	3	3	3	3	3	23	6	10	3	25	7	3
			ARI	1.00	1.00	1.00	1.00	1.00	0.13	0.33	0.77	0.87	0.41	0.21	0.09
			NMI	1.00	1.00	1.00	1.00	1.00	0.44	0.45	0.85	0.87	0.46	0.39	0.30
			AMI	1.00	1.00	1.00	1.00	1.00	0.38	0.43	0.85	0.87	0.42	0.38	0.30
		Fixed number of clusters (average linkage)	number of clusters	3	3	3	3	3	7	7	7	3	6	21	16
			ARI	1.00	1.00	0.94	0.00	1.00	0.40	0.11	0.00	0.64	0.01	0.21	0.00
			NMI	1.00	1.00	0.89	0.00	1.00	0.38	0.12	0.02	0.75	0.05	0.39	0.01
			AMI	1.00	1.00	0.89	0.00	1.00	0.35	0.09	0.00	0.75	0.02	0.38	0.00
		Silhouette derived number of clusters (average linkage)	number of clusters	3	3	4	3	3	3	3	3	5	3	3	3
			ARI	1.00	1.00	1.00	0.00	1.00	0.24	0.07	0.00	0.98	0.00	0.00	0.00
			NMI	1.00	1.00	1.00	0.00	1.00	0.19	0.06	0.01	0.98	0.01	0.00	0.00
			AMI	1.00	1.00	1.00	0.00	1.00	0.17	0.05	0.00	0.97	0.00	0.00	0.00
	OT (eps = .5)	Ground cost		correlation	correlation	correlation	correlation	cosine	cosine	cosine	cosine	cosine	cosine	correlation	correlation
		Fixed number of clusters (complete linkage)	number of clusters	3	3	3	3	3	7	7	7	3	6	21	16
			ARI	1.00	1.00	1.00	1.00	1.00	0.20	0.36	0.89	1.00	0.00	0.41	0.63
			NMI	1.00	1.00	1.00	1.00	1.00	0.37	0.52	0.90	1.00	0.06	0.65	0.74
			AMI	1.00	1.00	1.00	1.00	1.00	0.33	0.50	0.90	1.00	0.03	0.64	0.74
		Silhouette derived number of clusters (complete linkage)	number of clusters	3	3	3	3	3	9	3	8	3	3	3	3
			ARI	1.00	1.00	1.00	1.00	1.00	0.29	0.46	0.90	1.00	0.00	0.40	0.09
			NMI	1.00	1.00	1.00	1.00	1.00	0.44	0.38	0.92	1.00	0.01	0.51	0.30
			AMI	1.00	1.00	1.00	1.00	1.00	0.40	0.37	0.92	1.00	0.00	0.51	0.30
		Fixed number of clusters (average linkage)	number of clusters	3	3	3	3	3	7	7	7	3	6	21	16
			ARI	1.00	1.00	1.00	1.00	1.00	0.42	0.20	0.80	1.00	0.00	0.61	0.30
			NMI	1.00	1.00	1.00	1.00	1.00	0.35	0.24	0.87	1.00	0.03	0.70	0.56
			AMI	1.00	1.00	1.00	1.00	1.00	0.32	0.21	0.87	1.00	0.00	0.70	0.56
		Silhouette derived number of clusters (average linkage)	number of clusters	3	3	3	3	3	3	3	8	3	3	3	3
			ARI	1.00	1.00	1.00	1.00	1.00	0.26	0.02	0.93	1.00	0.00	0.20	0.00
			NMI	1.00	1.00	1.00	1.00	1.00	0.18	0.03	0.94	1.00	0.01	0.39	0.00
			AMI	1.00	1.00	1.00	1.00	1.00	0.16	0.02	0.94	1.00	0.00	0.39	0.00
Spectral clustering	Pearson	Fixed number of clusters	number of clusters	3	3	3	3	3	7	7	7	3	6	21	16
			ARI	1.00	1.00	1.00	1.00	1.00	0.16	0.13	0.91	1.00	0.63	0.49	0.70
			NMI	1.00	1.00	1.00	1.00	1.00	0.34	0.37	0.90	1.00	0.69	0.81	0.80
			AMI	1.00	1.00	1.00	1.00	1.00	0.30	0.34	0.89	1.00	0.68	0.80	0.79
		Silhouette derived number of clusters	number of clusters	3	3	3	3	3	13	3	7	3	19	10	3
			ARI	1.00	1.00	1.00	1.00	1.00	0.11	0.32	0.91	1.00	0.46	0.76	0.30
			NMI	1.00	1.00	1.00	1.00	1.00	0.41	0.41	0.90	1.00	0.47	0.80	0.56
			AMI	1.00	1.00	1.00	1.00	1.00	0.36	0.40	0.89	1.00	0.44	0.80	0.56
	OT (eps = .5)	Ground cost		correlation	correlation	correlation	correlation	cosine	cosine	cosine	cosine	cosine	cosine	correlation	correlation
		Fixed number of clusters	number of clusters	3	3	3	3	3	7	7	7	3	6	21	16
			ARI	1.00	1.00	1.00	1.00	1.00	0.22	0.19	0.96	1.00	0.69	0.48	0.69
			NMI	1.00	1.00	1.00	1.00	1.00	0.36	0.45	0.95	1.00	0.76	0.81	0.81
			AMI	1.00	1.00	1.00	1.00	1.00	0.32	0.43	0.95	1.00	0.75	0.80	0.80
		Silhouette derived number of clusters	number of clusters	3	3	3	3	3	6	3	7	3	7	3	3
			ARI	1.00	1.00	1.00	1.00	1.00	0.23	0.36	0.96	1.00	0.51	0.37	0.30
			NMI	1.00	1.00	1.00	1.00	1.00	0.33	0.45	0.95	1.00	0.63	0.53	0.56
			AMI	1.00	1.00	1.00	1.00	1.00	0.30	0.44	0.95	1.00	0.62	0.53	0.56
Leiden clustering	PCA + euclidean	Silhouette derived number of clusters	resolution	0.91	0.45	0.25	0.64	0.51	0.32	0.25	0.25	1.04	0.25	1.37	0.71
			number of clusters	11	12	8	12	3	4	2	7	7	3	32	18
			ARI	0.39	0.35	0.28	0.34	0.90	0.03	0.03	0.64	0.53	0.29	0.34	0.58
			NMI	0.64	0.63	0.56	0.62	0.87	0.15	0.17	0.81	0.73	0.40	0.74	0.79
			AMI	0.64	0.62	0.56	0.62	0.87	0.12	0.16	0.81	0.72	0.39	0.73	0.79
	OT (eps = .5)	Ground cost		correlation	correlation	correlation	correlation	cosine	cosine	cosine	cosine	cosine	cosine	correlation	correlation
		Silhouette derived number of clusters	resolution	0.25	0.25	0.32	0.25	0.25	0.51	0.25	0.38	0.25	0.84	0.25	0.38
			number of clusters	3	3	5	3	3	6	2	9	3	4	16	14
			ARI	1.00	1.00	0.44	1.00	1.00	0.25	0.85	0.76	1.00	0.52	0.64	0.85
			NMI	1.00	1.00	0.69	1.00	1.00	0.44	0.65	0.90	1.00	0.65	0.82	0.86
			AMI	1.00	1.00	0.69	1.00	1.00	0.41	0.64	0.90	1.00	0.65	0.82	0.86
Louvain clustering	PCA + euclidean	Silhouette derived number of clusters	resolution	0.91	0.45	0.25	0.64	0.51	0.32	0.25	0.25	1.04	0.25	1.37	0.71
			number of clusters	11	12	8	12	3	4	2	7	7	3	32	18
			ARI	0.39	0.35	0.28	0.34	0.90	0.03	0.03	0.64	0.53	0.29	0.34	0.58
			NMI	0.64	0.63	0.56	0.62	0.87	0.15	0.17	0.81	0.73	0.40	0.74	0.79
			AMI	0.64	0.62	0.56	0.62	0.87	0.12	0.16	0.81	0.72	0.39	0.73	0.79
	OT (eps = .5)	Ground cost		correlation	correlation	correlation	correlation	cosine	cosine	cosine	cosine	cosine	cosine	correlation	correlation
		Silhouette derived number of clusters	resolution	0.25	0.25	0.32	0.25	0.25	0.51	0.25	0.38	0.25	0.84	0.25	0.38
			number of clusters	3	3	5	3	3	6	2	9	3	4	16	14
			ARI	1.00	1.00	0.44	1.00	1.00	0.25	0.85	0.76	1.00	0.52	0.64	0.85
			NMI	1.00	1.00	0.69	1.00	1.00	0.44	0.65	0.90	1.00	0.65	0.82	0.86
			AMI	1.00	1.00	0.69	1.00	1.00	0.41	0.64	0.90	1.00	0.65	0.82	0.86

Appendix B

Unsupervised ground metric learning with Wasserstein Singular Sectors

B.1 Computation Times and Numerical Resources

CPU computations were performed on a Dell Latitude 5420 with an 8 core 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz CPU. GPU computations were performed on Nvidia V100 SXM2 32 Go GPUs.

50 pairs of Wasserstein power iterations for the synthetic datasets in Section 3.2.3 run in about three minutes on CPU.

15 pairs of Sinkhorn ($\varepsilon = 0.1$) power iterations for the single-cell dataset in Section 3.5 run in about 1h50mn on GPU.

B.2 Proof of Theorem 3.2.3 (Existence of Singular Vectors)

We consider the set

$$\mathcal{K}_n^\rho \stackrel{\text{def.}}{=} \{\mathbf{D} \in \mathcal{D}_n \text{ s.t. } \|\mathbf{D}\|_\infty = 1, \forall i \neq j, \mathbf{D}_{i,j} \geq \rho\}$$

Let $(\mathbf{C}, \mathbf{D}) \in \mathcal{K}_m^\rho \times \mathcal{K}_n^\rho$. A classical result states that

$$\frac{\rho}{2}\|\mathbf{a} - \mathbf{a}'\|_1 \leq \mathrm{W}_\mathbf{C}(\mathbf{a}, \mathbf{a}') \leq \tfrac{1}{2}\|\mathbf{a} - \mathbf{a}'\|_1$$

Thus with $(\mathbf{C}', \mathbf{D}') = \left(\dfrac{\Phi_\mathbf{B}(\mathbf{D})}{\|\Phi_\mathbf{B}(\mathbf{D})\|_\infty}, \dfrac{\Phi_\mathbf{A}(\mathbf{C})}{\|\Phi_\mathbf{A}(\mathbf{C})\|_\infty}\right)$,

$$\frac{\rho}{2}\|\mathbf{b}_k - \mathbf{b}_l\|_1 + \tau R(\mathbf{b}_k - \mathbf{b}_l) \leq \Phi_\mathbf{B}(\mathbf{D})_{k,l} \leq \tfrac{1}{2}\|\mathbf{b}_k - \mathbf{b}_l\|_1 + \tau R(\mathbf{b}_k - \mathbf{b}_l)$$

and

$$\frac{\rho}{2}\|\mathbf{a}_i - \mathbf{a}_j\|_1 + \tau R(\mathbf{a}_i - \mathbf{a}_j) \leq \Phi_\mathbf{A}(\mathbf{C})_{i,j} \leq \tfrac{1}{2}\|\mathbf{a}_i - \mathbf{a}_j\|_1 + \tau R(\mathbf{a}_i - \mathbf{a}_j).$$

Equivalence of norms implies $k_R^- \|\cdot\|_1 \leq R(\cdot) \leq k_R^+ \|\cdot\|_1$. With $\gamma_\mathbf{A} \stackrel{\text{def.}}{=} \dfrac{\min_{i\neq j}\|\mathbf{a}_i - \mathbf{a}_j\|_1}{\max_{i\neq j}\|\mathbf{a}_i - \mathbf{a}_j\|_1}$ and $\gamma_\mathbf{B} \stackrel{\text{def.}}{=}$

$$\frac{\min_{k\neq l}\|\mathbf{b}_k-\mathbf{b}_l\|_1}{\max_{k\neq l}\|\mathbf{b}_k-\mathbf{b}_l\|_1},$$

$$\forall k\neq l, C'_{k,l}\geq \frac{\rho+2\tau k_R^-}{1+2\tau k_R^+}\gamma_{\mathbf{B}} \quad \text{and} \quad \forall i\neq j, D'_{i,j}\geq \frac{\rho+2\tau k_R^-}{1+2\tau k_R^+}\gamma_{\mathbf{A}}.$$

This shows that for

$$0<\rho\leq \min\left(\frac{2\gamma_{\mathbf{A}}\tau k_R^-}{2\tau k_R^+ +1-\gamma_{\mathbf{A}}}, \frac{2\gamma_{\mathbf{B}}\tau k_R^-}{2\tau k_R^+ +1-\gamma_{\mathbf{B}}}\right)$$

one has $(\mathbf{C}',\mathbf{D}')\in \mathcal{K}_m^\rho\times\mathcal{K}_n^\rho$. So for such ρ, the function $(\mathbf{C}',\mathbf{D}')\mapsto\left(\frac{\Phi_{\mathbf{B}}(\mathbf{D})}{\|\Phi_{\mathbf{B}}(\mathbf{D})\|_\infty}, \frac{\Phi_{\mathbf{A}}(\mathbf{C})}{\|\Phi_{\mathbf{A}}(\mathbf{C})\|_\infty}\right)$ is a continuous map from the locally contractible compact $\mathcal{K}_m^\rho\times\mathcal{K}_n^\rho$ to itself so using the Brouwer theorem, it has a fixed point, which is a pair of singular vectors of $(\Phi_{\mathbf{A}},\Phi_{\mathbf{B}})$.

B.3 Proof of Proposition 3.2.6 (Uniqueness of Singular Vectors for Large τ)

Denoting $s=1/\tau$, we consider the map

$$\Psi_{\mathbf{A}}(\mathbf{C})\overset{\text{def.}}{=}\frac{\|\mathbf{C}\|_\infty\mathbf{U}+s\mathbf{W}}{\left\|\|\mathbf{C}\|_\infty\mathbf{U}+s\mathbf{W}\right\|_\infty}$$

on the set of $\|\mathbf{C}\|_\infty=1$, where $\mathbf{U}\overset{\text{def.}}{=}\left(R\left(\mathbf{a}_i-\mathbf{a}_j\right)\right)_{i,j}$ is constant and $\mathbf{W}\overset{\text{def.}}{=}\left(\mathrm{W}_{\mathbf{C}}\left(\mathbf{a}_i,\mathbf{a}_j\right)\right)_{i,j}$. Since $\mathrm{W}_{\mathbf{C}}$ is 1-Lipschitz, and $\|\mathbf{C}\|_\infty=1$, one needs to study the contractance of

$$\mathbf{W}\mapsto\frac{\mathbf{U}+s\mathbf{W}}{\|\mathbf{U}+s\mathbf{W}\|_\infty}.$$

One can explicitly compute the derivative of this map, which is $O(s)$, so that $\Psi_{\mathbf{A}}$ is itself contractant for s small enough. This shows that $(\mathbf{C},\mathbf{D})\mapsto(\Psi_{\mathbf{B}}(\mathbf{D}),\Psi_{\mathbf{A}}(\mathbf{C}))$ is also contractant, which implies uniqueness of the singular vector and linear convergence for $\|\cdot\|_\infty$ of the power iterations.

B.4 Additional Figure for the Numerical Illustration of Section 3.2.3

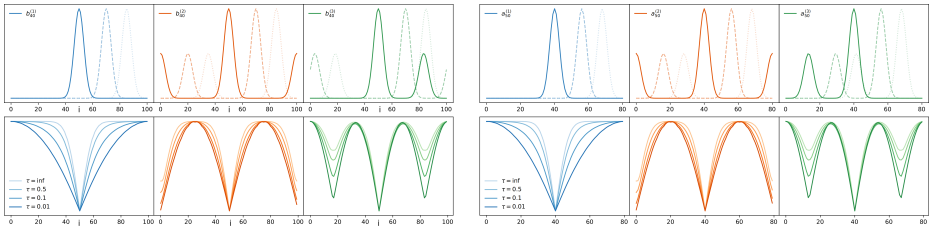
Figure B.1 shows that as predicted, the singular vectors $\mathbf{C}_p$ and $\mathbf{D}_p$ are identical up to rescaling.

B.5 Proof of Convergence for the Stochastic Power Iterations of Section 3.3.1

Proof. We follow steps similar to the proof of [Nemirovski et al., 2009] for projected stochastic gradient descent. The theorem supposes τ large enough, so we can consider a pair of unique Wasserstein Singular Vectors $(\mathbf{C}^\star,\mathbf{D}^\star)$. We study the quantity $\mathbb{E}[u_t]$ where $u_t\overset{\text{def.}}{=}\|\mathbf{C}_t-\mathbf{C}^\star\|_2^2$. The proof for $\|\mathbf{D}_t-\mathbf{D}^\star\|_2^2$ is identical. We consider a constant scaling factor $\tilde{\lambda}$. Let us start by defining

$$(\mathbf{T}_t)_{k,l}=\begin{cases}(\mathbf{C}_t)_{k,l}-\Phi_{\mathbf{B}}(\mathbf{D}_t)_{k,l}/\tilde{\lambda} \text{ if } (k,l)=(k_t,l_t)\\ 0 \text{ otherwise}\end{cases}$$

Figure B.1: Additional Figure for the Numerical Illustration of Section 3.2.3



(left, top) Translated histograms template defining $\mathbf{B}_p$; **(left, bottom)** One element of the singular vector $\mathbf{C}_p$; **(right, top)** Translated histograms template defining $\mathbf{A}_p$; **(right, bottom)** One element of the singular vector $\mathbf{D}_p$.

and $\mathbf{T}^\star \overset{\text{def.}}{=} p(\mathbf{C}^\star - \Phi_{\mathbf{B}}(\mathbf{D}^\star)/\tilde{\lambda})$ where $p = \frac{1}{m^2}$ is the probability for an element to be updated.

By definition of the power iterations, $\mathbf{C}_{t+1} = \Pi(\mathbf{C}_t - \alpha_t \mathbf{T}_t)$.

By definition of the Wasserstein Singular Vectors, $\mathbf{C}^\star = \Pi(\mathbf{C}^\star - \alpha_t \mathbf{T}^\star)$.

Thus,

$$\|\mathbf{C}_{t+1} - \mathbf{C}^\star\|_2^2 = \|\Pi(\mathbf{C}_t - \alpha_t \mathbf{T}_t) - \Pi(\mathbf{C}^\star - \alpha_t \mathbf{T}^\star)\|_2^2.$$

The value of $\tilde{\lambda}$ proposed in the theorem ensures that $\|\mathbf{C}_t - \alpha_t \mathbf{T}_t\|_\infty \geq 1$ and $\|\mathbf{C}^\star - \alpha_t \mathbf{T}^\star\|_\infty \geq 1$.

The theorem of projection on a convex (the unit sphere for the norm $\|\cdot\|_\infty$) then ensures that

$$\|\mathbf{C}_{t+1} - \mathbf{C}^\star\|_2^2 \leq \|(\mathbf{C}_t - \mathbf{C}^\star) - \alpha_t(\mathbf{T}_t - \mathbf{T}^\star)\|_2^2.$$

Decomposing the squared norm, we get

$$\|(\mathbf{C}_t - \mathbf{C}^\star) - \alpha_t(\mathbf{T}_t - \mathbf{T}^\star)\|_2^2 = u_t - 2\alpha_t \langle \mathbf{C}_t - \mathbf{C}^\star, \mathbf{T}_t - \mathbf{T}^\star \rangle + \alpha_t^2 \|\mathbf{T}_t - \mathbf{T}^\star\|_2^2.$$

The middle term can be simplified when taking its expectation:

$$\mathbb{E}_t[\langle \mathbf{C}_t - \mathbf{C}^\star, \mathbf{T}_t - \mathbf{T}^\star \rangle] = pu_t - \frac{p}{\tilde{\lambda}} \langle \mathbf{C}_t - \mathbf{C}^\star, \Phi_{\mathbf{B}}(\mathbf{D}_t) - \Phi_{\mathbf{B}}(\mathbf{D}^\star) \rangle$$

Since the regularization terms in $\Phi_{\mathbf{B}}$ cancel out,

$$\mathbb{E}_t[\langle \mathbf{C}_t - \mathbf{C}^\star, \mathbf{T}_t - \mathbf{T}^\star \rangle] = pu_t - \frac{p}{\tilde{\lambda}} \langle \mathbf{C}_t - \mathbf{C}^\star, \mathrm{W}_{\mathbf{B}}(\mathbf{D}_t) - \mathrm{W}_{\mathbf{B}}(\mathbf{D}^\star) \rangle \geq (1 - L/\tilde{\lambda}) pu_t,$$

for some constant L, since $\mathrm{W}_{\mathbf{B}}$ is 1-Lipschitz with regard to the infinite norm as proved earlier. The last term can be bounded as well:

$$\|\mathbf{T}_t - \mathbf{T}^\star\|_2^2 \leq \|\mathbf{T}_t\|_2^2 + \|\mathbf{T}^\star\|_2^2 \leq M.$$

We have finally

$$\mathbb{E}_t[u_{t+1}] \leq \left(1 - 2\alpha_t p\left(1 - L/\tilde{\lambda}\right)\right) \times u_t + \alpha_t^2 M.$$

In the next steps we assume τ big enough for $p\left(1 - L/\tilde{\lambda}\right)$ to be positive and name the quantity Q.

Let us note that an overly pessimistic upper bound for L is m, which would require a very large value of τ. However, this upper-bound of L is obtained by juggling between different norms. In practice, the algorithm converges for arbitrarily small values of τ. This suggests a much smaller constant, that does not depend on the data's dimensionality.

Taking the expectation over all times t,

$$\mathbb{E}[u_{t+1}] \leq (1 - 2\alpha_t Q) \times \mathbb{E}[u_t] + \alpha_t^2 M.$$

Reformulating,

$$2\alpha_t Q \mathbb{E}[u_t] \leq \mathbb{E}[u_t] - \mathbb{E}[u_{t+1}] + \alpha_t^2 M.$$

Summing along $t = 1...T$,

$$\min_{t=1...T} \mathbb{E}[u_t] \leq \left(\sum_{t=1}^{T} \alpha_t\right)^{-1} \left(\frac{u_1}{2Q} + \frac{M}{2Q}\sum_{t=1}^{T} \alpha_t^2\right).$$

For $\alpha_t = 1/\sqrt{t}$, we thus have classically a convergence rate of $\mathcal{O}(\log(t)/\sqrt{t})$ □

B.6 Sinkhorn Algorithm

The Sinkhorn cost can be computed by the dual formula

$$\mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}_i, \mathbf{a}_j) = \varepsilon\left(\langle \log(\mathbf{u}), \mathbf{a}_i\rangle + \langle \log(\mathbf{v}), \mathbf{a}_j\rangle - \langle \mathbf{K}\mathbf{v}, \mathbf{u}\rangle\right),$$

where $\mathbf{K} \overset{\text{def.}}{=} \exp(-\frac{\mathbf{C}}{\varepsilon})$ and $(\mathbf{u}, \mathbf{v})$ are obtained by iterating the following Sinkhorn fixed point

$$\mathbf{u} \leftarrow \frac{\mathbf{a}_i}{\mathbf{K}^\top \mathbf{v}} \quad \text{and} \quad \mathbf{v} \leftarrow \frac{\mathbf{a}_j}{\mathbf{K}\mathbf{u}}.$$

This allows one to compute with precision ε the n^2 entries of $\Phi_{\mathbf{A}}^{\varepsilon}(\mathbf{C})$ in $\mathcal{O}((mn)^2/\varepsilon^2)$ operations [Altschuler et al., 2017], using a parallelizable algorithm that is well suited for GPU computations.

B.7 Proof of Proposition 3.4.4 (Connection with PCA)

Proof. We define the operator mapping correlation kernels to Euclidean distance

$$\Delta(\mathbf{K}) \overset{\text{def.}}{=} -(\mathbf{K} + \mathbf{K}^\top) + \mathrm{diag}(\mathbf{K})\mathbf{1}_n^T + \mathbf{1}_n \mathrm{diag}(\mathbf{K})^\top.$$

One has the convenient formula

$$\Phi_{\mathbf{A}}^{\infty}(\mathbf{C}) = -\Delta(\mathbf{AC}\,\mathbf{A}^\top).$$

Let the centering operator $\mathbf{J} \stackrel{\text{def.}}{=} \mathbf{Id}_m - \mathbf{1}_{m\times m}/m$, which satisfies $\mathbf{J}^2 = \mathbf{J}$ and $\ker(\mathbf{J}) = \mathrm{Span}(\mathbf{1}_m)$. Let $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^m$ be a pair of singular vectors of $\mathbf{AJ}$, so that there exists $\lambda \in \mathbb{R}$ such that

$$\mathbf{AJv} = \lambda\mathbf{u} \quad \text{and} \quad \mathbf{JA}^\top\mathbf{u} = \lambda\mathbf{v}.$$

By linearity and using the fact that

$$\ker(\Phi_{\mathbf{A}}^{\infty}) = \{\mathbf{C}\,\text{s.t.}\,\mathbf{C} = -\mathbf{C}^\top\} \cup \{\mathbf{a}\mathbf{1}_m^\top + \mathbf{1}_m\mathbf{b}^\top\}$$

one has

$$\begin{aligned}
\Phi_{\mathbf{A}}^{\infty}(\Delta(\mathbf{vv}^\top)) &= -2\Phi_{\mathbf{A}}^{\infty}(\mathbf{vv}^\top) \\
&= 2\Delta(\mathbf{A}^\top\mathbf{vv}^\top\mathbf{A}) \\
&= 2\Delta(\mathbf{JA}^\top\mathbf{vv}^\top\mathbf{AJ}) \\
&= 2\lambda^2\Delta(\mathbf{uu}^\top),
\end{aligned}$$

where we used the fact that $\Delta(\mathbf{JKJ}) = \Delta(\mathbf{K})$. The same reasoning for $\Phi_{\mathbf{B}}^{\infty}$ yields the advertised result. □

B.8 Details on Data Processing

We recovered the 'pbmc3k' dataset using the function `pbmc3k` of Scanpy [Wolf et al., 2018]. Cell types were annotated using the Azimuth [Hao et al., 2021] web tool, which projects it onto large-scale reference atlases. We removed the cluster 'other T cells' and cells for which the annotation was less than 90% confident. Cells were selected using a standard quality filtering pipeline. The data was CPM-normalized, log1p-transformed, and then the 1000 most varying genes were selected. To those genes we added the canonical markers given in the documentation of Azimuth [Hao et al., 2021].

Appendix C

Paired single-cell multi-omics data integration with Mowgli

Supplementary Information for "*Paired single-cell multi-omics data integration with Mowgli*"

The Supplementary Information is organized as follows:

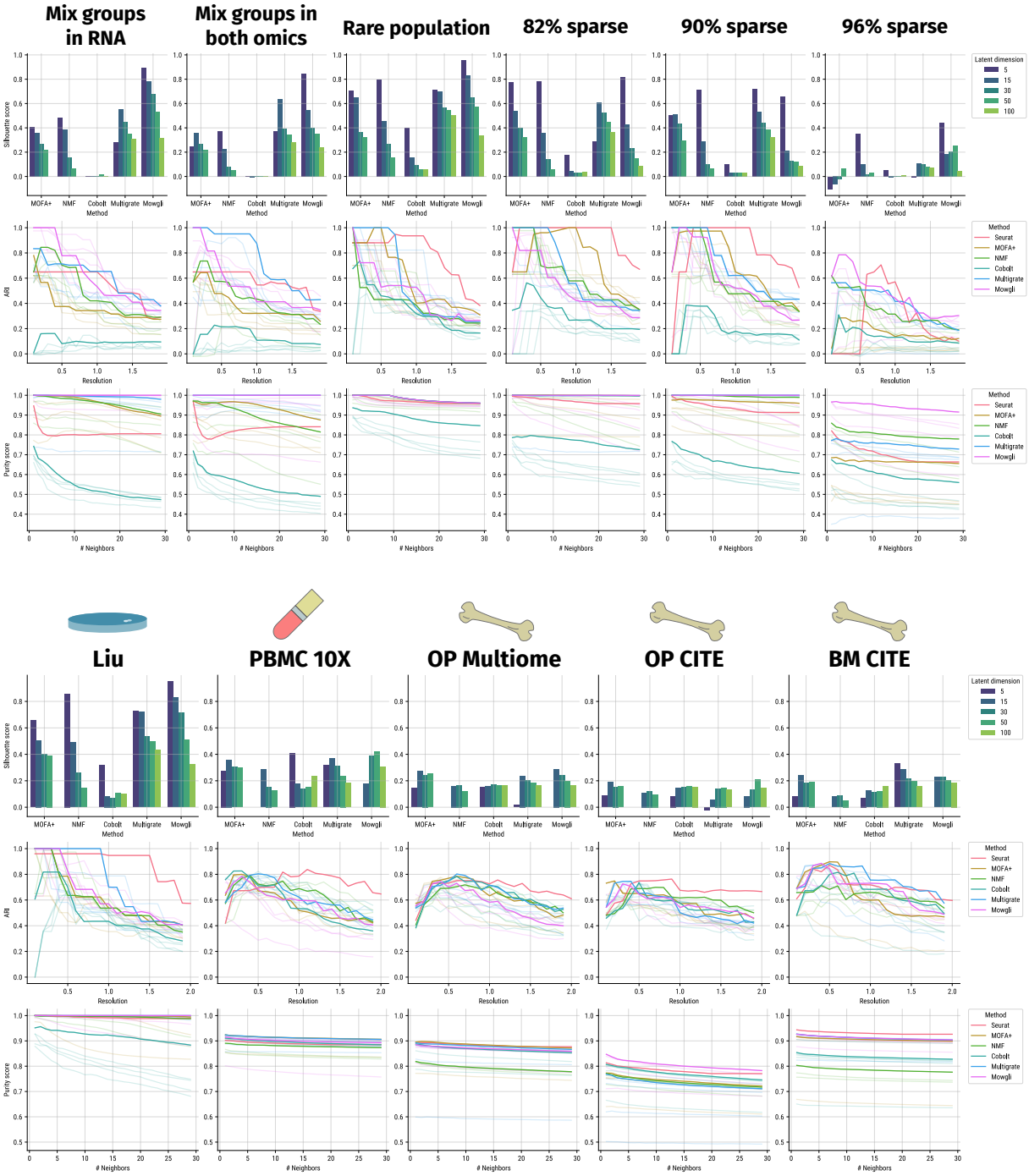
- Page 2 of this PDF: Supplementary Table 1
- Page 3 of this PDF: Supplementary Table 2
- Pages 4 to 19 of this PDF: Supplementary Figures 1 to 16
- Pages 20 to 26 of this PDF: Supplementary Note 1
- Pages 27 to 43 of this PDF: Supplementary Note 2
- Page 44 of this PDF: Supplementary References

Dataset name	Technology	Modalities	Organism	Tissue	Number of cells	Number of annotations	Reference
Liu	scCAT-seq	RNA, ATAC	Human	Cell lines	206	3	Liu et al. [1]
10X PBMC	10X Mutliome	RNA, ATAC	Human	PBMC	9320	14	10X Genomics
OP Multiome	10X Multiome	RNA, ATAC	Human	Bone Marrow	6137	21	Open Problems [2]
OP CITE	CITE-seq	RNA, ADT	Human	Bone Marrow	4249	35	Open Problems [2]
BM CITE	CITE-seq	RNA, ADT	Human	Bone Marrow	29803	27	Stuart et al. [3]
TEA	TEA-seq	RNA, ATAC, ADT	Human	PBMC	7084	no annotations	Swanson et al. [4]

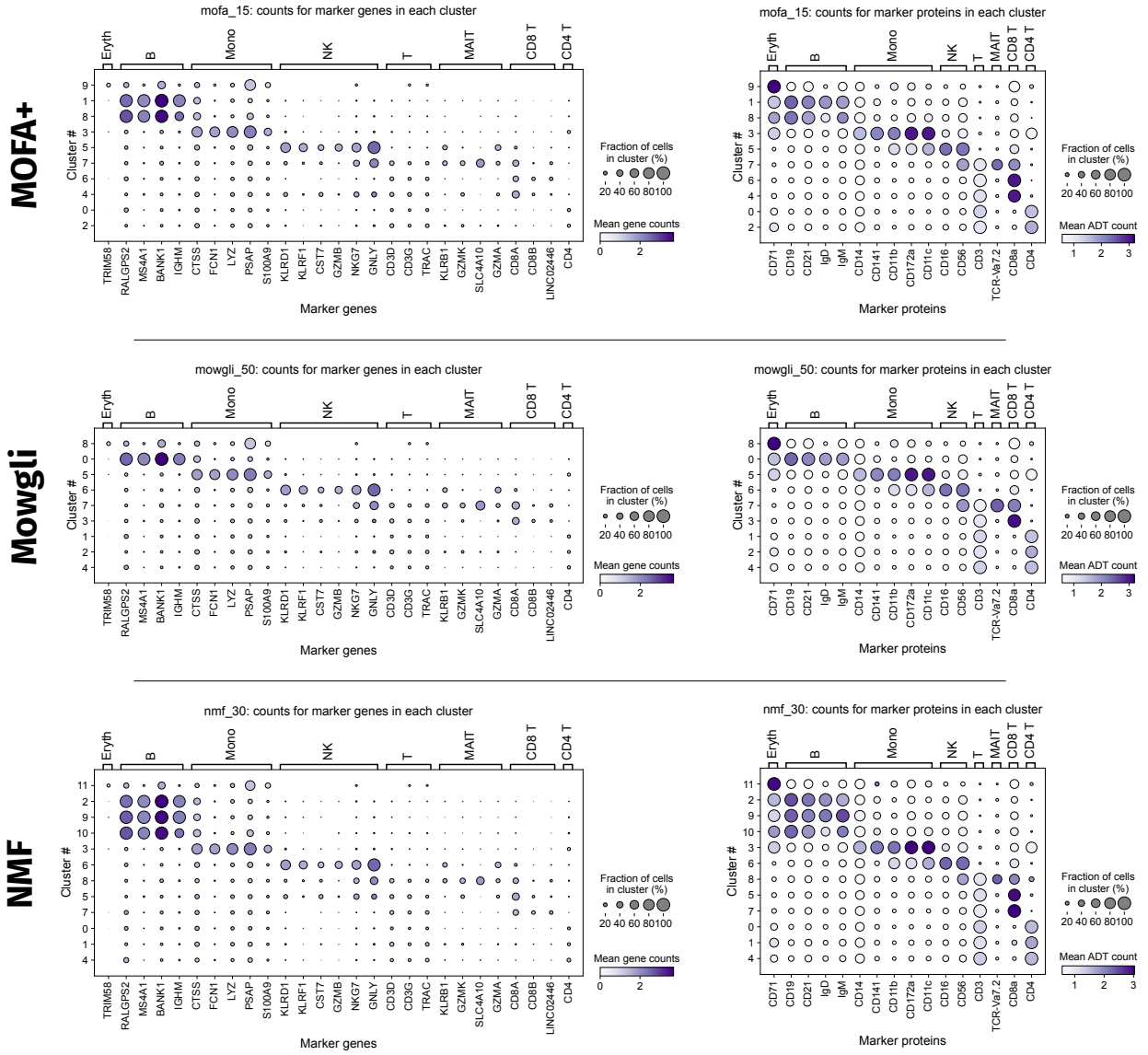
Supplementary Table 1. List of the datasets used in this paper and their characteristics.

Method	CPU or GPU	Dataset				
		Liu	10X PBMC	OP Multiome	OP CITE	BM CITE
Mowgli	GPU (Nvidia A100)	1mn08s	25mn00s	1h42mn	19mn47s	33mn09s
MOFA+	CPU (Intel i5)	24s	17mn29s	21mn48s	3mn24s	23mn4s
Seurat	CPU (Intel i5)	3s	1mn36s	51s	32s	5mn37s
integrative NMF	CPU (Intel i5)	6s	1mn15s	1mn31s	1mn01s	3mn10s
Cobolt	GPU (Nvidia A100)	59s	27mn00s	19mn07s	13mn28s	1h16mns
Multigrate	GPU (Nvidia A100)	42s	03mn55s	3mn04s	01mn43s	7mn18s

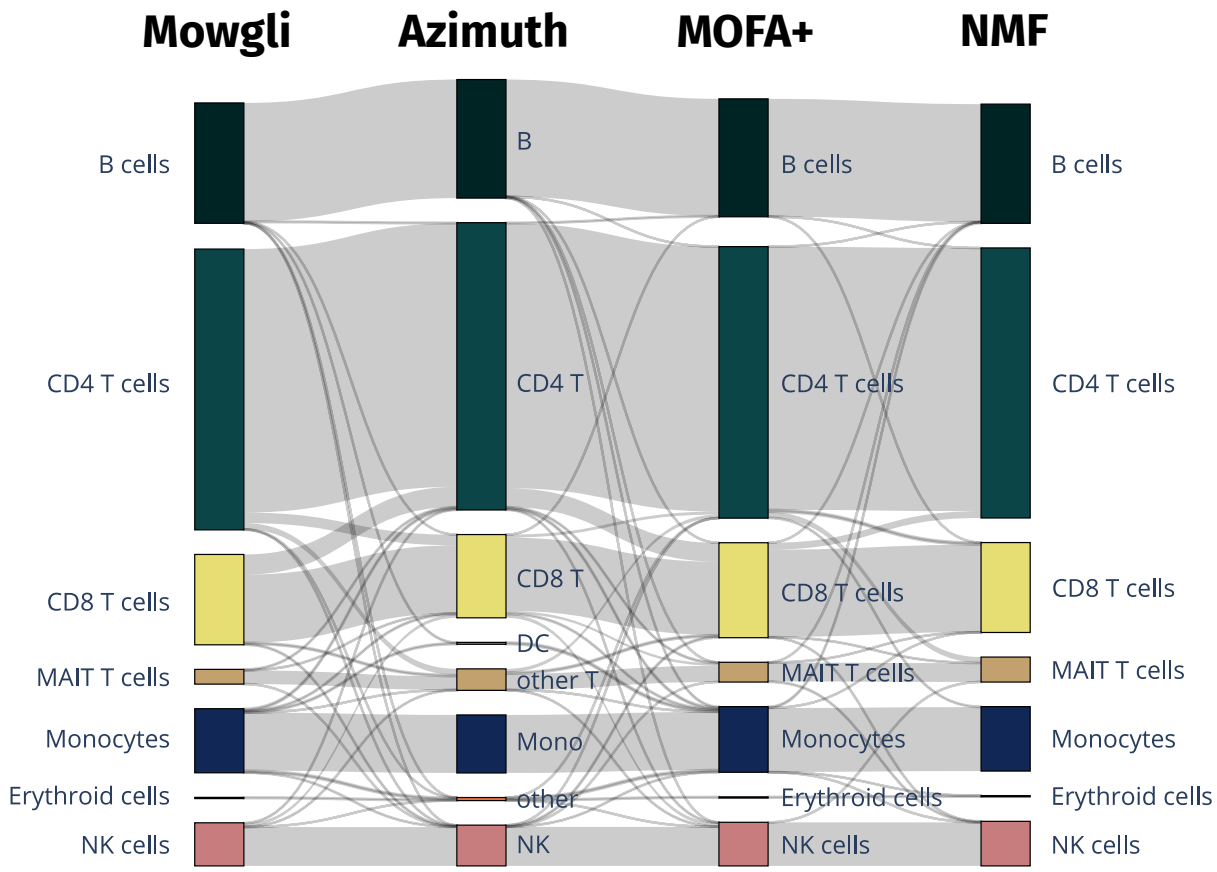
Supplementary Table 2. Compared running times of benchmarked methods across datasets



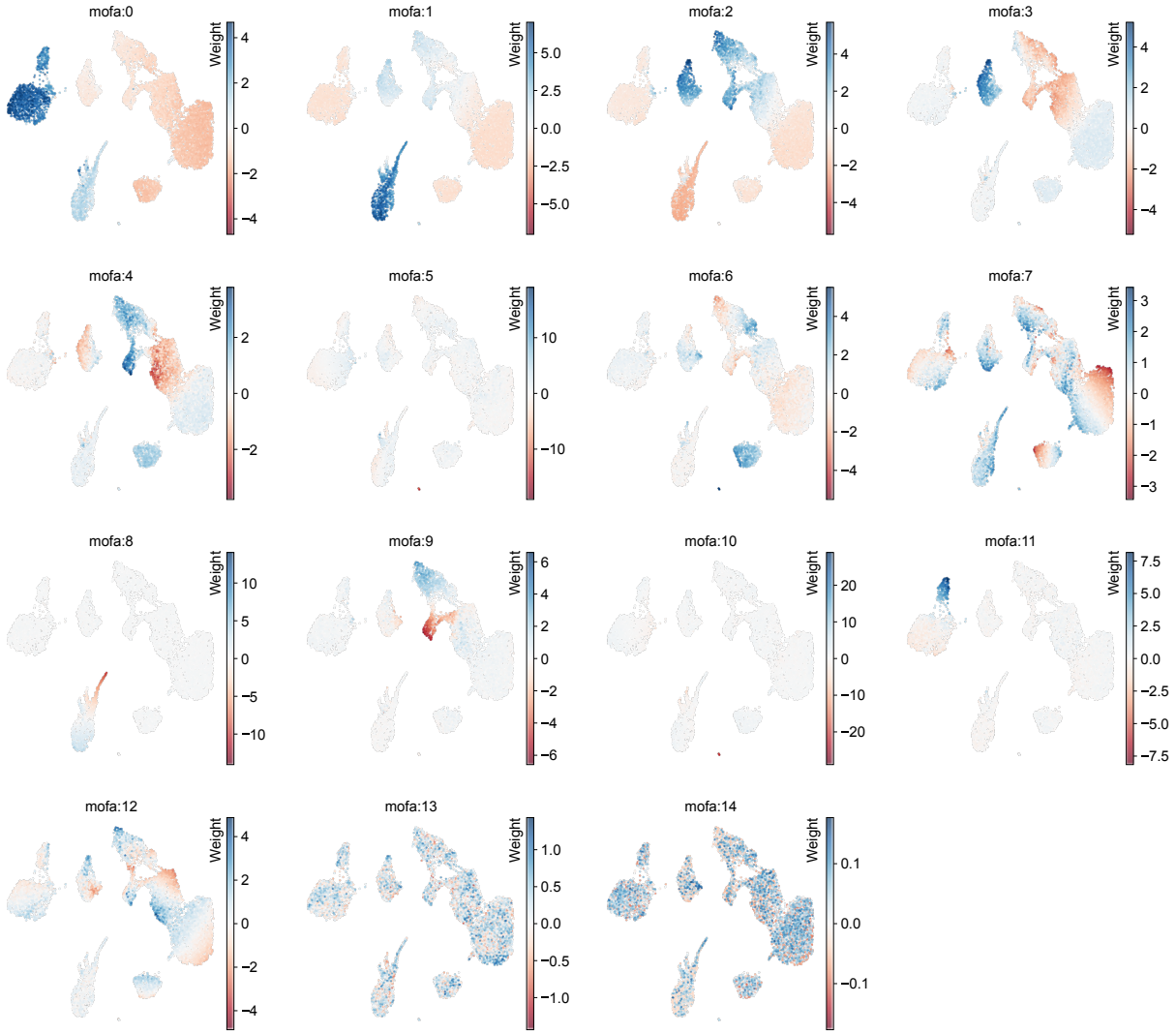
Supplementary Figure 1. Evaluation metrics across datasets and methods for the different numbers of latent dimensions. For each Adjusted Rand Index and purity score subplot, we highlight the best performing number of latent dimensions.



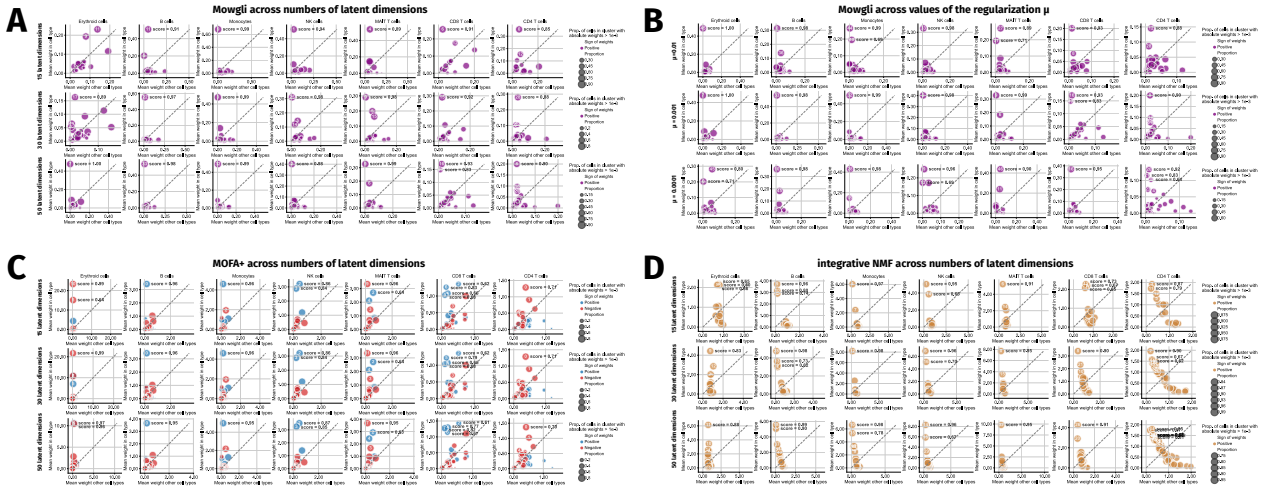
Supplementary Figure 2. Gene and protein markers for coarse annotation of TEA-seq data. Expression of gene and protein markers in the coarse Leiden clusters in the TEA-seq PBMC dataset, for Mowgli, NMF and MOFA+.



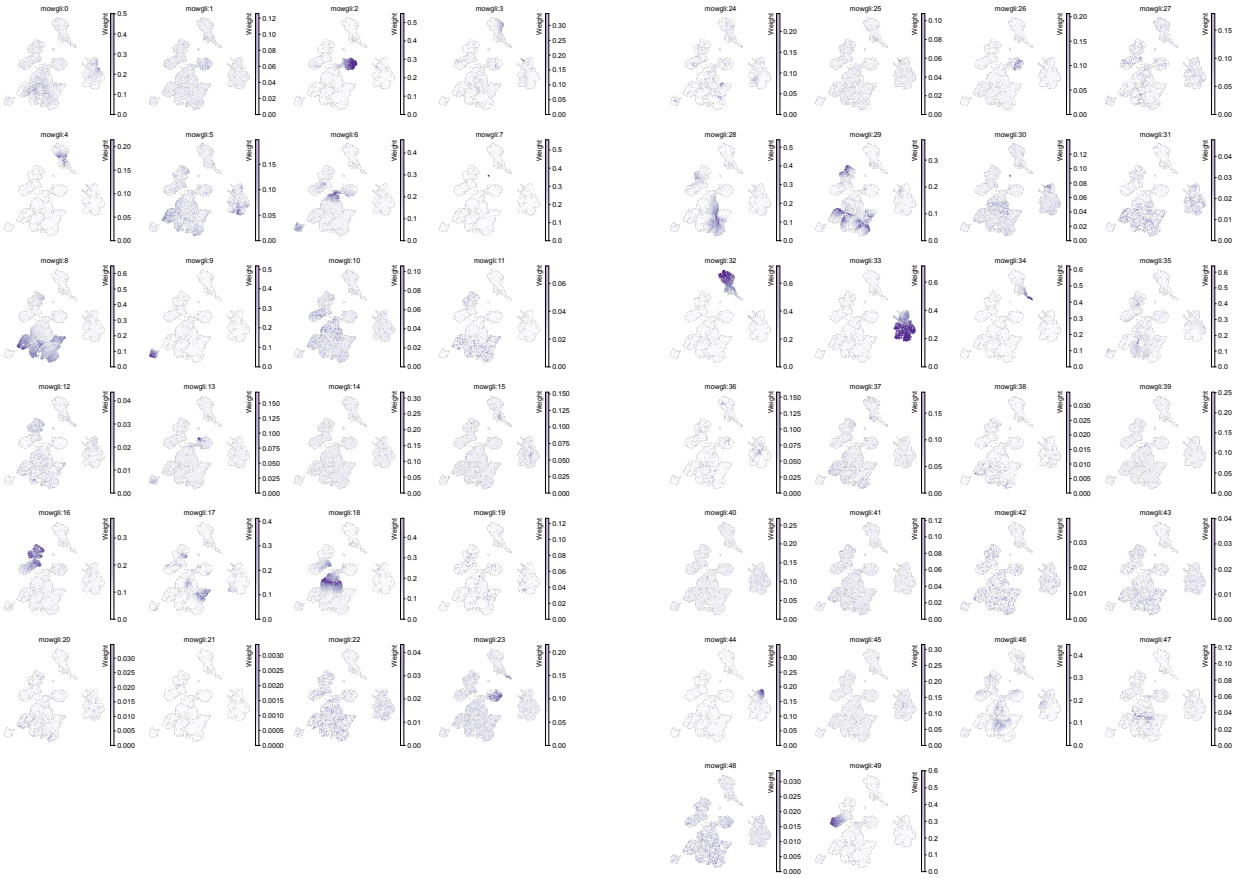
Supplementary Figure 3. Agreement between coarse annotations of TEA-seq data using different methods. We display a Sankey diagram showing the agreement between Mowgli's annotation (first), Azimuth's annotation (second), MOFA+'s annotation (third) and integrative NMF's annotation (fourth).



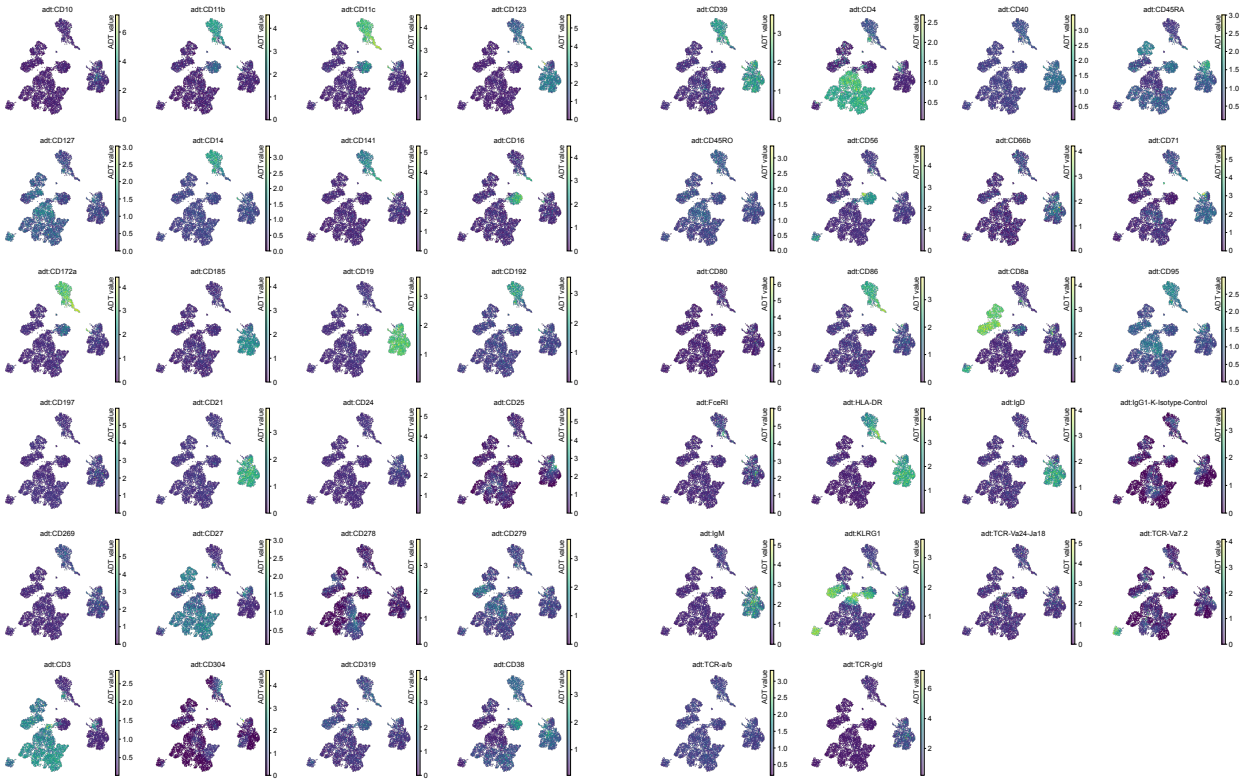
Supplementary Figure 4. MOFA+ factors across cells. For each factor of MOFA+, we display a UMAP plot of MOFA+'s embedding, each cell colored by factor weight.



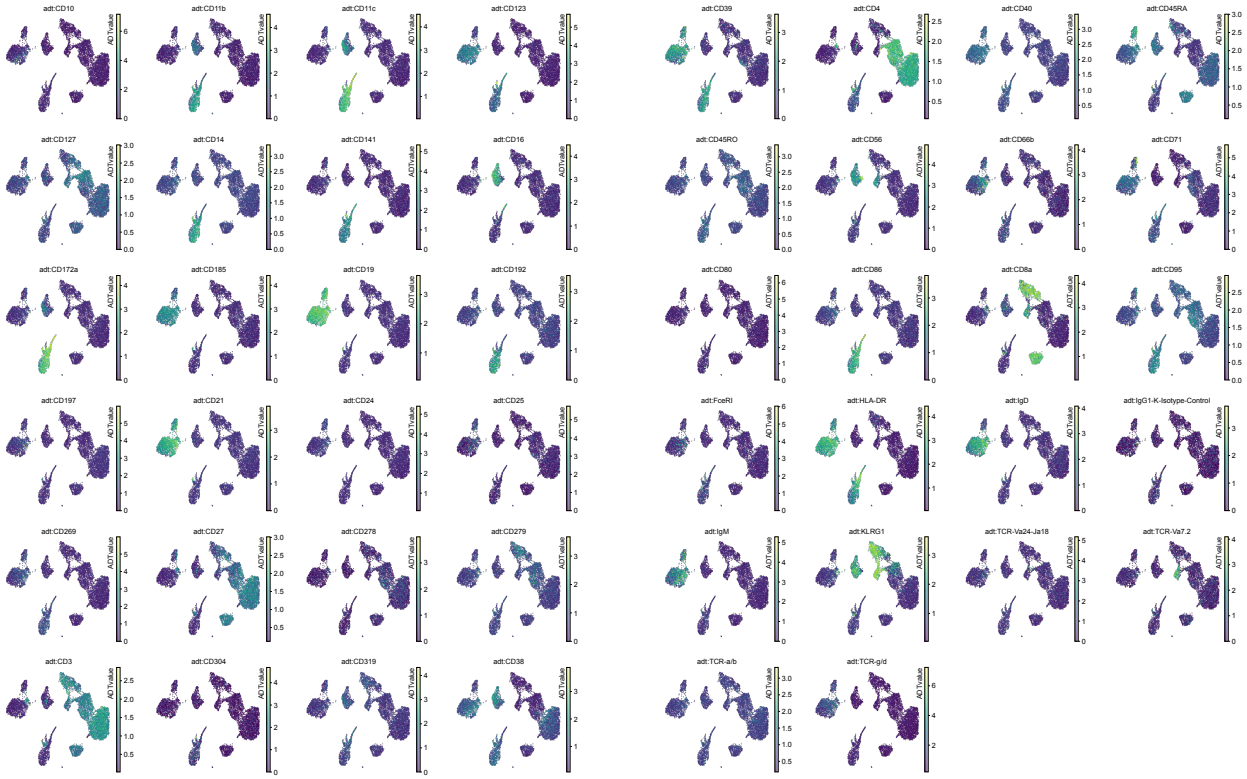
Supplementary Figure 5. Specificity scores on TEA-seq data. (A) Specificity of Mowgli's factors for 15, 30 and 50 latent dimensions. (B) Specificity of Mowgli's factors for μ equal to 0.01, 0.001, 0.0001. (C) Specificity of MOFA+'s factors for 15, 30 and 50 latent dimensions. (D) Specificity of integrative NMF for 15, 30, and 50 latent dimensions.



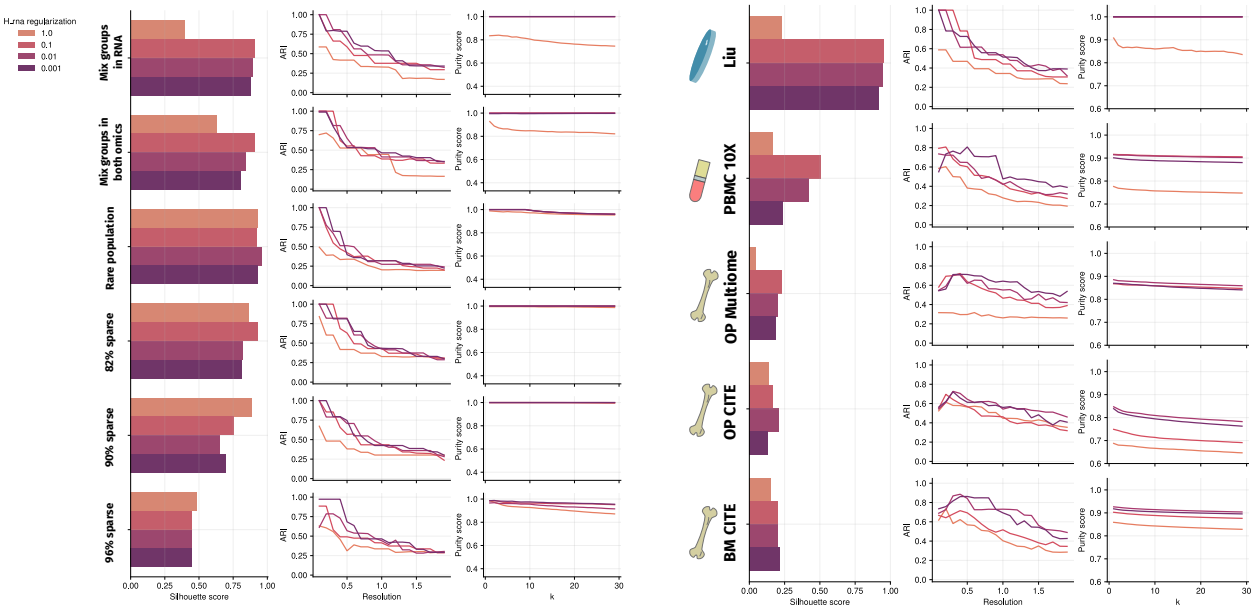
Supplementary Figure 6. Mowgli factors across cells. For each of Mowgli's factors we display a UMAP plot of Mowgli's embedding, each cell colored by factor weight.



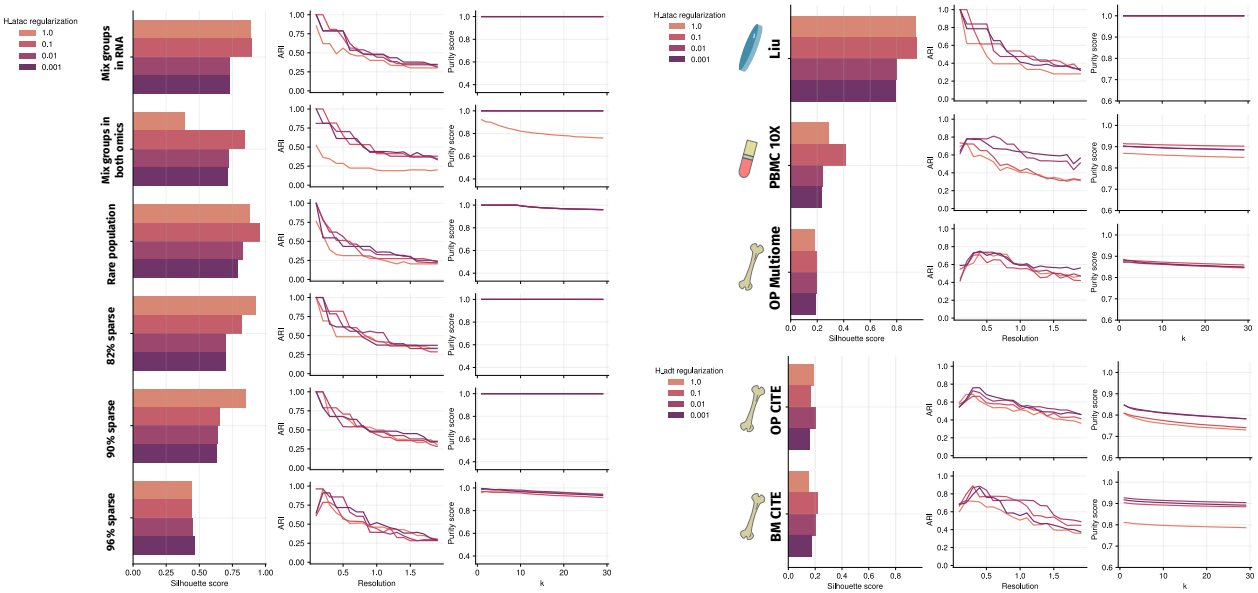
Supplementary figure 7. ADT profile and Mowgli embedding. For each protein we display a UMAP plot of Mowgli cell embeddings, colored by protein counts.



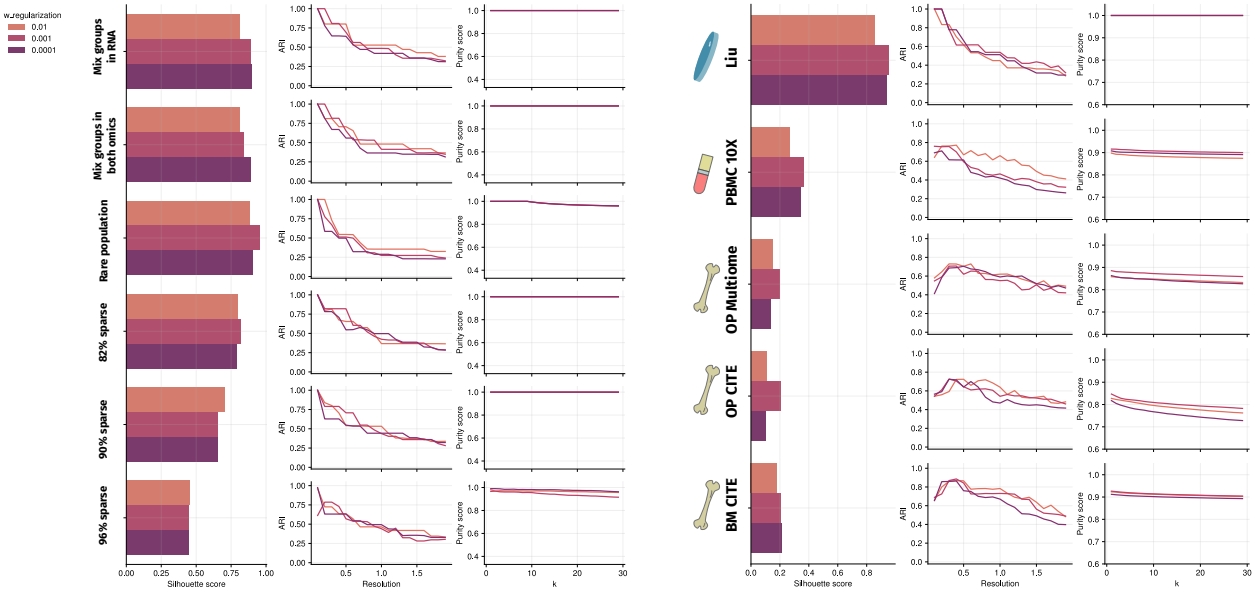
Supplementary Figure 8. ADT profile and MOFA+ embedding. For each protein we display a UMAP plot of MOFA+ cell embeddings, colored by protein counts.



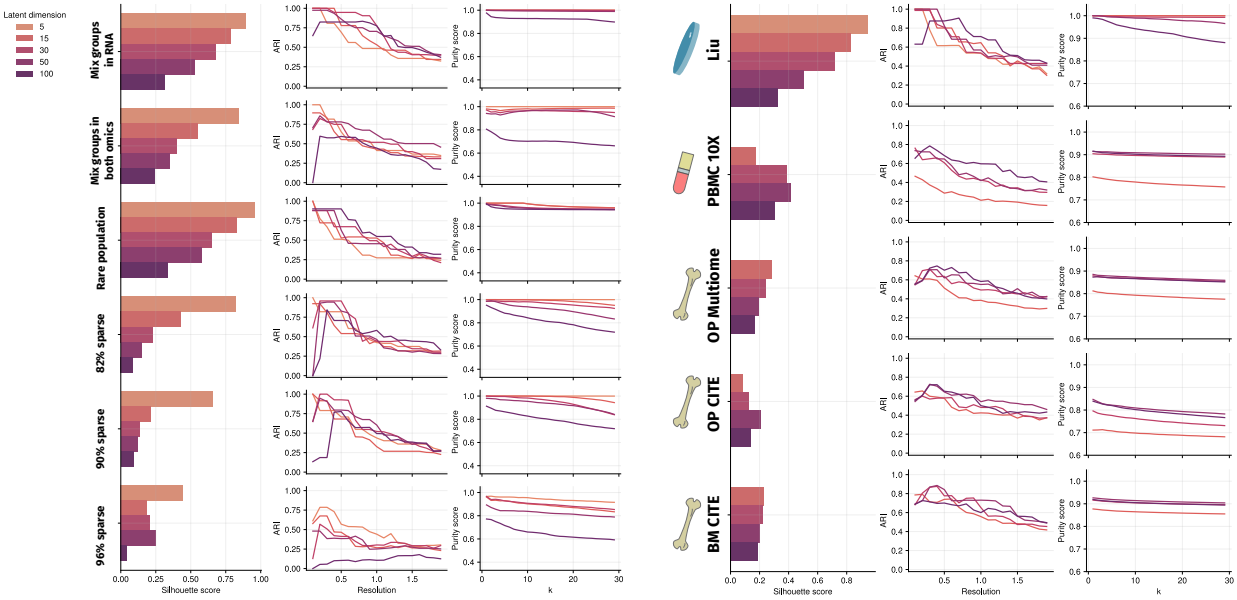
Supplementary Figure 9. Mowgli's clustering and embedding quality. Evaluated across scores, datasets and ρ_{RNA}



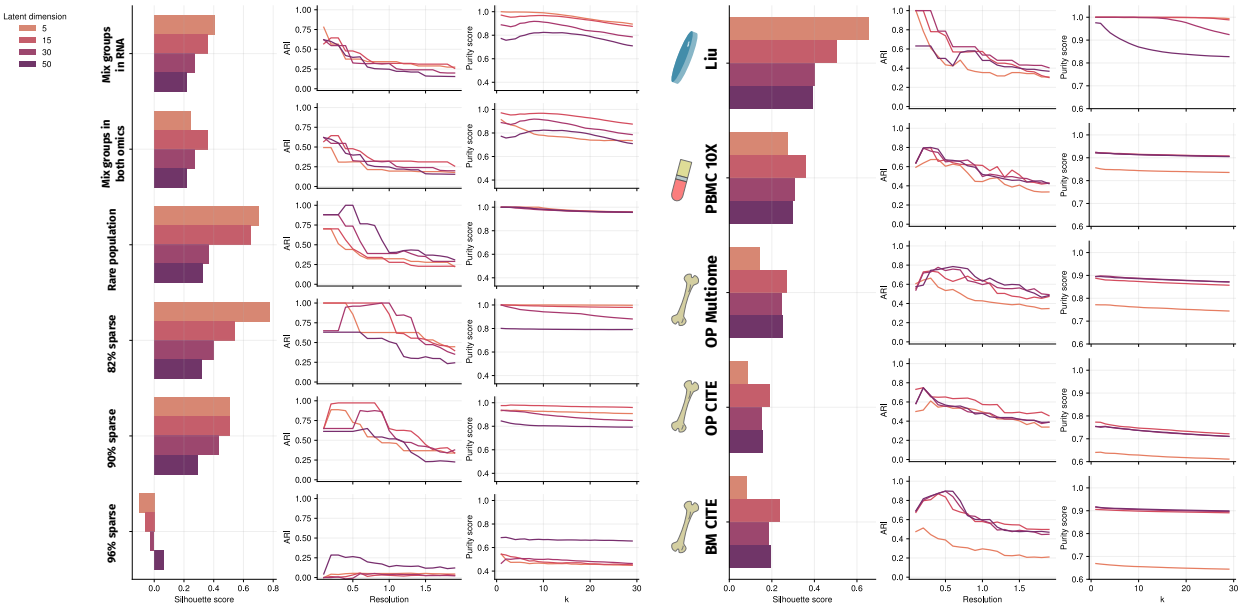
Supplementary Figure 10. Mowgli's clustering and embedding quality. Evaluated across scores, datasets and ρ_{ATAC} or ρ_{ADT}.



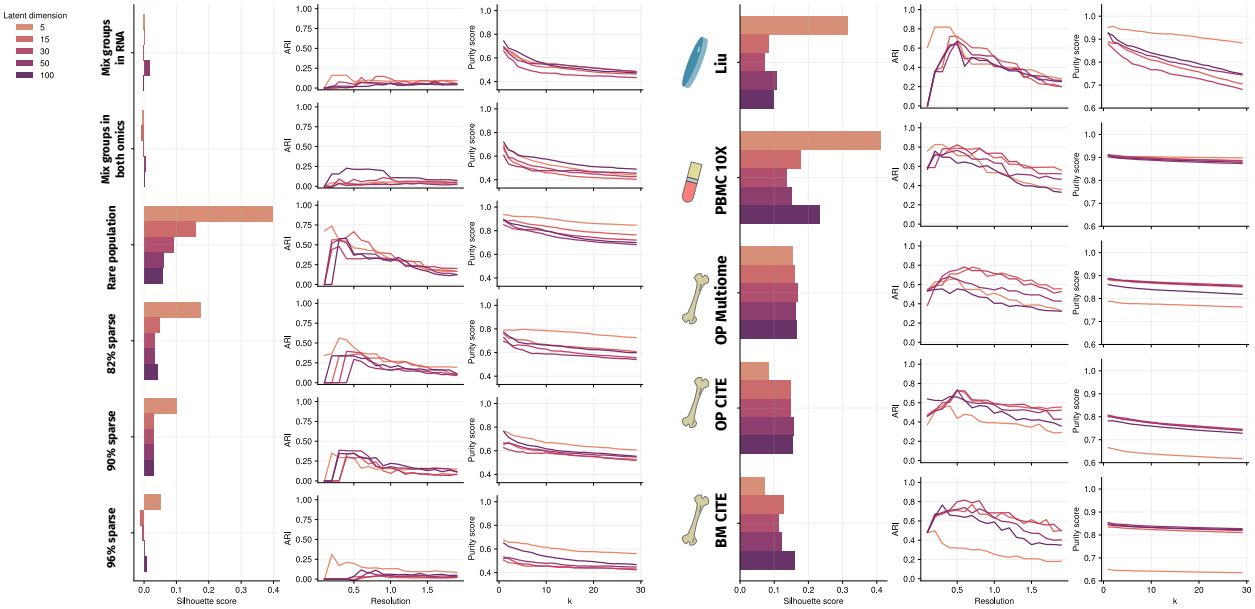
Supplementary figure 11. Mowgli's clustering and embedding quality. Evaluated across datasets, scores and μ



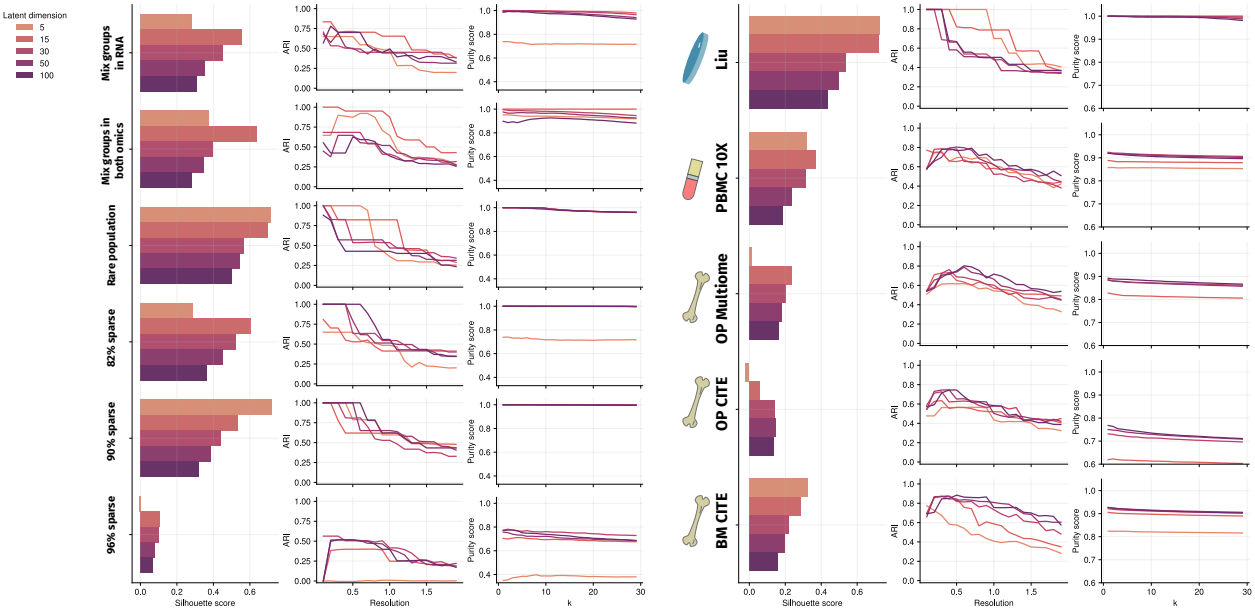
Supplementary Figure 12. Mowgli's clustering and embedding quality. Evaluated across datasets, scores and latent dimensions.



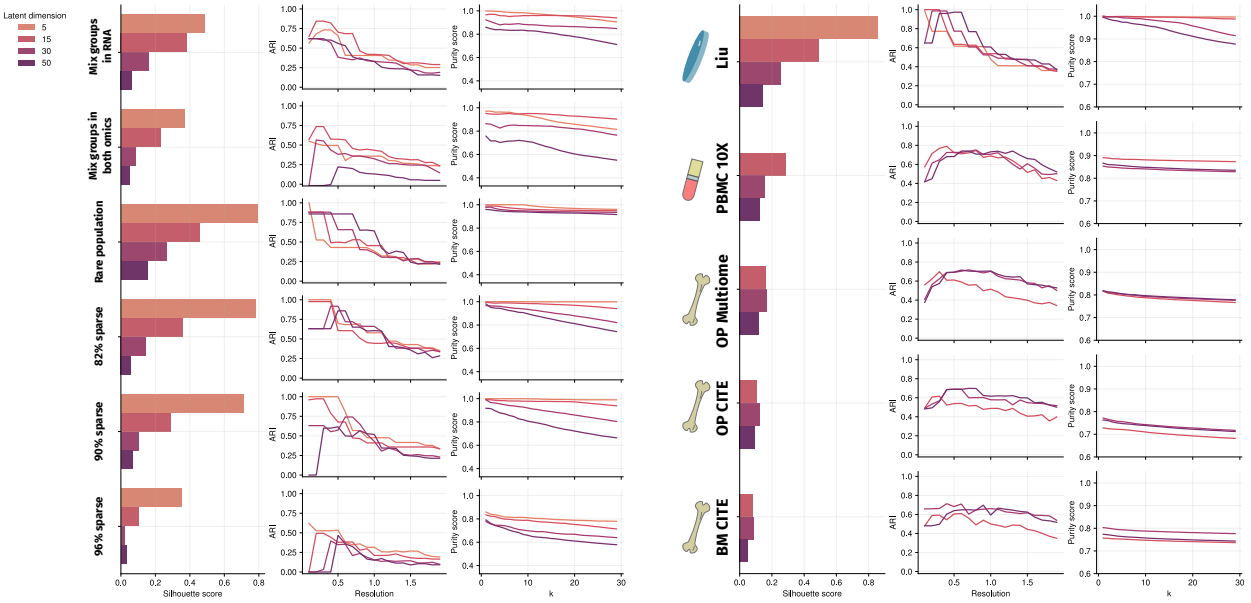
Supplementary Figure 13. MOFA+'s clustering and embedding quality. Evaluated across datasets, scores and latent dimensions.



Supplementary Figure 14. Cobolt's clustering and embedding quality. Evaluated across scores, datasets and latent dimensions.



Supplementary Figure 15. Multigrate's clustering and embedding quality. Evaluated across scores, datasets and latent dimensions.



Supplementary Figure 16. Integrative NMF's clustering and embedding quality. Evaluated across scores, datasets and latent dimensions.

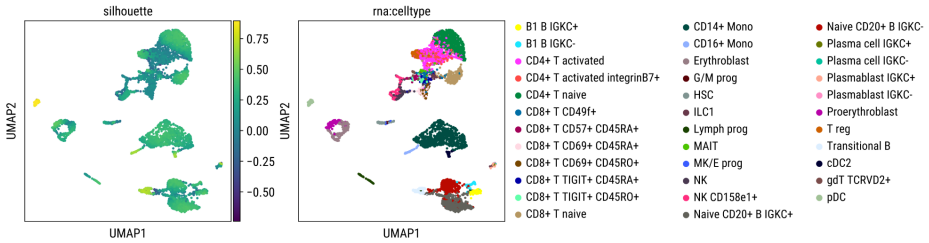
Supplementary Note 1 - Testing the influence of ground truth cell-type annotation on Mowgli's measured performance in real datasets

The metrics used to evaluate embedding and clustering quality in section 3 of Results rely on a ground truth cell-type annotation. Our benchmark is thus limited by the reliability and the granularity of these annotations. We illustrate these challenges using two specific edge cases which negatively affect our benchmark. First, we identify an inconsistency in the annotation of one batch in the *OP CITE* dataset. Second, we show that the granularity of the annotation in *BM CITE* limits the measured performance of Mowgli.

This supplementary note is structured as follows:

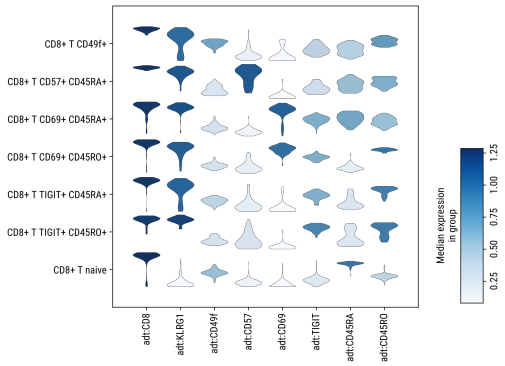
1. Influence of ground truth annotation in the *OP CITE* dataset
2. Influence of ground truth annotation in the *BM CITE* dataset

1. Influence of ground truth annotation in the *OP CITE* dataset



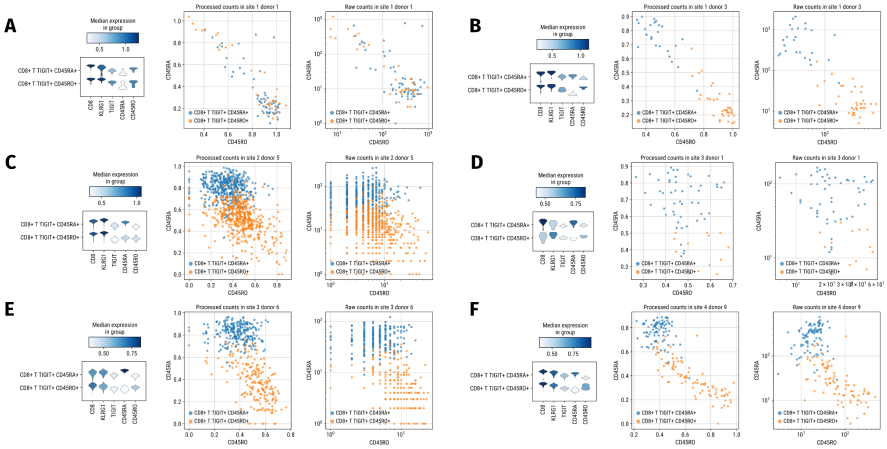
Supplementary Figure 17. Overview of Mowgli's silhouette score across cell types in *OP CITE*. UMAP plot colored by cell-wise silhouette score (left) and cell-type annotation (right) for the *OP CITE* dataset.

Supplementary Figure 17 displays the cell-wise silhouette score and the cell-type annotation for the OP CITE dataset. Subsets of CD8+ T cells have overall low silhouette scores and thus contribute negatively to Mowgli's measured performance. Among the CD8+ T cells, there are two subpopulations CD8+TIGIT+CD45RA+ and CD8+TIGIT+CD45RO+ separated by their levels of CD45RA and CD45RO.



Supplementary Figure 18. Compared ADT profiles of CD8+ T cells in *OP CITE*. ADT profiles of CD8+ T cell subtypes in the *OP CITE* dataset's first batch (site 1, donor 1).

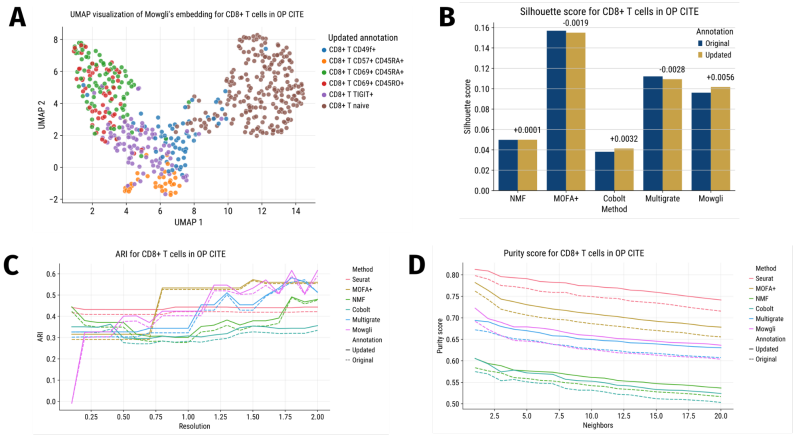
We compared the ADT profiles of CD8+ T cell subpopulations in order to investigate the reasons behind their low silhouette scores (see Supplementary Figure 18). Surprisingly, Supplementary Figure 18 shows comparable levels of CD45RO and CD45RA in the CD8+TIGIT+CD45RA+ and the CD8+TIGIT+CD45RO+ cell types.



Supplementary Figure 19. Compared values of CD45RA and CD45RO in the OP CITE dataset. (A to F) Compared values of CD45RA and CD45RO in the CD8+TIGIT+CD45RA+ and the CD8+TIGIT+CD45RO+ cell types, across batches of the OP CITE dataset.

Supplementary Figure 19 displays the relative values of CD45RA and CD45RO in these two cell types as scatterplots both for normalized and raw data, and confirms that CD45RA and CD45RO do not allow distinguishing these cell types. Interestingly, this is only the case in "site 1 donor 1", the batch we used for our benchmark. Panels B to F of Supplementary Figure 19 show that in all other batches containing

these cell types, CD45RO and CD45RA patterns are undoubtedly distinct. This inconsistency in the ground truth cell-type annotation negatively affects the evaluation of the various methods in our benchmark. Indeed, cells with similar single-cell profiles but with different annotations will be embedded close together.

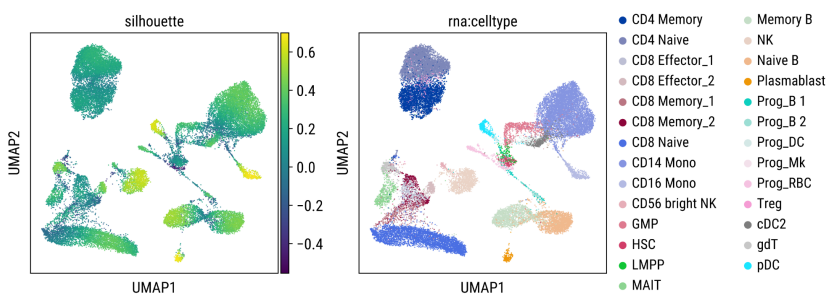


Supplementary Figure 20. Impact of imperfect ground truth annotation on evaluation metrics. (A) UMAP visualization of Mowgli's embedding for CD8+ T cells in the OP CITE data, colored by updated annotation. (B) Silhouette score for CD8+ T cells in OP CITE, before and after updating the annotation. (C) Adjusted Rand Index for CD8+ T cells in OP CITE, before and after updating the annotation. (D) Purity score for CD8+ T cells in OP CITE, before and after updating the annotation.

We investigated whether imperfect ground truth annotation impacts the evaluation metrics evenly for all methods. For this, we updated the annotation by fusing the CD8+TIGIT+CD45RA+ and CD8+TIGIT+CD45RO+ subtypes into a single CD8+TIGIT+ subtype. Panel A of Supplementary Figure 20 displays a UMAP visualization of Mowgli's embedding for the different CD8+ T cell subtypes in the *OP CITE* dataset, colored according to the updated annotation. Panel B of Supplementary Figure 20 shows silhouette scores computed across CD8+ T cells for different methods. The updated annotation leads to an increase in silhouette score for NMF, Cobolt, and Mowgli, and to a decrease in silhouette score for MOFA+ and Multigrate. Panels C and D of Supplementary Figure 20 show that the updated annotation leads to an increase in ARI and purity score for all methods. Supplementary Figure 20 shows that changes in ground truth annotation have an uneven impact on evaluation across methods and metrics. This limits the reliability of benchmarking in the presence of imperfect annotation, which should be complemented by an in-depth

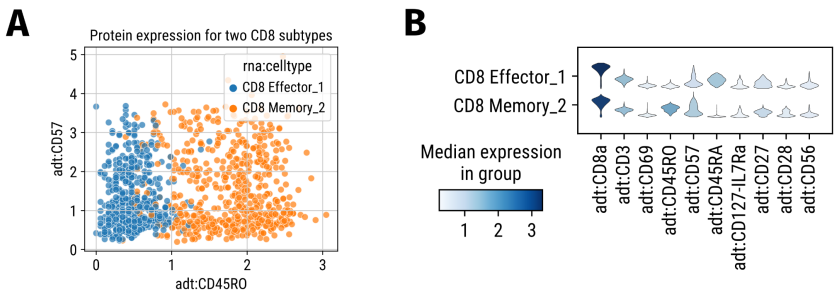
biological analysis of results, as proposed in this Supplementary Note and in the Results.

2. Influence of ground truth annotation in the *BM CITE* dataset



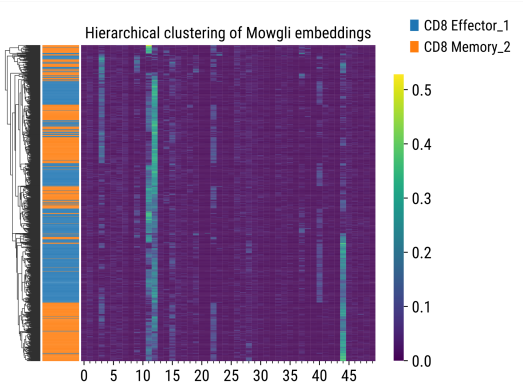
Supplementary Figure 21. Overview of Mowgli's silhouette score across cell types in *BM CITE*. Cell-wise silhouette score (left) and cell-type annotation (right) for the BM CITE dataset.

Supplementary Figure 21 displays the cell-wise silhouette score and the cell-type annotation for the BM CITE dataset. The CD8 Effector_1 and CD8 Memory_2 cell types, which were computationally derived in [5], have overall low silhouette scores and thus contribute negatively to Mowgli's measured performance.



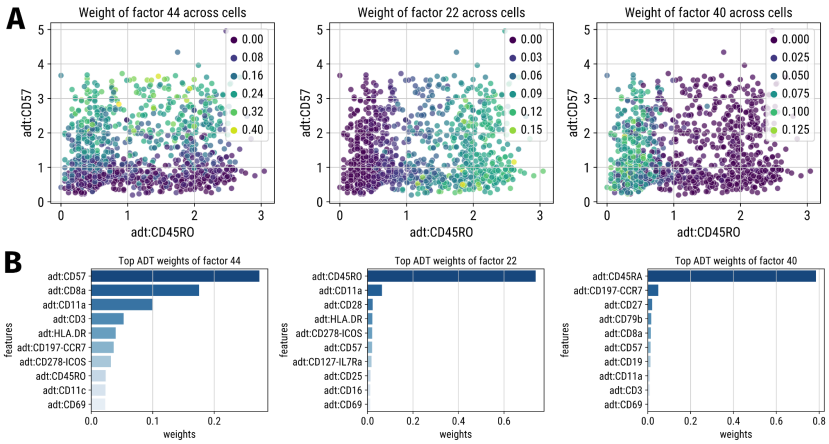
Supplementary Figure 22. Compared ADT profiles of the CD8 Effector_1 and CD8 Memory_2 populations of the BM CITE dataset. (A) Compared values of CD45RO and CD57 in the CD8 Effector_1 and CD8 Memory_2 cell types of the BM CITE dataset. (B) Compared ADT profiles of CD8 Effector_1 and CD8 Memory_2 in the BM CITE dataset.

Supplementary Figure 22 shows that CD45RO levels clearly distinguish CD8 Effector_1 and CD8 Memory_2 cells.



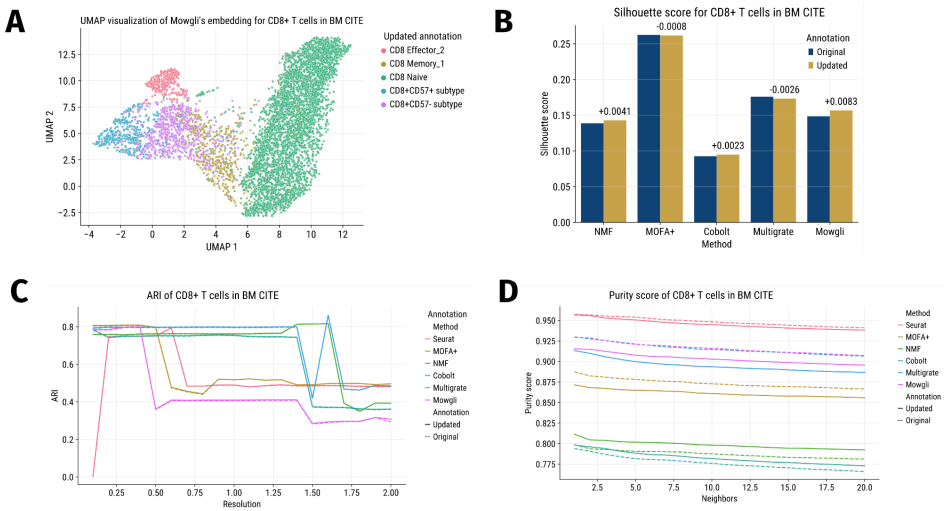
Supplementary Figure 23. Hierarchical clustering of CD8 subtypes using Mowgli embeddings. Hierarchical clustering on Mowgli's embeddings of CD8 Effector_1 and CD8 Memory_2 cells in the BM CITE dataset.

Despite this clear separation, hierarchical clustering on Mowgli's embeddings mixes CD8 Effector_1 and CD8 Memory_2 cells, which explains low silhouette scores (Supplementary Figure 23). One strong signal driving this clustering is factor 44, even though factors 22 and 40 would have allowed the clustering of CD8 Memory_2 and CD8 Effector_1 cells separately.



Supplementary Figure 24. Compared ADT profiles and Mowgli factor weights for the CD8 Effector_1 and CD8 Memory_2 populations of the BM CITE dataset. (A) Compared values of CD45RO and CD57 in the CD8 Effector_1 and CD8 Memory_2 cell types of the BM CITE dataset. Cells are colored by their weight of Mowgli factors 44, 22, and 40 (B) Top ADT weights in Mowgli factors 44, 22, and 40.

Supplementary Figure 24 shows that together, factors 44, 22, and 40 effectively split the CD8 Effector_1 and CD8 Memory_2 populations into four subpopulations: CD57+CD8+ Effector_1, CD57-CD8+ Effector_1, CD57+CD8+ Memory_2, and CD57-CD8+ Memory_2. While this granularity is not present in the annotation, it is mentioned in the original publication [5]. Indeed, [5] explicitly highlights the relevance of CD57 as a marker of effector cells and displays its expression in the associated Figure S2, where CD57 clearly splits CD8 Effector_1 and CD8 Memory_2 into CD57+ and CD57- subpopulations. In this example, our benchmark penalizes Mowgli even though it detects more granularity than the original annotation.



Supplementary Figure 25. Impact of imperfect ground truth annotation on evaluation metrics. (A) UMAP visualization of Mowgli's embedding for CD8+ T cells in the BM CITE data, colored by updated annotation. (B) Silhouette score for CD8+ T cells in BM CITE, before and after updating the annotation. (C) Adjusted Rand Index for CD8+ T cells in BM CITE, before and after updating the annotation. (D) Purity score for CD8+ T cells in BM CITE, before and after updating the annotation.

As in section 1 of this Supplementary Note, we investigated whether imperfect ground truth annotation impacts the evaluation metrics evenly for all methods. For this, we updated the annotation by reorganizing cells from the CD8 Effector_1 and CD8 Memory_2 subtypes into a CD57+ and a CD57- subtype. Panel A of Supplementary Figure 25 displays a UMAP visualization of Mowgli's embedding for the different CD8+ T cell subtypes in the *BM CITE* dataset, colored according to the updated annotation. Panel B of Supplementary Figure 25 shows silhouette scores

computed across CD8+ T cells for different methods. As in *OP CITE*, the updated annotation leads to an increase in silhouette score for NMF, Cobolt, and Mowgli, and to a decrease in silhouette score for MOFA+ and Multigrate. Panel C of Supplementary Figure 25 shows no influence of the updated annotation on ARI. Panel D of Supplementary Figure 25 shows that the updated annotation leads to an increase in purity score for Cobolt and NMF, and a decrease in purity score for Seurat, MOFA+, Multigrate, and Mowgli. As observed in the *OP CITE* dataset, changes in ground truth annotation have an uneven impact on evaluation across methods and metrics for the *BM CITE* dataset.

Supplementary Note 2 - Comparing the interpretability of Mowgli and MOFA+'s factors for the detection of specific subtypes in the TEA-seq PBMC dataset

To complement Figure 5b and section 5 of Results, we quantitatively analyzed factors of Mowgli associated with specific immune subpopulations in the *TEA-seq PBMC* dataset: CD56 bright NK cells, classical and nonclassical monocytes, naive CD4 T cells, plasmacytoid dendritic cells, and conventional dendritic cells.

We assessed the extent of this association through (i) the ranking of markers in the factor's top proteins, (ii) the Pearson correlation across cells between the factor and a protein marker, and (iii) scatterplots with compared values of the factor and of protein markers. We used these same quantitative assessments to identify if MOFA+ presents factors associated with the considered subpopulations.

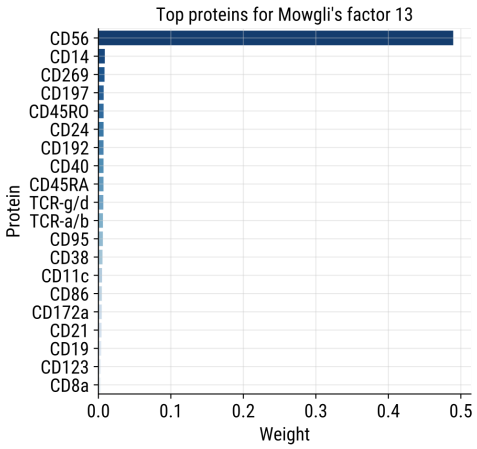
In nonclassical monocytes, both MOFA+ and Mowgli present a factor closely associated with the subpopulation. In the other considered subpopulations, Mowgli's signal is more biologically relevant than MOFA+'s.

This supplementary note is structured as follows:

1. CD56 bright NK cells
2. Plasmacytoid dendritic cells
3. Naive CD4 T cells
4. Conventional dendritic cells
5. Nonclassical monocytes
6. Classical monocytes

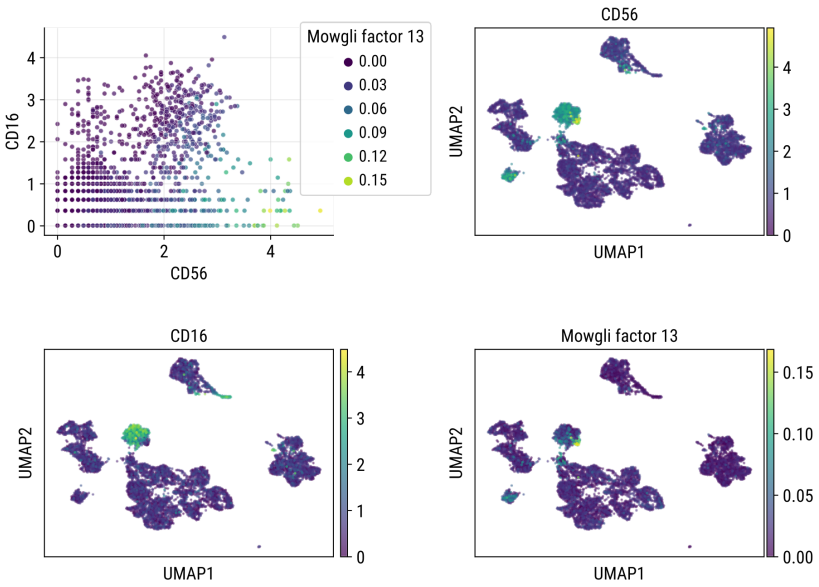
1. CD56 bright NK cells

CD56 bright NK cells have CD56-bright and CD16-negative phenotypes. Mowgli's factor 13 is largely driven by CD56 and has a low contribution of CD16, and is thus coherent with the biological signature of CD56 bright NK cells (see Supplementary Figure 26).



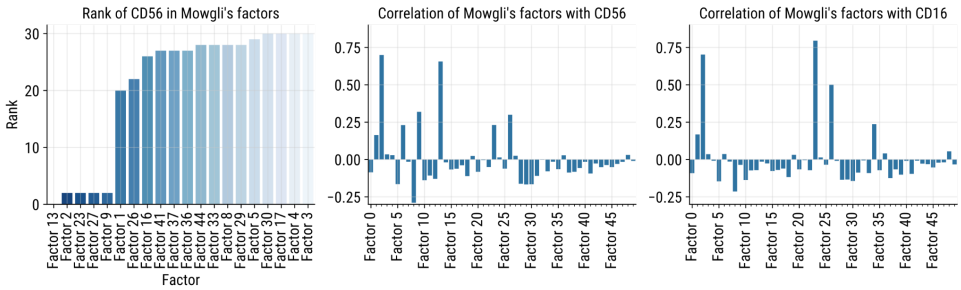
Supplementary Figure 26. Protein weights for Mowgli's factor 13. Top ADT weights for Mowgli's factor 13 in the TEA dataset.

In addition, high weights for Mowgli's factor 13 coincide with low CD16 and high CD56 (see Supplementary Figure 27).



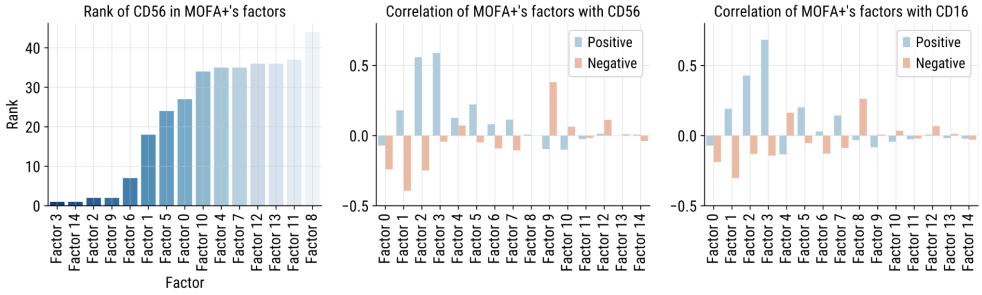
Supplementary Figure 27. Compared ADT profiles and Mowgli factor 13. Compared values of CD16, CD56, and weight of Mowgli's factor 13.

Mowgli's factor 13 is the only one with a high Pearson correlation with CD56 and a low absolute Pearson correlation with CD16 (see Supplementary Figure 28). In addition, CD56 ranks first in factor 13's top genes, which is the best rank across factors.



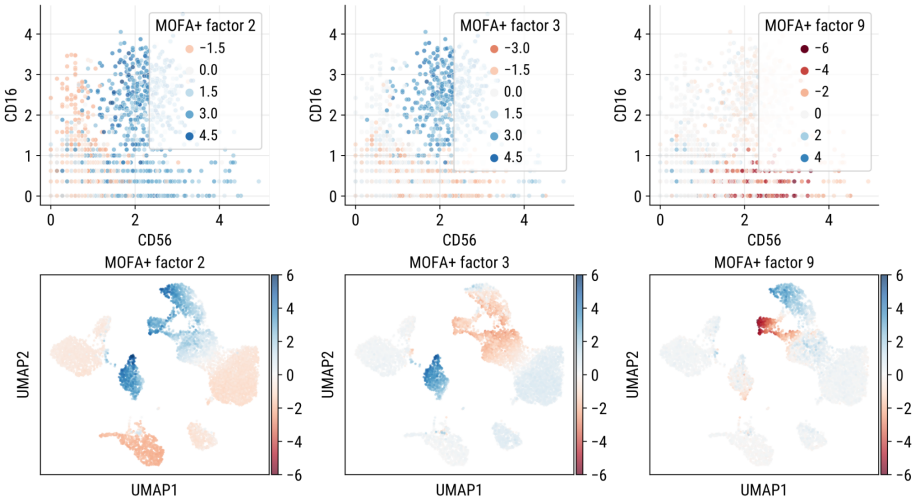
Supplementary Figure 28. Correlation between CD56, CD16, and Mowgli's factors. Compared rank of CD56 in Mowgli's factors, and correlation of CD56 and CD16 with Mowgli's factors.

By contrast, MOFA+ has no factors that are both highly correlated with CD56 and uncorrelated with CD16 (see Figure Supplementary Figure 29).



Supplementary Figure 29. Correlation between CD56, CD16, and MOFA+'s factors. Compared rank of CD56 in MOFA+'s factors, and correlation of CD56 and CD16 with MOFA+'s factors.

Instead, MOFA+'s factors where CD56 has a good rank or factors highly correlated with CD56 coincide with general CD56+ populations: NK cells (MOFA+'s factors 2 and 3, positive weights) and MAIT T-cells (MOFA+'s factor 9, negative weights) (see Supplementary Figure 30).

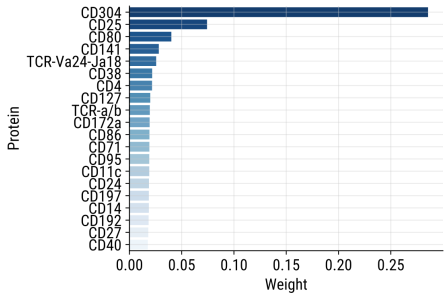


Supplementary Figure 30. Compared ADT profiles and MOFA+ factors. Compared values of CD16, CD56, and weights of MOFA+'s factors 2, 3, and 9.

In conclusion, Mowgli's factor 13 is coherent with the biological signature of CD56 bright NK cells and allows one to distinguish CD56 bright NK cells easily. This is impossible using a unique factor from MOFA+. Indeed, MOFA+'s factors match broad CD56+ populations at best.

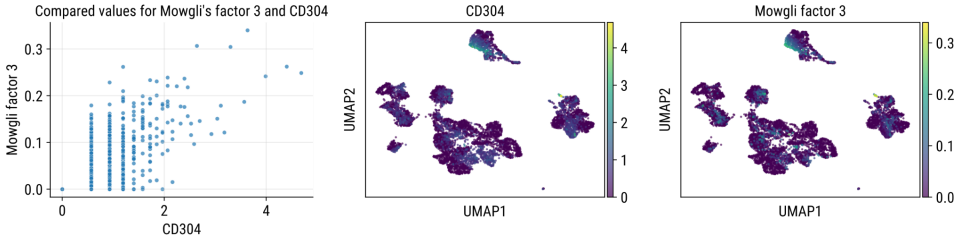
2. Plasmacytoid dendritic cells

Plasmacytoid dendritic cells (pDC) have a distinctive CD304-positive and CD14-negative phenotype [6]. Mowgli's factor 3 is largely driven by CD304, and has a low contribution of CD14, and is thus coherent with the biological signature of pDCs (see Supplementary Figure 31).



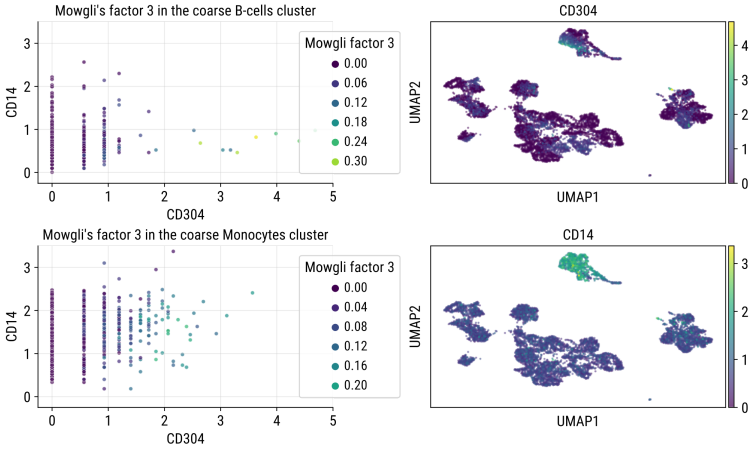
Supplementary Figure 31. Protein weights for Mowgli's factor 3. Top ADT weights for Mowgli's factor 3 in the TEA dataset.

High weights for Mowgli's factor 3 coincide with high CD304 (see Supplementary Figure 32).



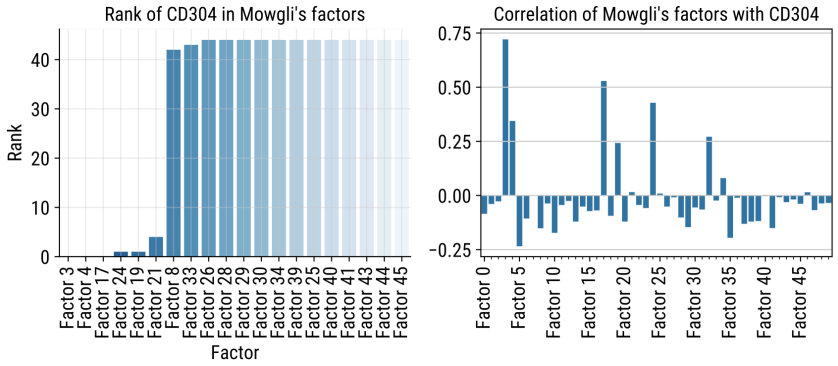
Supplementary Figure 32. Compared ADT profiles and Mowgli factor 3. Compared values of CD304 and weight of Mowgli's factor 3.

In the coarse B-cells cluster, cells with high weights of Mowgli's factor 3 have low CD14 and high CD304, thus corresponding to the pDC phenotype. In the coarse Monocytes cluster, cells with high weights of Mowgli's factor 3 have higher levels of CD14 and lower levels of CD304 than in the B-cells cluster, thus suggesting a CD304+ monocyte subpopulation rather than pDCs (see Supplementary Figure 33). Note also that the highest values of Mowgli's factor 3 are achieved in the coarse B-cells cluster.



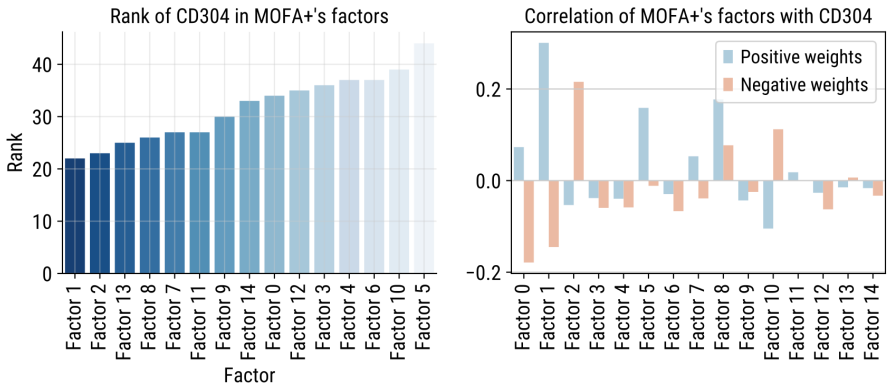
Supplementary Figure 33. Compared ADT profiles and Mowgli factor 3 in subtypes. Compared values of CD304, CD14, and weight of Mowgli's factor 3 in the coarse B-cells and Monocytes subclusters.

Mowgli's factor 3 is the factor most correlated with CD304, and CD304 ranks first among its top proteins (see Supplementary Figure 34).



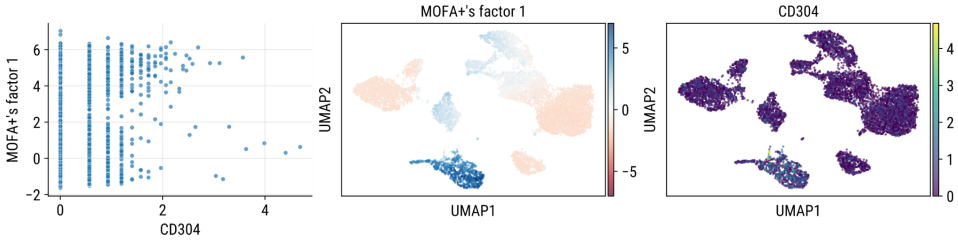
Supplementary Figure 34. Correlation between CD304 and Mowgli's factors. Compared rank of CD304 in Mowgli's factors, and correlation of CD304 with Mowgli's factors.

By contrast, MOFA+ has no factors highly correlated with CD304 or factors where CD304 has a low rank (see Supplementary Figure 35).



Supplementary Figure 35. Correlation between CD304 and MOFA+'s factors. Compared rank of CD304 in MOFA+'s factors, and correlation of CD304 with MOFA+'s factors.

MOFA+'s factor 1 is the most highly correlated with CD304 and the one where CD304 ranks best, but it coincides with monocytes rather than pDCs (see Supplementary Figure 36).

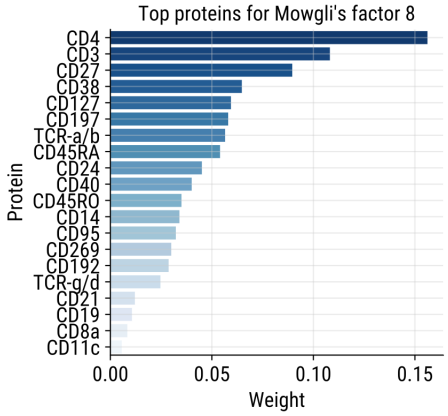


Supplementary Figure 36. Compared ADT profiles and MOFA+ factor 1. Compared values of CD304 and weights of MOFA+'s factor 1.

In conclusion, Mowgli's factor 3 allows one to identify pDCs easily, which is impossible with a unique factor from MOFA+. Indeed, MOFA+ presents no factor closely associated with pDCs, and the best candidate coincides with the broad population of monocytes instead.

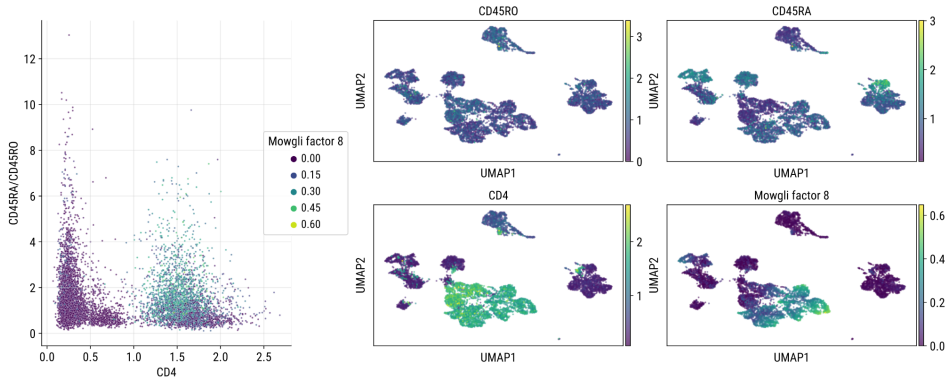
3. Naive CD4 T cells

Naive CD4 T cells are CD4-positive and present higher levels of CD45RA than CD45RO. Mowgli's factor 8 is driven by CD4 and has a higher weight for CD45RA than CD45RO (see Supplementary Figure 37).



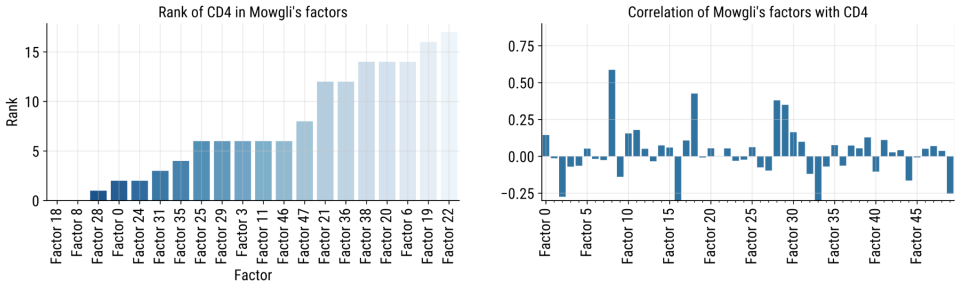
Supplementary Figure 37. Protein weights for Mowgli's factor 8. Top ADT weights for Mowgli's factor 8 in the TEA dataset.

High weights for Mowgli's factor 8 coincide with high CD4, high CD45RA, and low CD45RO (see Supplementary Figure 38).



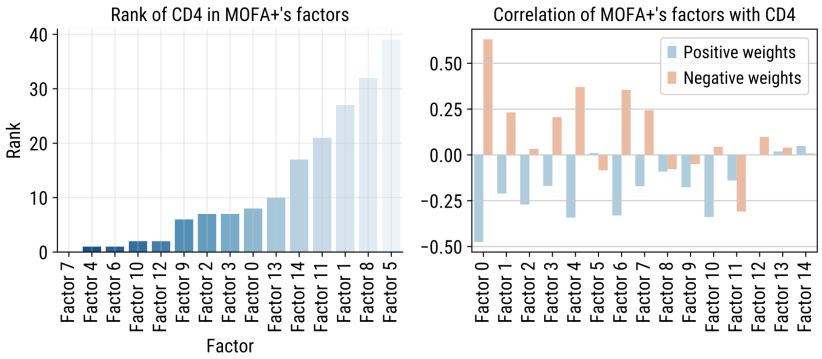
Supplementary Figure 38. Compared ADT profiles and Mowgli factor 8. Compared values of CD4, CD45RA/CD45RO, and weight of Mowgli's factor 8.

CD4 ranks first among the top proteins of Mowgli's factor 8, and factor 8 also has the highest Pearson correlation with CD4 (see Supplementary Figure 39).

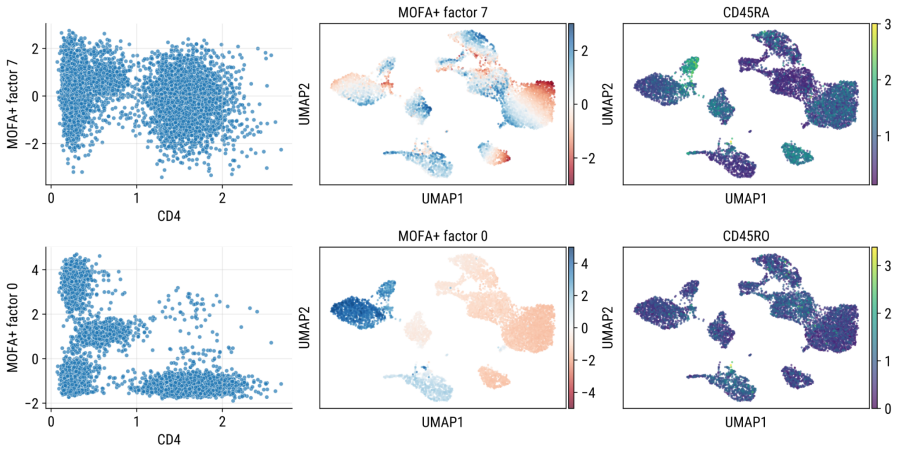


Supplementary Figure 39. Correlation between CD4 and Mowgli's factors. Compared rank of CD4 in Mowgli's factors, and correlation of CD4 with Mowgli's factors.

MOFA+'s factor with the lowest rank for CD4 (factor 7) does not correspond to a CD4 T cell subset. In particular, it does not allow to distinguish CD45RA+CD4+ T cells from CD45RO+CD4+ T cells. MOFA+'s factor most highly correlated with CD4 (factor 0, negative weights) coincides with the general T cell and NK cell populations instead (see Supplementary Figures 40, 41).



Supplementary Figure 40. Correlation between CD4 and MOFA+'s factors. Compared rank of CD4 in MOFA+'s factors, and correlation of CD4 with MOFA+'s factors.

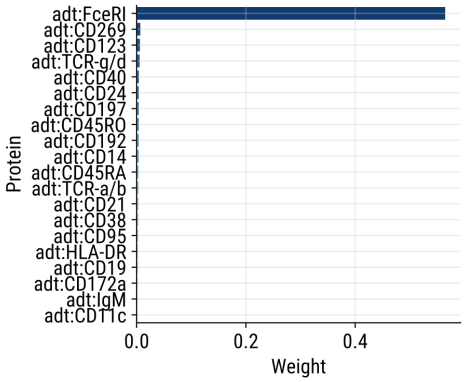


Supplementary Figure 41. Compared ADT profiles and MOFA+ factors 3 and 7. Compared values of CD4, CD45RA, CD45RO, and weights of MOFA+'s factors 0 and 7.

In conclusion, Mowgli's factor 8 allows one to distinguish naive CD4 T cells easily, which is not possible with a unique factor from MOFA+.

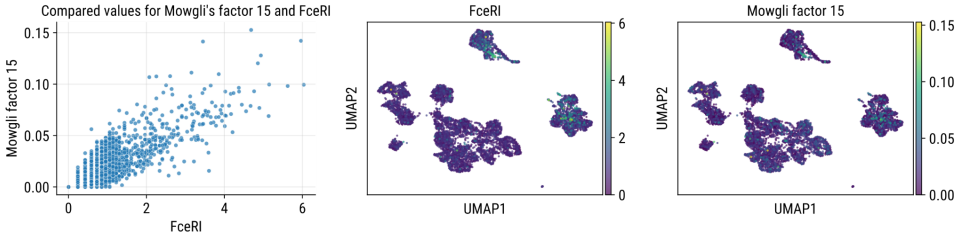
4. Conventional dendritic cells

Conventional dendritic cells (cDC) have a distinctive FceRI phenotype. Mowgli's factor 15 is largely driven by FceRI (see Supplementary Figure 42).



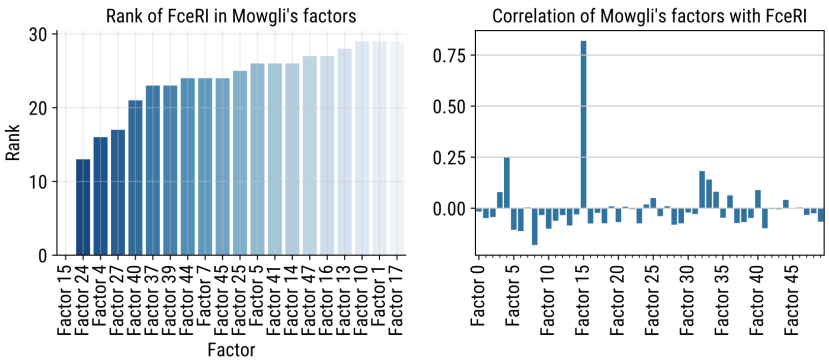
Supplementary Figure 42. Protein weights for Mowgli's factor 15. Top ADT weights for Mowgli's factor 15 in the TEA dataset.

High weights for Mowgli's factor 15 coincide with high FceRI (see Supplementary Figure 43).



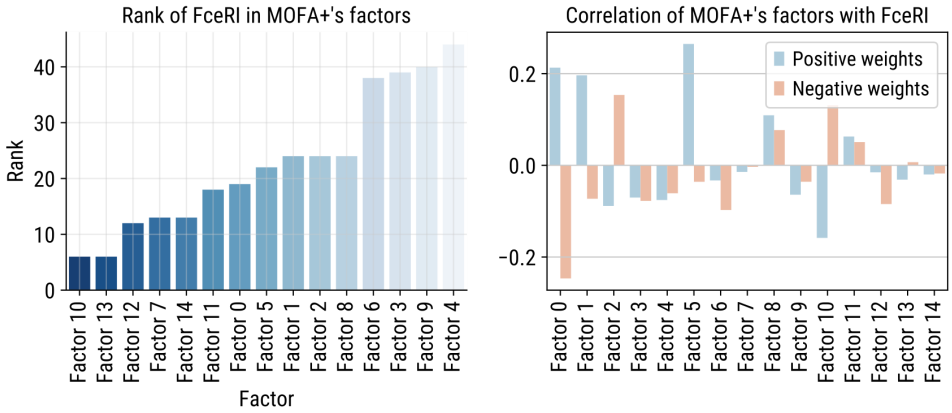
Supplementary Figure 43. Compared ADT profiles and Mowgli factor 15. Compared values of FceRI and weight of Mowgli's factor 15.

Mowgli's factor 15 is the only factor with a high Pearson correlation with FceRI. In addition, FceRI ranks first among factor 15's top proteins, and factor 15 is the only one with a low rank of FceRI (see Supplementary Figure 44).



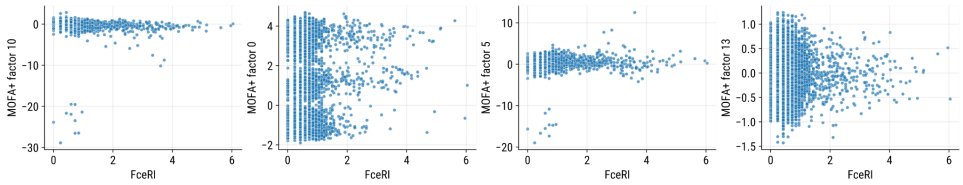
Supplementary Figure 44. Correlation between FceRI and Mowgli's factors. Compared rank of FceRI in Mowgli's factors, and correlation of FceRI with Mowgli's factors.

By contrast, MOFA+ has no factors highly correlated with FceRI or with a low rank of FceRI (see Figure Supplementary Figure 45).



Supplementary Figure 45. Correlation between FceRI and MOFA+'s factors. Compared rank of FceRI in MOFA+'s factors, and correlation of FceRI with MOFA+'s factors.

Even for MOFA+'s factors with the lowest rank for FceRI (factors 10 and 13) or the highest correlation with FceRI (factors 0 and 5), a scatterplot shows no pattern of association between the factor and the cDC phenotype (see Supplementary Figure 46).

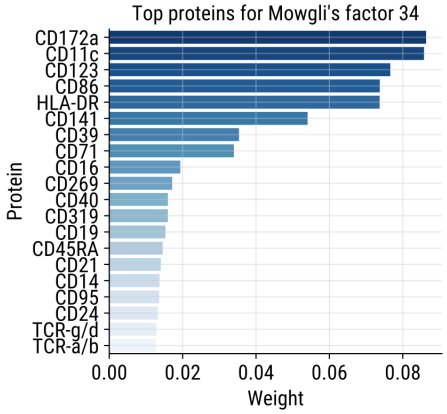


Supplementary Figure 46. Compared ADT profiles and MOFA+ factors 0, 5, 10 and 13. Compared values of FceRI and weights of MOFA+'s factors 0, 5, 10, and 13.

In conclusion, Mowgli's factor 15 allows distinguishing cDCs easily, which is impossible using a unique factor from MOFA+.

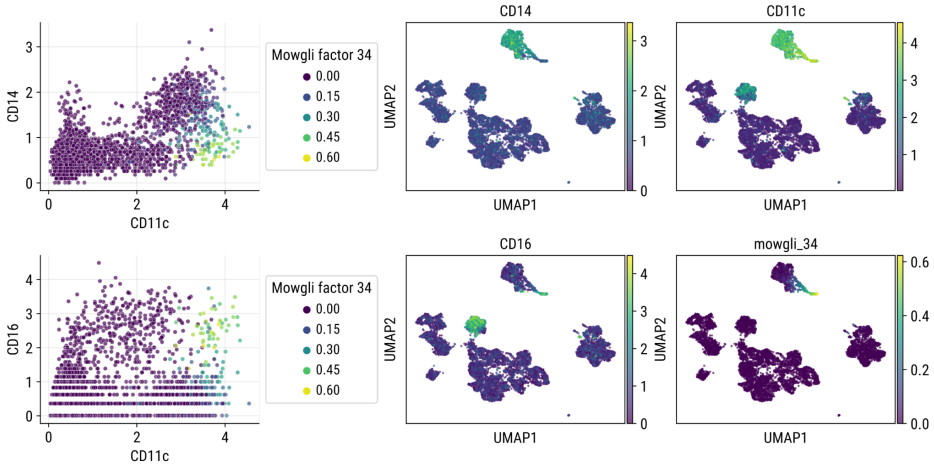
5. Nonclassical monocytes

Non-classical monocytes are CD14-low and CD16-high. Mowgli's factor 34 is driven by CD172a and CD11c, proteins present on the surface of monocytes, and has higher weights for CD16 than CD14 (see Supplementary Figure 47).



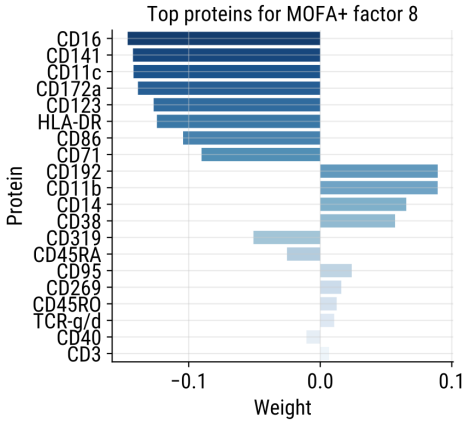
Supplementary Figure 47. Protein weights for Mowgli's factor 34. Top ADT weights for Mowgli's factor 34 in the TEA dataset.

High weights for Mowgli's factor 34 coincide with high CD11c, high CD16, and low CD14 (see Supplementary Figure 48).



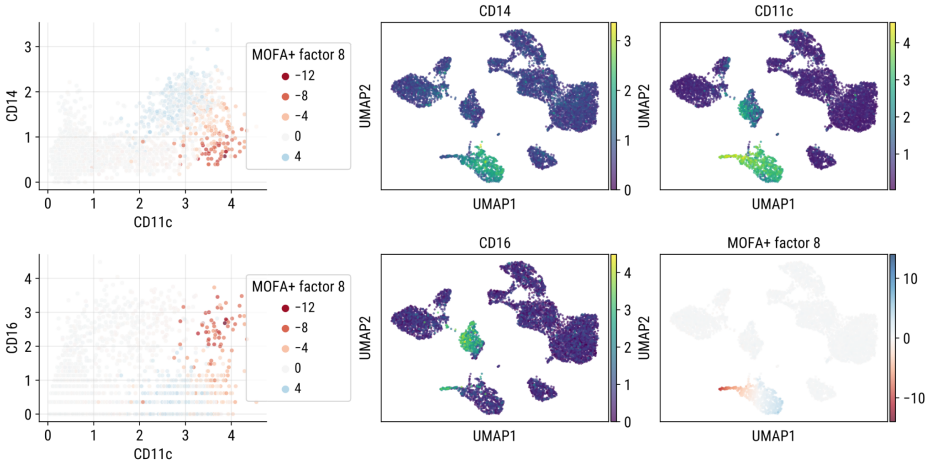
Supplementary Figure 48. Compared ADT profiles and Mowgli factor 34. Compared values of CD14, CD16, CD11c, and weight of Mowgli's factor 34.

MOFA+'s factor 8 has CD16 and CD11c among the top proteins, and CD14's contribution has an opposite sign to CD16 (see Supplementary Figure 49).



Supplementary Figure 49. Protein weights for MOFA+'s factor 8. Top absolute ADT weights for MOFA+'s factor 8.

Negative weights for MOFA+'s factor 8 coincide with high CD11c, high CD16, and low CD14 (see Supplementary Figure 50).

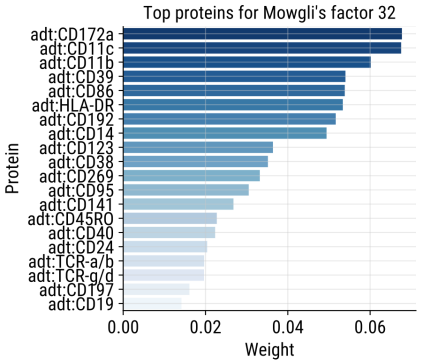


Supplementary Figure 50. Compared ADT profiles and MOFA+ factor 8. Compared values of CD14, CD16, CD11c, and weights of MOFA+'s factor 8.

In conclusion, both Mowgli's factor 13 and MOFA+'s factor 8 allow one to distinguish nonclassical monocytes from classical monocytes.

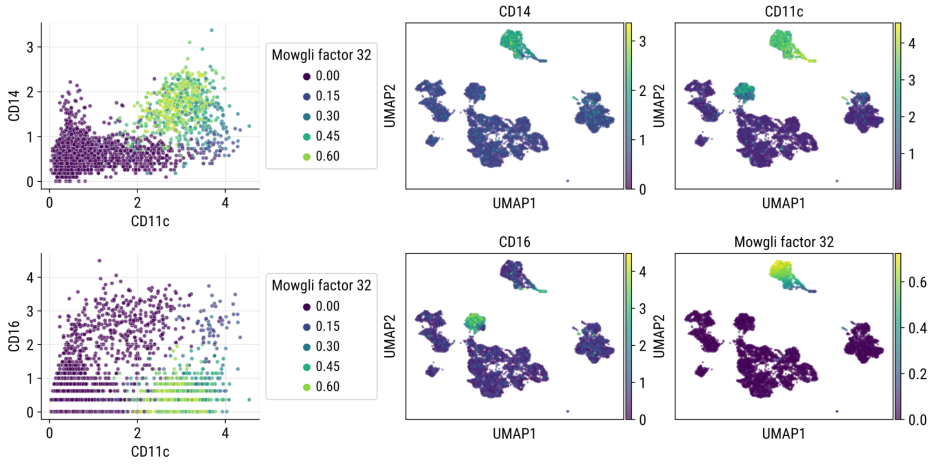
6. Classical monocytes

Classical monocytes are CD14-positive and CD16-negative. Mowgli's factor 32 is driven by CD172a, CD11b, and CD11c (proteins present on the surface of classical monocytes) and has higher weights for CD14 than CD16 (see Supplementary Figure 51).



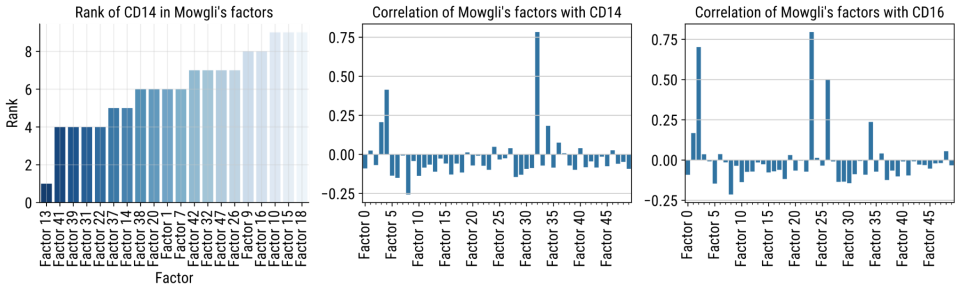
Supplementary Figure 51. Protein weights for Mowgli's factor 32. Top ADT weights for Mowgli's factor 32 in the TEA dataset.

High weights for Mowgli's factor 32 coincide with high CD11c, high CD14, and low CD16 (see Supplementary Figure 52).



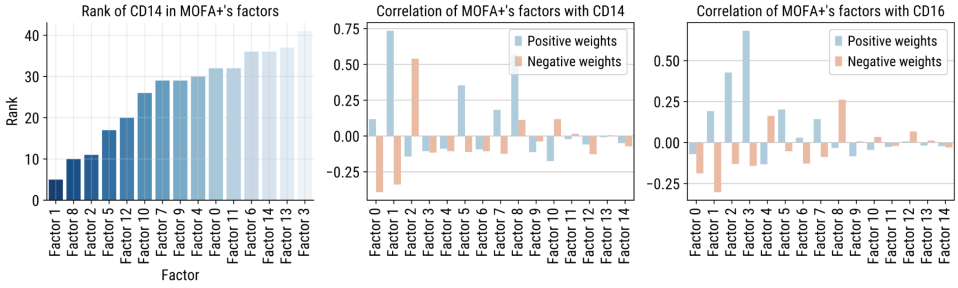
Supplementary Figure 52. Compared ADT profiles and Mowgli factor 32. Compared values of CD14, CD16, CD11c, and weight of Mowgli's factor 32.

CD14 ranks fifth among the top proteins for Mowgli's factor 32, and factor 32 has the highest Pearson correlation with CD14. In addition, factor 32 has one of the lowest absolute Pearson correlations with CD16 (see Supplementary Figure 53).



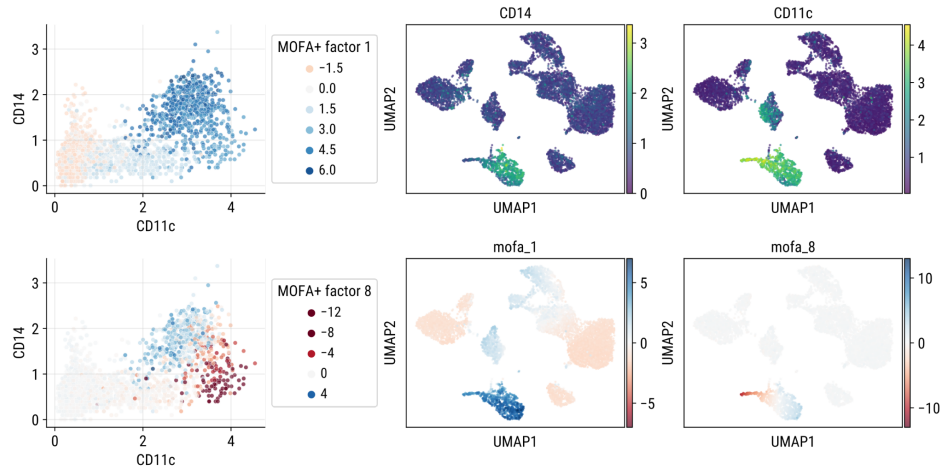
Supplementary Figure 53. Correlation between CD14, CD16, and Mowgli's factors. Compared rank of CD14 in Mowgli's factors, and correlation of CD14 and CD16 with Mowgli's factors.

MOFA+'s factors 1 and 8 have the lowest rank for CD14. They are the most highly correlated with CD14 and have a low correlation with CD16 (see Supplementary Figure 54).



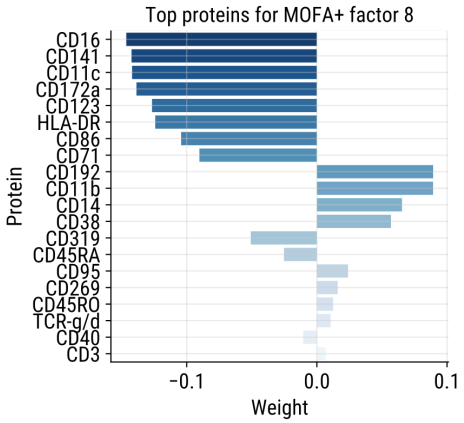
Supplementary Figure 54. Correlation between CD14, CD16, and MOFA+'s factors. Compared rank of CD14 in MOFA+'s factors, and correlation of CD14 and CD16 with MOFA+'s factors.

Positive weights for MOFA+'s factor 1 coincide with the broad population of monocytes. Positive weights for MOFA+'s factor 8 coincide with classical monocytes (see Supplementary Figure 55). Note that as described previously, negative weights for factor 8 coincide with nonclassical monocytes.



Supplementary Figure 55. Compared ADT profiles and MOFA+ factors 1 and 8. Compared values of CD14, CD11c, and weights of MOFA+'s factors 1 and 8.

While factor 8 correctly splits nonclassical and classical monocytes, positive weights for MOFA+'s factor 8 do not reflect the biological signature of classical monocytes. Indeed, factor 8 has negative weights for CD11c and CD172a (see Supplementary Figure 56), even though these proteins are present on the surface of both classical and nonclassical monocytes.



Supplementary Figure 56. Protein weights for MOFA+'s factor 8. Top absolute ADT weights for MOFA+'s factor 8.

In conclusion, MOFA+'s factor 8 allows one to distinguish classical monocytes from nonclassical monocytes, but Mowgli's factor 32 is more clearly associated with the biological signature of classical monocytes than MOFA+'s factor 8.

Supplementary References

1. Liu, L. *et al.* Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Nat. Commun. **10**, 470 (2019).
2. Luecken, M. *et al.* A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. in Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (eds. Vanschoren, J. & Yeung, S.) vol. 1 (2021).
3. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. Cell **177**, 1888-1902.e21 (2019).
4. Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. eLife **10**, e63632 (2021).
5. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. Cell **184**, 3573-3587.e29 (2021).
6. Collin, M. *et al.* Human dendritic cell subsets. Immunology **140**(1), 22-30 (2013).

Appendix D

Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein

D.1 Mouse proliferation gene set

The following list of genes was retrieved from https://moscot.readthedocs.io under `utils.data.proliferation_markers`. The original source for this genes is [Tirosh et al., 2016b].

Mcm4	Ccnb2	Gmnn	Wdr76	Gins2	Cdc45	Nuf2
Smc4	Cenpa	Cdc6	Ung	Cdc20	Ckap5	Anln
Gtse1	Cenpe	Pold3	Hn1	Tpx2	Ctcf	Nasp
Ttk	Cdca8	Ckap2l	Cks2	Rrm1	Clspn	Hmmr
Rangap1	Ckap2	Fam64a	Kif20b	Cdc25c	Chaf1b	Rfc2
Top2a	Rad51	Ubr7	Cdk1	Cdca3	Tacc3	Prim1
Hmgb2	Pcna	Fen1	Slbp	Ect2	Mcm5	Uhrf1
Ccne2	Ube2c	Bub1	Aurkb	Pola1	Anp32e	
G2e3	Lbr	Brip1	Kif11	Ndc80	Casp8ap2	
Tmpo	Cenpf	Atad2	Cks1b	Rpa2	Tubb4b	
Nusap1	Birc5	Psrc1	Blm	Nek2	Kif23	
Ncapd2	Dtl	Rrm2	Msh2	Mki67	Exo1	
Mcm2	Dscc1	Tipin	Gas2l3	Mcm6	Tyms	
Kif2c	Cbx5	E2f8	Rad51ap1	Cdca7	Hjurp	
Aurka	Mlf1ip	Cdca2	Usp1	Dlgap5	Hells	

D.2 Mouse apoptosis markers

The following list of genes was retrieved from https://moscot.readthedocs.io under `utils.data.apoptosis_markers`. The original source for this genes is the `HALLMARK_P53_PATHWAY` pathway from https://www.gsea-msigdb.org/.

Ercc5	Serpinb5	Krt17	Pmm1	Sphk1	Cdkn2a	Nhlh2
Pcna	Bmp2	Inhbb	Steap3	Dgka	Tpd52l1	Def6

Trib3	Procr	Btg2	Phlda3	Cdkn2aip	Sesn1	Cdkn2b
Blcap	Ada	Tnni1	Rgs16	Hmox1	Foxo3	Jun
Fgf13	Irak1	Ier5	Slc19a2	Rrad	Ddit4	Slc35d1
Tspyl2	Sat1	Adck3	Ephx1	Cdh13	Zfp365	Plk3
Zmat3	Hspa4l	Ptpn14	Atf3	Vwa5a	Osgin1	Rnf19b
Slc7a11	Tm4sf1	Notch1	Rxra	Zbtb16	Cgrrf1	Fuca1
Rap2b	Fbxw7	Ralgds	Ak1	Rps27l	Abhd4	Wrap73
S100a4	S100a10	Stom	Ddb2	Mapkapk3	Kif13b	Rchy1
Triap1	Prkab1	Cd82	Il1a	Traf4	Hint1	Ip6k2
Lrmp	Tm7sf3	Trafd1	Pom121	Tgfb1	Sertad3	Pdgfa
Cebpa	Klk8	Vamp8	Retsat	Bax	Ppp1r15a	Tprkb
Rpl18	Aen	Mxd1	Sec61a1	Rrp8	Ccp110	Xpc
Prmt2	Nupr1	H2afj	Ldhb	Btg1	Ctsd	Mknk2
Mdm2	Cd81	Dram1	Hras	Ddit3	Perp	Apaf1
Gls2	Rps12	Ikbkap	Txnip	Cdk5r1	Gm2a	Tcn2
Ppm1d	Hist3h2a	Lif	Tsc22d1	Rad51c	Alox8	Upp1
Tob1	Trp53	Ccng1	St14	Hexim1	Fdxr	Cyfip2
Plk2	Sdc1	Tax1bp3	Gnb2l1	Ccnk	Slc3a2	Gpx2
Jag2	Fas	Zfp36l1	Rad9a	Ndrg1	Itgb4	Fos
Plxnb2	Rhbdf2	Vdr	Baiap2	Csrnp2	Dcxr	Acvr1b
Sp1	Ninj1	Abat	Nol8	Socs1	F2r	Abcc5
Cdkn1a	Trp63	Tap1	Fam162a	Ier3	App	Polh
Dnttip2	Clca2	Wwp1	Klf4	Ankra2	Ei24	Hbegf
Sfn	Epha2	Mxd4	Iscu	Ccnd3	Hdac3	
Rb1	Gadd45a	Tgfa	Ccnd2	Rab40c	Ctsf	
Ptpre	Eps8l2	Nudt15	Casp1	Bak1	Hist1h1c	

D.3 Axolotl neural stem cell gene set

The following list of genes was retrieved from [Wei et al., 2022].

AMEX60DD001640	AMEX60DD002950	AMEX60DD005040	AMEX60DD009861	AMEX60DD011133
AMEX60DD016851	AMEX60DD022108	AMEX60DD029807	AMEX60DD031236	AMEX60DD033253
AMEX60DD034100	AMEX60DD036933	AMEX60DD038483	AMEX60DD039579	AMEX60DD041912
AMEX60DD043225	AMEX60DD043789	AMEX60DD044011	AMEX60DD044899	AMEX60DD045542
AMEX60DD045792	AMEX60DD049907	AMEX60DD050822	AMEX60DD051580	AMEX60DD052006
AMEX60DD055073	AMEX60DDU001006420			

D.4 Human apoptosis gene set

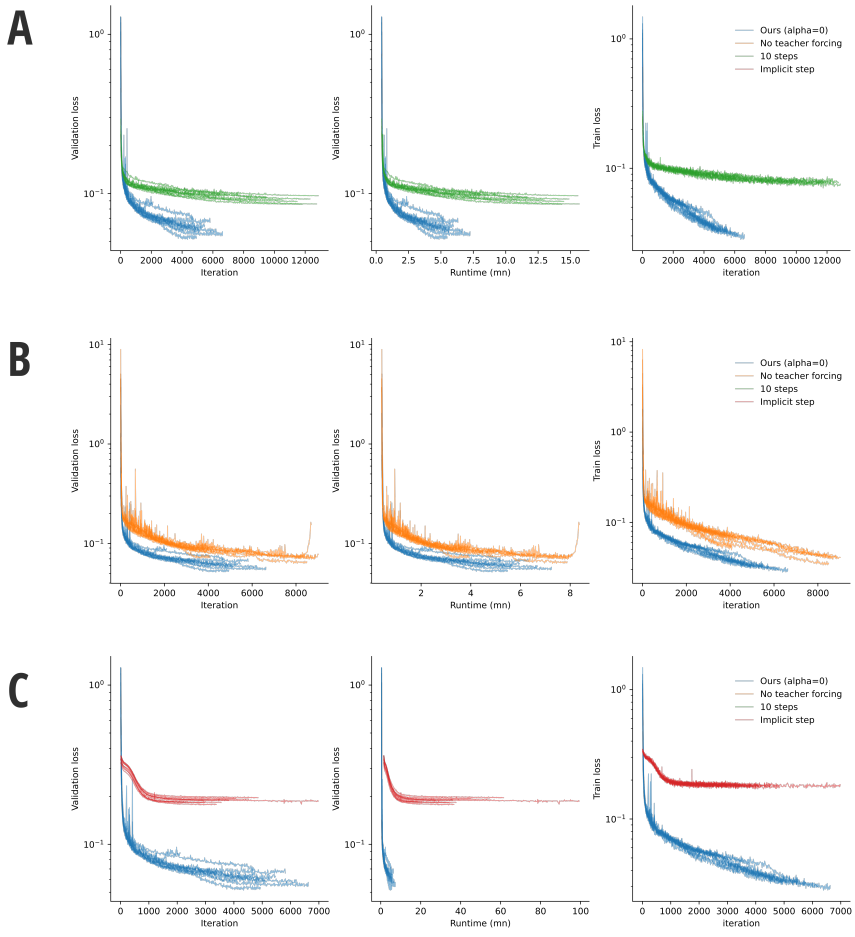
The following list of genes was retrieved from https://moscot.readthedocs.io under `utils.data.apoptosis_markers`. The original source for this genes is the `HALLMARK_APOPTOSIS` pathway from https://www.gsea-msigdb.org/.

ADD1	AIFM3	ANKH	ANXA1	APP	ATF3	AVPR1A	BAX	CASP2	TNFSF10
BCAP31	BCL10	BCL2L1	BCL2L10	BCL2L11	BCL2L2	BGN	BID	CASP3	TOP2A
BIK	BIRC3	BMF	BMP2	BNIP3L	BRCA1	BTG2	BTG3	CASP1	TSPO
CASP4	CASP6	CASP7	CASP8	CASP9	CAV1	CCNA1	CCND1	CCND2	TXNIP

CD14	CD2	CD38	CD44	CD69	CDC25B	CDK2	CDKN1A	SOD2	VDAC2
CDKN1B	CFLAR	CLU	CREBBP	CTH	CTNNB1	CYLD	DAP	DAP3	WEE1
DCN	DDIT3	DFFA	DIABLO	DNAJA1	DNAJC3	DNM1L	DPYD	EBP	XIAP
EGR3	EMP1	ENO2	ERBB2	ERBB3	EREG	ETF1	F2	F2R	FAS
FASLG	FDXR	FEZ1	GADD45A	GADD45B	GCH1	GNA15	GPX1	GPX3	
GPX4	GSN	GSR	GSTM1	GUCY2D	H1-0	HGF	KRT18	LEF1	
HMGB2	HMOX1	HSPB1	IER3	IFITM3	IFNB1	IFNGR1	LGALS3	LMNA	
IGF2R	IGFBP6	IL18	IL1A	IL1B	IL6	IRF1	ISG20	JUN	
LUM	MADD	MCL1	MGMT	MMP2	NEDD9	NEFH	PAK1	PDCD4	
PDGFRB	PEA15	PLAT	PLCB2	PLPPR4	PMAIP1	PPP2R5B	RETSAT	RHOB	
RHOT2	RNASEL	ROCK1	SAT1	SATB1	SC5D	SLC20A1	SMAD7	SOD1	
PPP3R1	PPT1	PRF1	PSEN1	PSEN2	PTK2	RARA	RELA	SPTAN1	
SQSTM1	TAP1	TGFB2	TGFBR3	TIMP1	TIMP2	TIMP3	TNF	TNFRSF12A	

D.5 Supplementary figure 1

Figure D.1: Compared losses during training on the zebrafish atlas, for different choices of discretization.



(A) Training and validation losses along iterations and runtime, for the linear method (blue) and a version with 10 forward steps (green), as in [Hashimoto et al., 2016]; **(B)** Training and validation losses along iterations and runtime, for the linear method (blue) and a version with without teacher forcing (orange), as in [Hashimoto et al., 2016]; **(C)** Training and validation losses along iterations and runtime, for the linear method (blue) and a version with an ICNN implicit step (red), as in [Bunne et al., 2022].

RÉSUMÉ

Caractériser la diversité cellulaire (c'est-à-dire les types et les états cellulaires) et les mécanismes qui conduisent aux transitions entre les états cellulaires est un objectif de longue date de la biologie cellulaire, rendu de plus en plus atteignable par les avancées des techniques de séquençage omique à haut débit sur cellule unique. Ces technologies permettent un profilage moléculaire quantitatif à une résolution sans précédent, facilitant la découverte non biaisée de types de cellules auparavant inconnus à travers les tissus et l'élucidation des trajectoires de différenciation entre les états cellulaires en santé et en maladie. Récemment, la transcriptomique spatiale et le séquençage multi-omique pairé sur cellule unique ont permis le profilage conjoint de modalités complémentaires, offrant des reseignements sur les motifs spatiaux et les mécanismes de régulation. Cependant, l'analyse des données multimodales pairées est difficile en raison de la haute dimensionnalité, du bruit et des vastes différences entre les modalités.

Une stratégie prometteuse pour relever les défis de l'intégration des données multimodales pairées sur cellule unique réside dans le Transport Optimal (OT), une méthode géometriquement pertinente pour comparer des distributions de probabilité. Des avancées récentes ont fait de cette théorie séculaire en un outil puissant pour l'apprentissage automatique, particulièrement efficace dans l'analyse des données omiques sur cellule unique. Des applications fructueuses incluent la modélisation de la dynamique des populations et l'alignement des données à travers différentes modalités. Cependant, la plupart des applications existantes se concentrent sur de modalités *non pairées* plutôt que sur l'analyse conjointe de modalités *pairées*.

Cette thèse propose de nouvelles méthodes basées sur l'OT pour l'intégration multimodale pairée sur cellule unique et s'organise autour de deux axes de recherche principaux. Le premier axe se concentre sur la *réduction dimensionnelle conjointe* des données multi-omiques pairées sur cellule unique. Contrairement aux méthodes existantes, cette thèse considère la cellule comme une distribution discrète de caractéristiques moléculaires (telles que des gènes, des régions génomiques ou des protéines). L'OT fournit alors un moyen naturel de comparer les cellules en tirant parti de la géométrie unique des caractéristiques de chaque couche omique tout en atténuant le bruit inhérent aux données sur cellule unique. Cette nouvelle notion de similarité entre les cellules est utilisée comme une fonction d'objectif en apprentissage automatique pour intégrer les cellules dans une représentation de faible dimension interprétable.

Le deuxième axe se concentre sur l'*inférence de trajectoire* utilisant la transcriptomique spatiale. Cette thèse propose un modèle causal de différenciation cellulaire qui formalise le paysage épigénétique de Waddington par un réseau de neurones entraîné sur des données d'expression profilées dans l'espace et le temps. La méthode proposée exploite une extension quadratique de l'OT pour imposer une invariance rotationnelle et translationnelle aux coordonnées spatiales. Par opposition, les méthodes existantes ne peuvent pas tirer parti de l'information spatiale puisqu'elles reposent sur la formulation originale de l'OT, sensible ces isométries.

Les contributions méthodologiques de cette thèse sont évaluées de manière approfondie dans divers contextes biologiques tels que le système immunitaire et le développement, démontrant que les méthodes développées fournissent des reseignements biologiques précieux à travers les modalités.

MOTS CLÉS

multimodal; unicellulaire; intégration; méthodes; transport optimal

ABSTRACT

Characterizing cellular diversity (i.e., cell types and cell states) and the mechanisms driving transitions between cell states is a long-standing goal of cell biology, made increasingly attainable by advances in high-throughput single-cell omics sequencing techniques. These technologies deliver quantitative molecular profiling at unprecedented resolution, facilitating the unbiased discovery of previously unknown cell types across tissues and elucidating differentiation trajectories between cell states in health and disease. Recently, spatial transcriptomics and paired single-cell multi-omics sequencing have enabled the joint profiling of complementary modalities, yielding insights into spatial patterns and regulatory mechanisms. However, analyzing paired multimodal data is challenging due to high dimensionality, noise, and the vast differences between modalities.

A promising strategy to address the challenges of paired single-cell multimodal data integration lies in Optimal Transport (OT), a geometrically meaningful way to compare probability distributions. Recent breakthroughs have reframed this centuries-old theory as a powerful tool for machine learning, which has proven particularly effective in analyzing single-cell omics data. Successful applications include modeling population dynamics and aligning data across different modalities. However, most existing applications focus on *unpaired* modalities rather than the joint analysis of *paired* modalities.

This thesis proposes novel OT-based methods for paired single-cell multimodal integration and is organized along two main research directions. The first direction focuses on *joint dimensionality reduction* of paired single-cell multi-omics data. Unlike existing methods, this thesis views the cell as a discrete distribution of molecular features (such as genes, genomic regions, or proteins). OT then provides a natural way to compare cells by leveraging the unique geometry underlying the features in each omics layer while mitigating the noise inherent to single-cell data. This new notion of similarity between cells is used as a machine learning objective to embed cells into an interpretable low-dimensional representation.

The second direction focuses on *trajectory inference* using spatial transcriptomics. This thesis proposes a causal model of cellular differentiation that formalizes Waddington's epigenetic landscape as a neural network trained on gene expression data profiled through space and time. The proposed method leverages a quadratic extension of OT to impose rotational and translational invariance for spatial coordinates. In contrast, existing methods cannot leverage spatial information since they rely on the original formulation of OT, which is sensitive to such transformations.

This thesis's methodological contributions are evaluated thoroughly in diverse biological settings such as the immune system and development, demonstrating that the developed methods provide valuable biological insights across modalities.

KEYWORDS

multimodal; single-cell; integration; methods; optimal transport