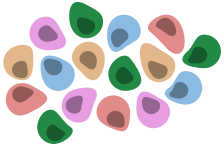
Development of methods for paired single-cell multimodal integration

Geert-Jan Huizing

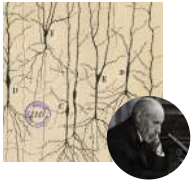
Thesis defense, September 27th, 2024

Ecole doctorale 515. Thesis directors: Laura Cantini & Gabriel Peyré

Uncovering cellular heterogeneity



~10^{13} cells in the human body, with vastly different functions.

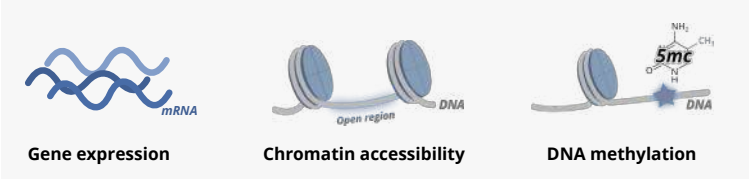


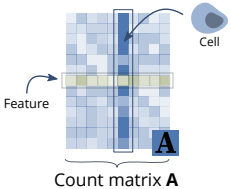
Early efforts to cartography cell identity relied on microscopy[1].

Recent initiatives measure the molecular profile of the cell[2].

[1] Ramón y Cajal, 1899; [2] Regev et al., *eLife*, 2017

Single-cell omics sequencing

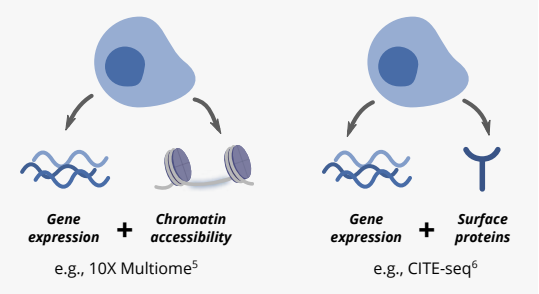




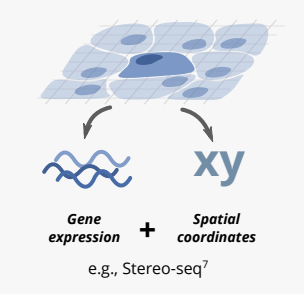
Single-cell sequencing technologies[3,4] deliver quantitative *omics* information as a count matrix.

[3] Nawy, *Nature methods*, 2014; [4] Preissl et al., *Nature Reviews Genetics*, 2023

Multimodal omics sequencing



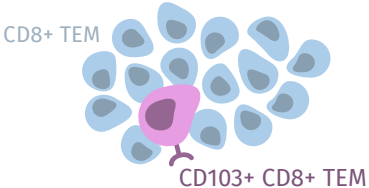
Single-cell multi-omics measure the cell at several molecular layers.



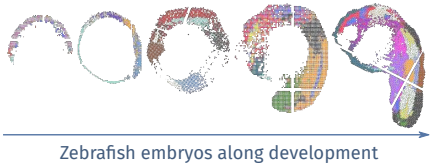
Spatial omics measure the cell without dissociating the tissue.

[5] 10X Genomics; [6] Stoeckius et al., *Nature methods*, 2017; [7] Chen et al., *Cell*, 2022

Applications of multimodal omics



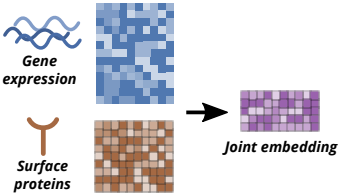
The joint profiling of **gene expression and surface proteins** enabled to identify a new subpopulation of CD8 TEM cells[8].



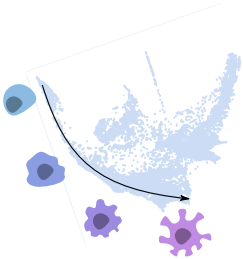
Spatial transcriptomics profiled across time have allowed to study development at unprecedented resolution[9].

[8] Hao et al., *Cell*, 2021; [9] Liu et al., *Developmental Cell*, 2022

Paired single-cell multimodal integration



Joint dimensionality reduction can discover patterns across modalities and identify cell subpopulations[10,11].

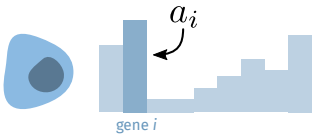


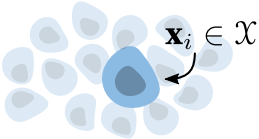
Trajectory inference, which can discover trends along a dynamic process such as development, aging or regeneration[12,13].

[10] Argelaguet et al., *Genome biology*, 2020; [11] Lotfollahi et al., *bioRxiv*, 2022; [12] Li et al., *Nature biotechnology*, 2023; [13] Klein et al., *bioRxiv*, 2023

Formalizing cells and populations

Unifying notation across this presentation: $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ with $\mathbf{a} \in \Sigma_n$ and $\mathbf{x}_i \in \mathcal{X}$





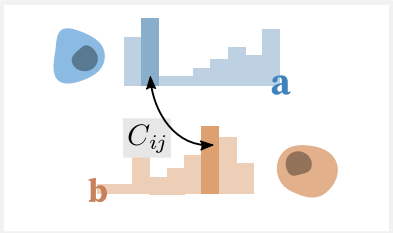
$$\mu = \sum_i a_i \delta_{\mathbf{x}_i} \text{ with } \underbrace{\mathbf{a} \in \Sigma_n}_{\text{cell}} \text{ and } \underbrace{\mathbf{x}_i \in \mathcal{X}}_{\text{gene}}$$

We formalize the cell as a histogram over the space of molecular features,

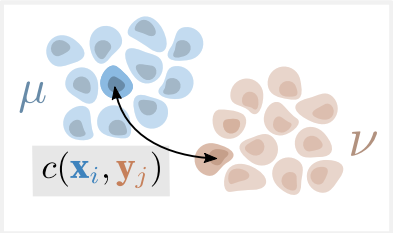
$$\mu = \sum_i a_i \delta_{\mathbf{x}_i} \text{ with } \mathbf{a} \in \Sigma_n \text{ and } \underbrace{\mathbf{x}_i \in \mathcal{X}}_{\text{cell}}$$

and populations of cells as a point cloud over some Euclidean space $\mathcal{X}$.

Optimal Transport compares distributions



Eulerian setting: $\mathrm{W}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$

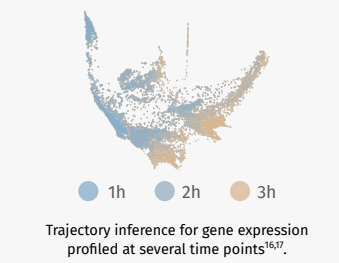


Lagrangian setting: $\mathcal{W}_c(\mu, \nu)$

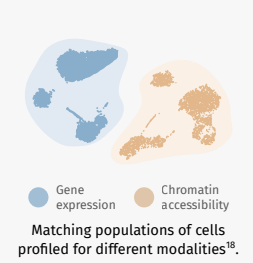
Optimal Transport[14,15] is a mathematical framework to compare probability distributions.

[14] Monge, *Mem. Math. Phys. Acad. Royale Sci.*, 1781; [15] Kantorovich, *Doklady Akademii Nauk*, 1942

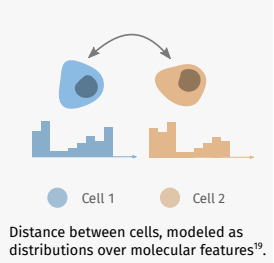
Applications of Optimal Transport in single-cell data



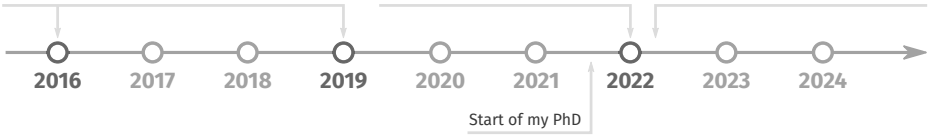
Trajectory inference for gene expression profiled at several time points[16,17].



Matching populations of cells profiled for different modalities[18].

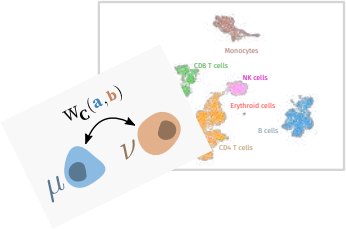


Distance between cells, modeled as distributions over molecular features[19].

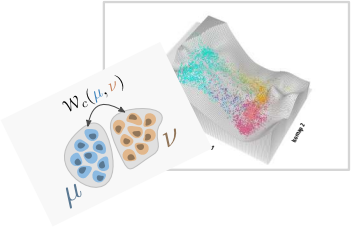


[16] Hashimoto et al., *ICML*, 2016; [17] Schiebinger et al., *Cell*, 2019; [18] Demetci et al., *Journal of Comp. Biology*, 2022; [19] Huizing et al., *Bioinformatics*, 2022

Aims of this thesis



Leverage a cost between features for **joint dimensionality reduction** of single-cell **multiomics.**

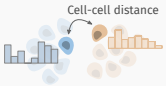


Leverage spatial information in **trajectory inference** with **spatial transcriptomics** through time.

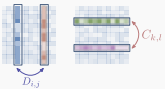
Contributions of this thesis

Eulerian setting

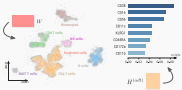
Lagrangian setting



I. *OT as a cell-cell distance*



II. *Wasserstein Singular Vectors*

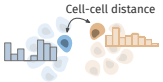


III. *OT-based joint dimensionality reduction*

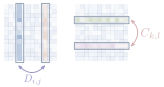
IV. *OT-based trajectory inference in space*

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

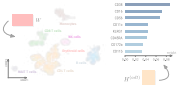
Work published as: G.-J. Huizing, G. Peyré, L. Cantini, *Bioinformatics*, 2022



I. *OT as a cell-cell distance*



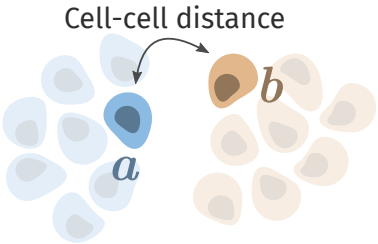
II. *Wasserstein Singular Vectors*



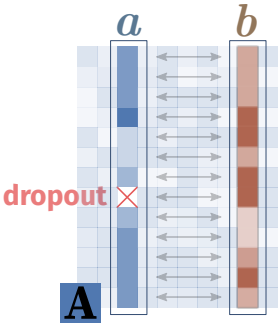
III. *OT-based joint dimensionality reduction*

IV. *OT-based trajectory inference in space*

Distances between cells

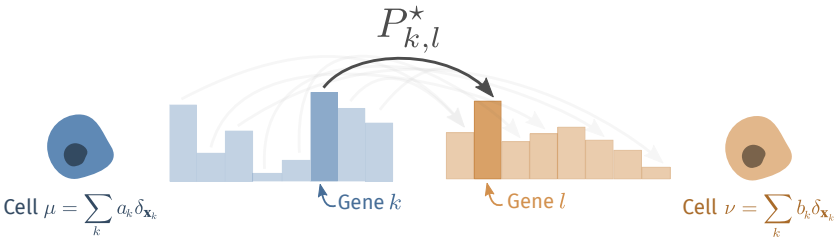


Distances between cells are crucial for downstream tasks (e.g., clustering).



Bin-bin distances like the Euclidean distance compare genes one by one.

Optimal Transport distances between cells

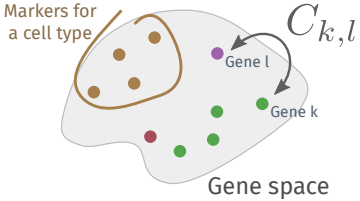


$$\mathrm{W}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathcal{U}(\mathbf{a},\mathbf{b})} \sum_{k,l} C_{k,l}\, P_{k,l} \quad \text{with } \mathcal{U}(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \{\mathbf{P}\in\mathbb{R}_+^{m\times m},\ \mathbf{P}\mathbf{1}_m = \mathbf{a}, \mathbf{P}^\top \mathbf{1}_m = \mathbf{b}\}$$

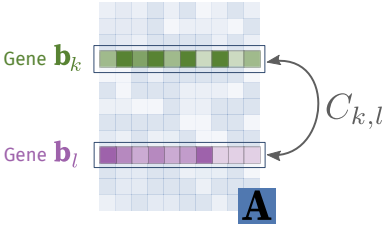
Optimal Transport[20,21] defines a distance between distributions as the cost of transporting mass from one to the other.

[20] Monge, *Mem. Math. Phys. Acad. Royale Sci.*, 1781; [21] Kantorovich, *Doklady Akademii Nauk*, 1942

Ground cost for single-cell omics



Similar features, like markers for the same cell type, should have a low cost.



$$1-\cos(\mathbf{b}_k,\mathbf{b}_l) = 1-\frac{\mathbf{b}_k^\top \mathbf{b}_l}{\|\mathbf{b}_k\|_2\|\mathbf{b}_l\|_2}$$

Here, we consider distances between rows of the cost matrix, but alternatives exist[24,25].

[24] Bellazi et al., *arXiv*, 2021; [25] Huizing et al., *ICML*, 2022

Entropic regularization of Optimal Transport

Entropy-regularized Optimal Transport[22] is a fast approximation of the previous problem. It can be computed efficiently using the GPU-enabled Sinkhorn algorithm.

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

$$\text{where } \mathrm{E} : \mathbf{P} \mapsto -\textstyle\sum_{k,l} P_{k,l}(\log P_{k,l} - 1)$$

The **Sinkhorn divergence**[23] elimitates the bias introduced by the entropic regularization.

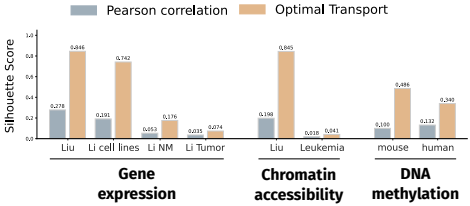
$$\overline{\mathrm{W}_{\mathbf{C}}^{\varepsilon}}(\mathbf{a},\mathbf{b}) \stackrel{\text{def.}}{=} \mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a},\mathbf{b}) - \tfrac{1}{2}\left(\mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{a},\mathbf{a}) + \mathrm{W}_{\mathbf{C}}^{\varepsilon}(\mathbf{b},\mathbf{b})\right).$$

[22] Cuturi, *NeurIPS*, 2013; [23] Genevay et al., *AISTATS*, 2018

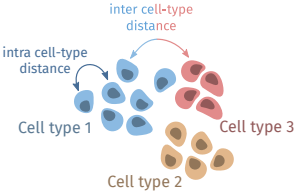
OT distances better reflect known heterogeneity

We consider several datasets across three omics annotated with cell types.

OT distances better reflect known heterogeneity

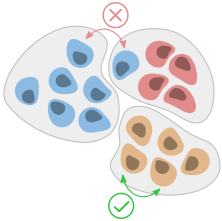


We consider several datasets across three omics annotated with cell types.



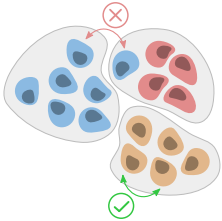
We compare OT to common bin-bin distances using the Silhouette score.

OT-based clustering outperforms the standard approach

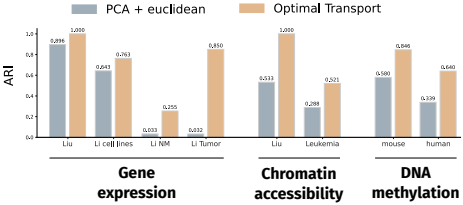


We also evaluate clustering performance using the Adjusted Rand Index.

OT-based clustering outperforms the standard approach

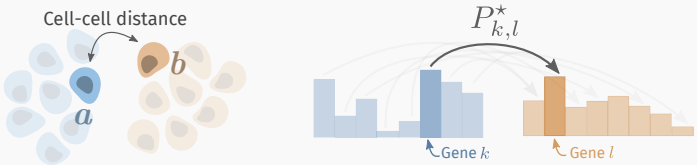


We also evaluate clustering performance using the Adjusted Rand Index.



We use Leiden clustering based on OT, and the standard Euclidean distance on PCA.

Conclusion



Optimal Transport as a distance between cells improves clustering performances.

cantinilab/OT-scOmics Public

This Python package will allow you to replicate the experiments from our research on applying Optimal Transport as a similarity metric in between single-cell omics data.

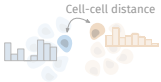
Python 36 7

```
pip install otscomics
```

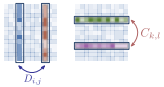
Open-source Python package

Unsupervised ground metric learning using Wasserstein Singular Vectors

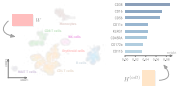
Work published as: G.-J. Huizing, L. Cantini, G. Peyré, *ICML*, 2022



I. *OT as a cell-cell distance*



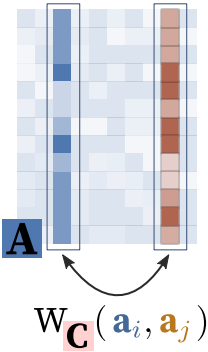
II. *Wasserstein Singular Vectors*



III. *OT-based joint dimensionality reduction*

IV. *OT-based trajectory inference in space*

Ground metric learning



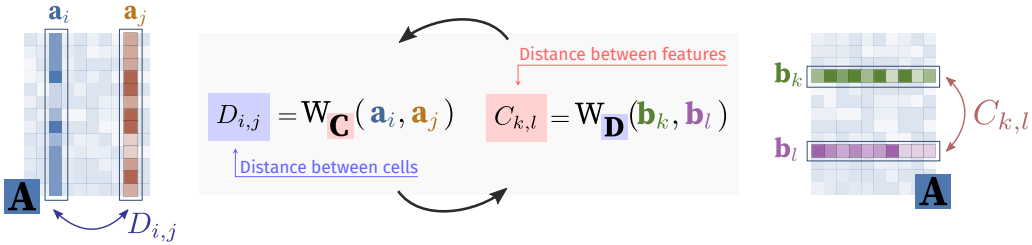
In the previous chapter, we computed cell-cell Optimal Transport distances using a cost $\mathbf{C}$.

In a supervised setting, it is possible to learn a ground cost that reflects labels[26,27].

We proposed the first **unsupervised** ground metric learning method.

[26] Cuturi and Avis, *JMLR*, 2014; [27] Wang and Guibas, *ECCV*, 2012

Bootstrapping intuition



The transposed problem defines a Wasserstein distance matrix between genes.

Wasserstein Singular Vectors

Let us define a Wasserstein distance map: $\Phi_{\mathbf{A}}(\mathbf{C})_{i,j} \stackrel{\text{def.}}{=} \mathrm{W}_{\mathbf{C}}(\mathbf{a}_i, \mathbf{a}_j)$

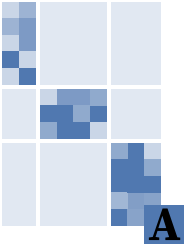
The fixed point condition for the previous bootstrapping algorithm is a **nonlinear singular vectors** problem:

Distance between cells

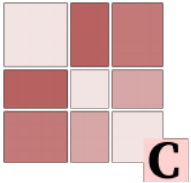
$$\Phi_{\mathbf{A}}(\mathbf{C}) = \gamma\,\mathbf{D}, \quad \Phi_{\mathbf{B}}(\mathbf{D}) = \lambda\,\mathbf{C} \quad \text{with } (\lambda, \gamma) \in (\mathbb{R}_+^*)^2 \qquad \textbf{(1)}$$

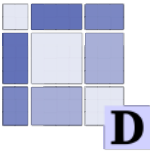
Distance between features

Wasserstein Singular Vectors of a block-diagonal matrix



Consider a block diagonal count matrix.





Pairs **C**,**D** of block constant and antidiagonal distance matrices are Wasserstein Singular Vectors.

Computing Wasserstein Singular Vectors

We now add a regularization term to the Wasserstein distance map:

$$\Phi_{\mathbf{A}}(\mathbf{C})_{i,j} \stackrel{\text{def.}}{=} \mathrm{W}_{\mathbf{C}}(\mathbf{a}_i, \mathbf{a}_j) + \tau \|\mathbf{a}_i - \mathbf{a}_j\|_1$$

Theorem. *When $\tau > 0$, there exist positive singular vectors $(\mathbf{C}, \mathbf{D})$ solving* (1).

To compute Wasserstein Singular Vectors, we can use the following power iterations:

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

In practice, power iterations converge even for $\tau = 0$.

Entropic regularization

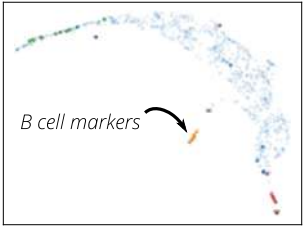
Sinkhorn divergence map: $\Phi^{\varepsilon}_{\mathbf{A}}(\mathbf{C})_{i,j} \overset{\text{def.}}{=} \overline{\mathrm{W}^{\varepsilon}_{\mathbf{C}}}(\mathbf{a}_i, \mathbf{a}_j) + \tau \| \mathbf{a}_i - \mathbf{a}_j \|_1$

$$\Phi_{\mathbf{A}} \xleftarrow{\varepsilon \to 0} \Phi^{\varepsilon}_{\mathbf{A}} \xrightarrow{\varepsilon \to \infty} \Phi^{\infty}_{\mathbf{A}}$$

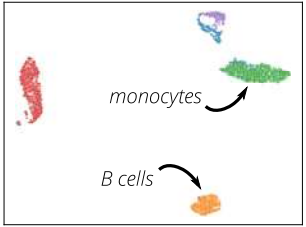
Wasserstein Singular Vectors	Sinkhorn Singular Vectors	In the limit case, connection with PCA

In the limit case, $\Phi^{\infty}_{\mathbf{A}} \overset{\text{def.}}{=} (-\frac{1}{2}\langle \mathbf{C}(\mathbf{a}_k - \mathbf{a}_l), \mathbf{a}_k - \mathbf{a}_l \rangle)_{k,l}$,
when **A** is bistochastic, we can explicit some singular vectors.

Sinkhorn Singular Vectors for single-cell gene expression



UMAP projection of gene-gene distances $\mathbf{C}_{k,l}$



UMAP projection of cell-cell distances $\mathbf{D}_{i,j}$

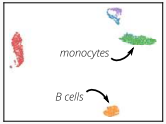
Silhouette score for cells

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**

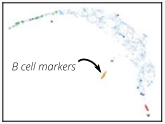
Silhouette score for marker genes

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

Conclusion



$$D_{i,j} = \frac{1}{\gamma}\mathrm{W}_{\mathbf{C}}(\mathbf{a}_i, \mathbf{a}_j), \quad C_{k,l} = \frac{1}{\lambda}\mathrm{W}_{\mathbf{D}}(\mathbf{b}_k, \mathbf{b}_l)$$



Unsupervised ground metric learning framed as a nonlinear singular vector problem.

CSDUlm/wsingular Public

Python package for the ICML 2022 paper "Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors".

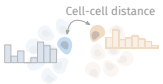
Python 9

```
pip install wsingular
```

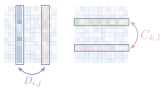
Open-source Python package

Paired single-cell multi-omics data integration with Mowgli

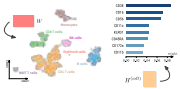
Work published as: G.-J. Huizing, I. M. Deutschmann, G. Peyré, L. Cantini, *Nature Comms*, 2023



I. *OT as a cell-cell distance*



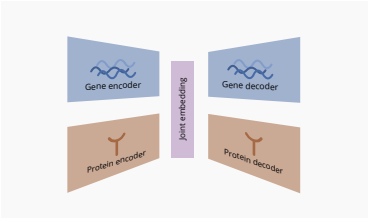
II. *Wasserstein Singular Vectors*



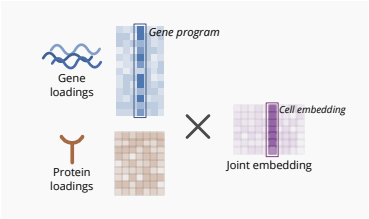
III. ***OT-based joint dimensionality reduction***

IV. *OT-based trajectory inference in space*

Joint dimensionality reduction methods



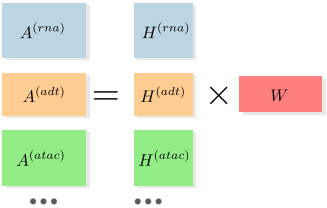
Deep learning methods[28,29] are expressive but lack interpretability.

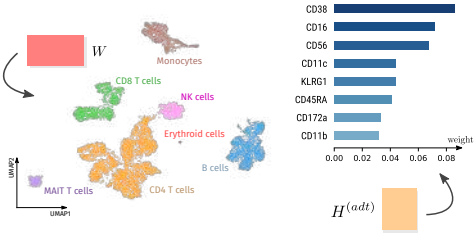


Matrix factorization methods[30,31] are less expressive but interpretable.

[28] Gong et al., *Genome biology*, 2021; [29] Lotfollahi et al., *bioRxiv*, 2022; [30] Argelaguet et al., *Genome biology*, 2020; [31] Chalise and Fridley, *PloS one*, 2017

Overview of our approach

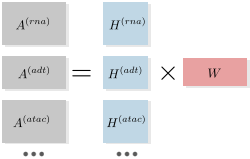




We developed **MOwgli**, a joint Nonnegative Matrix Factorization method with an OT objective.

The joint embedding allows clustering and visualization, and omics-specific loadings enable to interpret results.

Objective function



Sum over omics

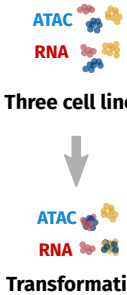
$$\min_{\mathbf{H}^{(p)},\mathbf{W}} \boxed{\sum_p} \left(\boxed{\sum_j \mathrm{W}^{\varepsilon}_{\mathbf{C}^{(p)}} \left(\mathbf{H}^{(p)} \mathbf{w}_j,\ \mathbf{a}_j^{(p)} \right)} - \alpha_p \boxed{\mathrm{E}\left(\mathbf{H}^{(p)}\right)} - \beta \boxed{\mathrm{E}\left(\mathbf{W}\right)} \right)$$

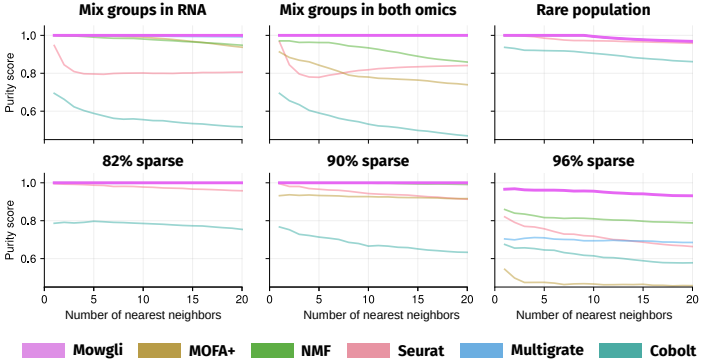
Reconstruction term Regularization terms

Our objective function extends **Wasserstein NMF**[32] to the multimodal setting.

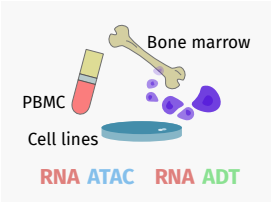
[32] Rolet et al., *AISTATS*, 2016

Improved embedding and clustering performance in controlled settings

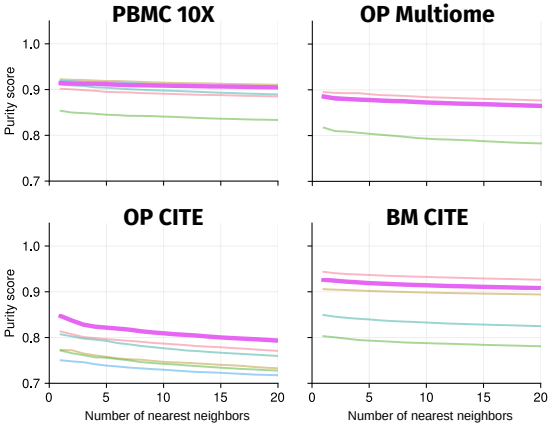




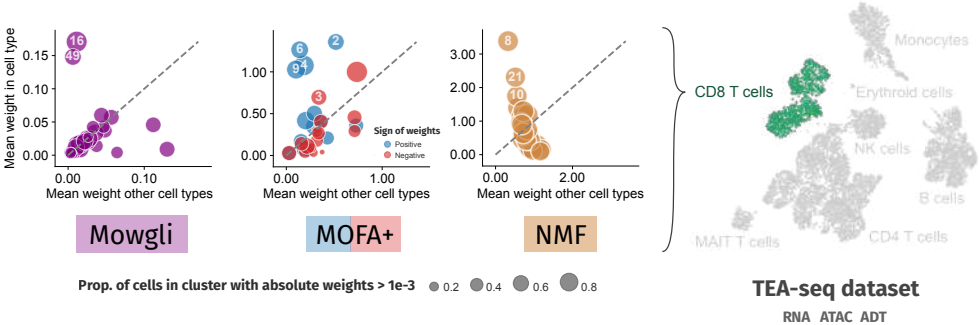
Competitive embedding and clustering performance in large datasets



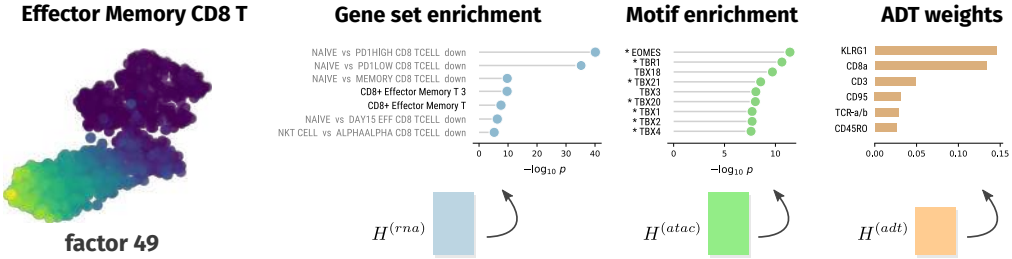
Several conditions and sequencing platforms



Cell-type specificity of the factors



Loadings allow to interpret specific factors



Conclusion

Interpretable OT-based joint dimensionality reduction for single-cell multiomics.

cantinilab/Mowgli Public

Single-cell multi-omics integration using Optimal Transport

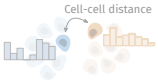
Python 36 2

```
pip install mowgli
```

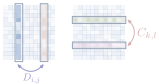
Open-source Python package

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

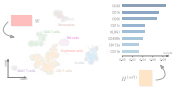
Preprint: G.-J. Huizing, G. Peyré, L. Cantini, *bioRxiv*, 2024



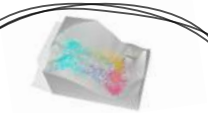
I. *OT as a cell-cell distance*



II. *Wasserstein Singular Vectors*



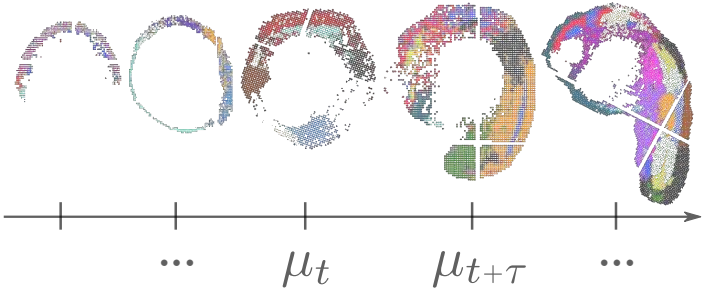
III. *OT-based joint dimensionality reduction*



IV. *OT-based trajectory inference in space*

Lagrangian setting

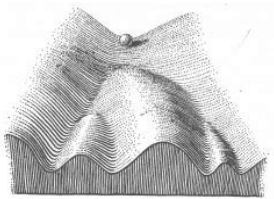
Spatial transcriptomics through time



Recent datasets have profiled spatial transcriptomics at several timepoints[33,34,35].

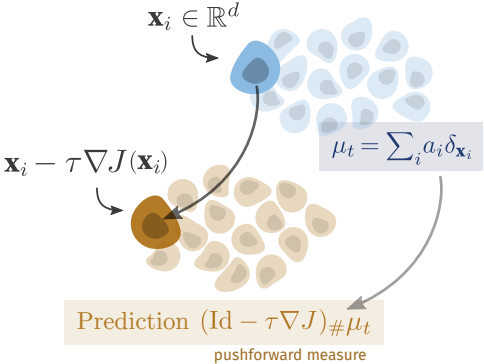
[33] Liu et al., *Developmental Cell*, 2022; [34] Chen et al., *Cell*, 2022; [35] Wei et al., *Science*, 2022

Potential landscapes



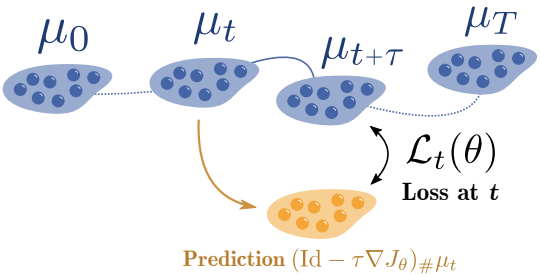
Potential $J : \mathbb{R}^d \to \mathbb{R}$

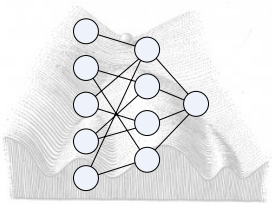
Recent trajectory inference methods[36,37,38] model cellular differentiation as the minimization of a **potential energy**.



[36] Hashimoto et al., *ICML*, 2016; [37] Yeo et al., *Nature communications*, 2021; [38] Bunne et al., *AISTATS*, 2022

Learning cell fate landscapes using time course data



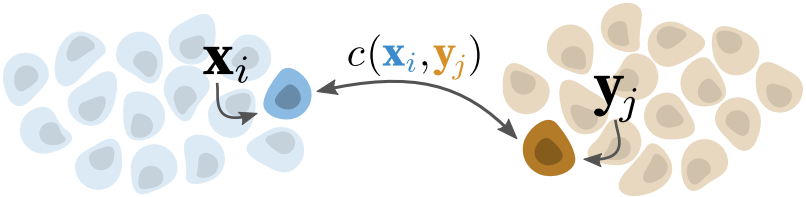


Potential J_θ = MLP

They learn a potential J_θ by comparing the model's predictions to the true distribution using Optimal Transport[39,30,41].

[39] Hashimoto et al., *ICML*, 2016; [40] Yeo et al., *Nature communications*, 2021; [41] Bunne et al., *AISTATS*, 2022

Sinkhorn loss

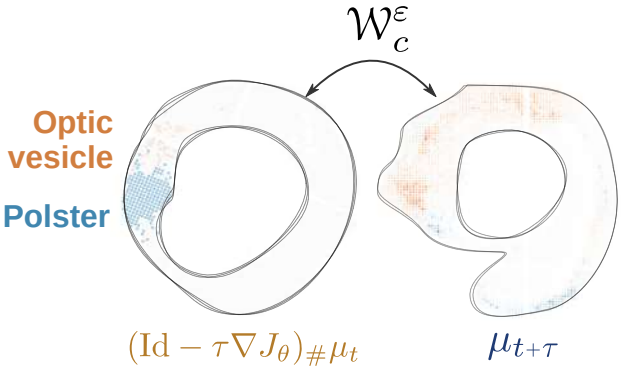


Ground truth $\mu_{t+\tau} = \sum_i a_i \delta_{\mathbf{x}_i}$

Prediction $\nu_{t+\tau} = (\mathrm{Id} - \tau \nabla J)_{\#} \mu_t$

$$\mathcal{W}_c^{\varepsilon}(\mu_{t+\tau}, \nu_{t+\tau}) \stackrel{\text{def.}}{=} \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j} c(\mathbf{x}_i, \mathbf{y}_j) P_{i,j} - \varepsilon \mathrm{E}(\mathbf{P})$$

The Sinkhorn loss is not spatially coherent



But the Sinkhorn loss matches regions that are not coherent in space.

Fused Gromov-Wasserstein loss

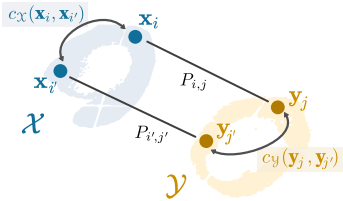
We propose using Fused Gromov-Wasserstein[42] to leverage space.

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

$$\boxed{\sum_{i,j} c(\mathbf{x}_i, \mathbf{y}_j) P_{i,j}} \qquad \boxed{\sum_{i,j,i',j'} \left| c_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_{i'}) - c_{\mathcal{Y}}(\mathbf{y}_j, \mathbf{y}_{j'}) \right|^2 P_{i,j} P_{i',j'}}$$

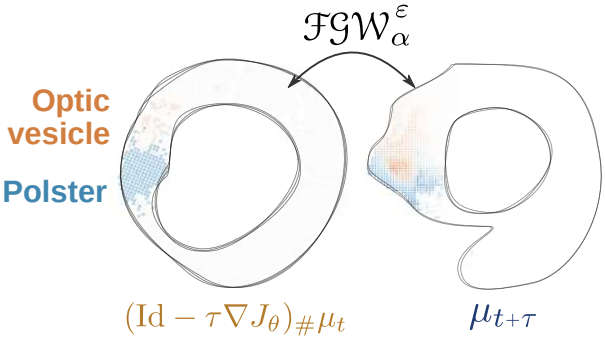
Linear term $\mathrm{L}(\mathbf{P})$
on gene expression

Quadratic term $\mathrm{Q}(\mathbf{P})$
on spatial coordinates



[42] Vayer et al., *arXiv*, 2018

Introducing spatial information with Fused Gromov-Wasserstein



Fused Gromov-Wasserstein leads to improved spatial coherence.

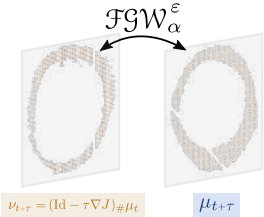
Benchmarking datasets: zebrafish, axolotl, and mouse

We consider three datasets[43,44,45]: zebrafish, mouse, and axolotl.

We split the data into a **train set**, **early test set**, and **late test set**.

[43] Liu et al., *Developmental Cell*, 2022; [44] Chen et al., *Cell*, 2022; [45] Wei et al., *Science*, 2022

Evaluation metrics



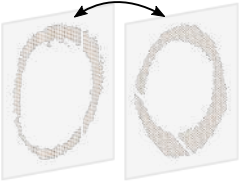
We evaluate predictions on the test set based on two metrics

$$\mathcal{FGW}_\alpha^\varepsilon(\mu_{t+\tau}, \nu_{t+\tau}) = \min_{\mathbf{P}\in\mathcal{U}(\mathbf{a},\mathbf{b})} (1-\alpha)\boxed{\mathrm{L}(\mathbf{P})} + \alpha\boxed{\mathrm{Q}(\mathbf{P})} - \varepsilon\mathrm{E}(\mathbf{P})$$

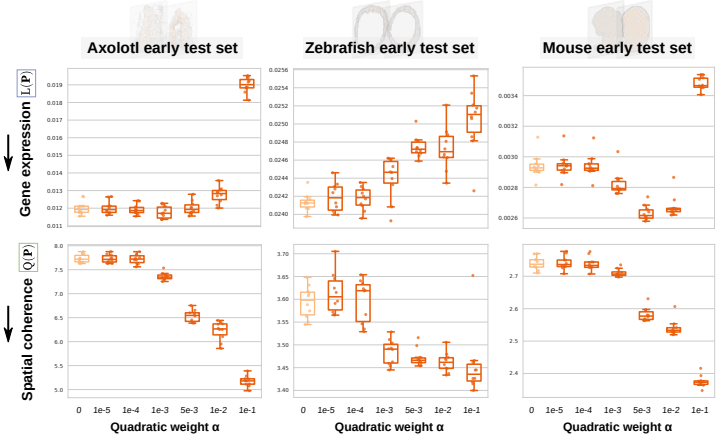
Gene expression **Spatial coherence**

Results on early time points

$$\mathcal{FGW}^{\varepsilon}_{\alpha} = (1-\alpha)\,\mathbf{L}(\mathbf{P}) + \alpha\,\mathbf{Q}(\mathbf{P}) - \varepsilon \mathrm{E}(\mathbf{P})$$

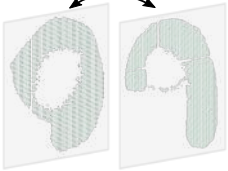


$(\mathrm{Id} - \tau\nabla J)_{\#}\mu_t$ $\mu_{t+\tau}$

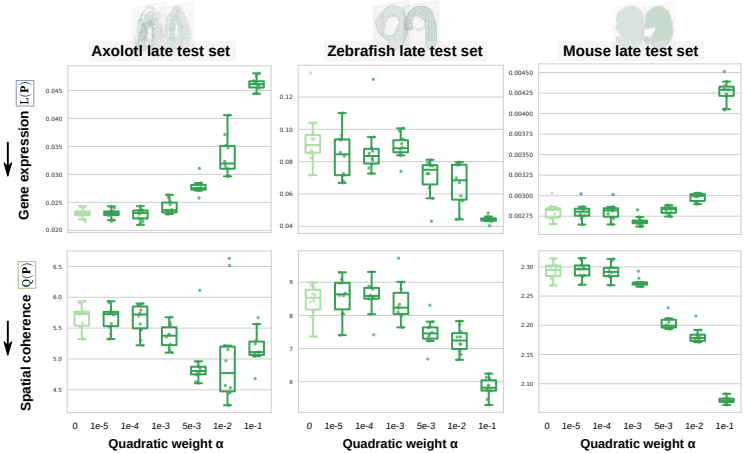


Results on late time points

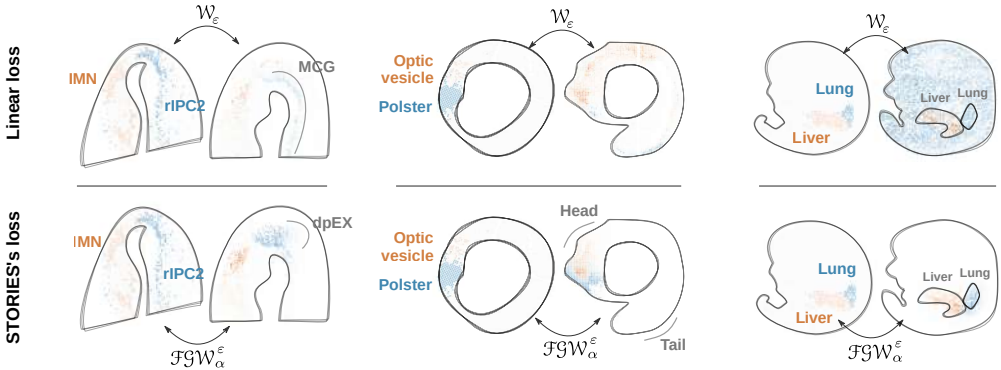
$$\mathcal{FGW}_\alpha^\varepsilon = (1-\alpha)\,\mathbf{L}(\mathbf{P}) + \alpha\,\mathbf{Q}(\mathbf{P}) - \varepsilon\mathbf{E}(\mathbf{P})$$



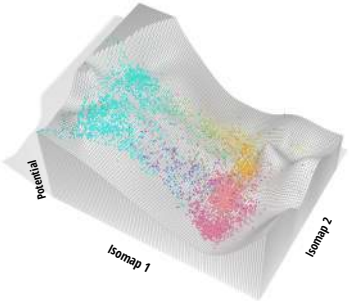
$(\mathrm{Id} - \tau\nabla J)_{\#}\mu_t$ $\mu_{t+\tau}$



Improved spatial coherence in the matchings



Learning a potential of axolotl neuron regeneration

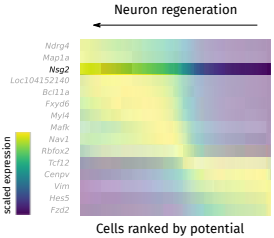


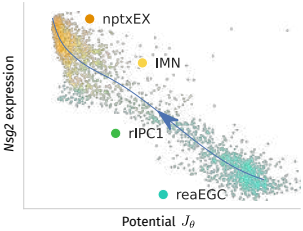
We computed a potential landscape of neuron regeneration in axolotls[46].

Gradients of this potential confirm a transition from EGCs to mature neurons.

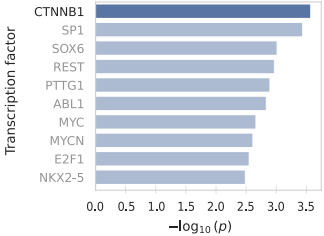
[46] Wei et al., *Science*, 2022

Analyzing the dynamics of neuron regeneration



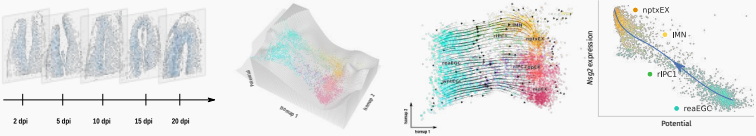


We identify potential driver genes by regressing their expression based on the potential.



Transcription Factors targeting these genes represent candidate regulators.

Conclusion



Trajectory inference for spatial transcriptomics profiled through time using FGW.

cantinilab/**stories** Public

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

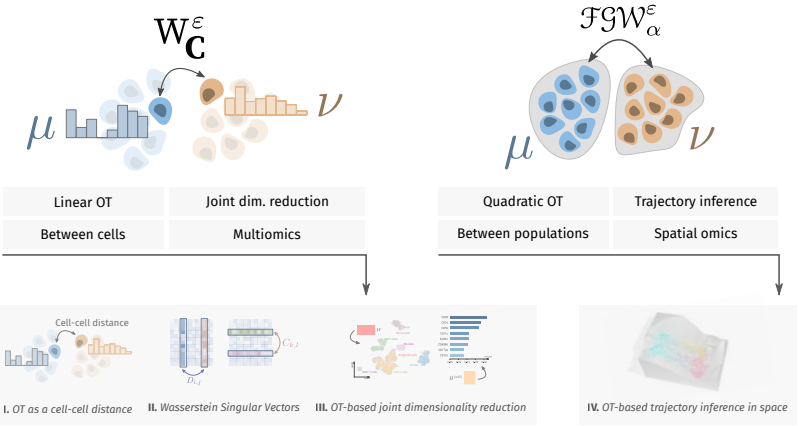
Python 8

```
pip install stories-jax
```

Open-source Python package

Conclusion

Contributions



[I] Huizing et al., *Bioinformatics*, 2022; [II] Huizing et al., *ICML*, 2022; [III] Huizing et al., *Nat. Comm.*, 2023; [IV] Huizing et al., *bioRxiv*, 2024

Perspectives: OT computation at the scale of single-cell atlases

The increased availability of large multimodal atlases motivates fast approximations of OT.

Graph diffusion-based[47]

Tree-based[48]

Neural network-based[49,50]

[47] Tong et al., *ICASSP*, 2022; [48] Düsterwald et al., *ICML GRaM Workshop*, 2024; [49] Courty et al., *ICLR*, 2018; [50] Haviv et al., *ICML*, 2024

Perspectives: Methodological improvements for further biological insights

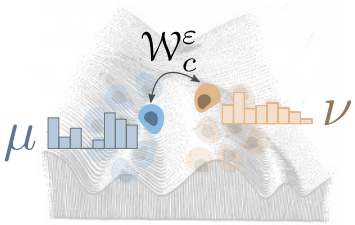
Finding rare cell types and complex cellular dynamics motivates further methodological improvements.

Nonlinear joint dimensionality reduction[51].

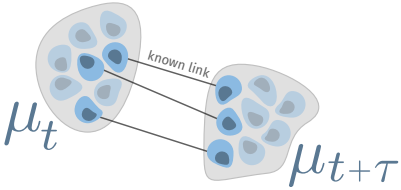
More general class of differentiation potentials[52,53].

[51] Schmitz et al., *SIAM Journal on Imaging Sciences*, 2018; [52] Weinreb et al., *PNAS*, 2018; [53] Terpin et al., *arXiv*, 2024

Perspectives: Unified OT framework for multimodal omics data



Hierarchical OT framework for population dynamics[54].



Ground metric learning from partial pairings in time-course data[55,56].

[54] Yurochkin et al., *NeurIPS*, 2019; [55] Dupuy and Galichon, *Journal of Political Economy*, 2014; [56] Chen et la., *Nature*, 2022

Acknowledgements

Jury members

Ahmed Mahfouz
Franck Picard
Andrea Rau
Laurence Calzone
Lorette Noiret
Laura Cantini
Gabriel Peyré

ML4IG lab • Institut Pasteur, CSD • ENS, and more!