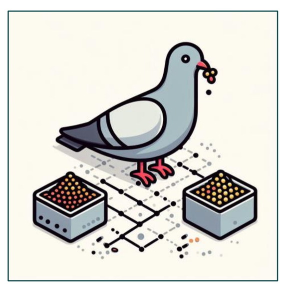
Global Ground Metric Learning with Applications to scRNA data

Damin Kühn, *Michael T. Schaub*
RWTH Aachen University



`pip install ggml-ot`

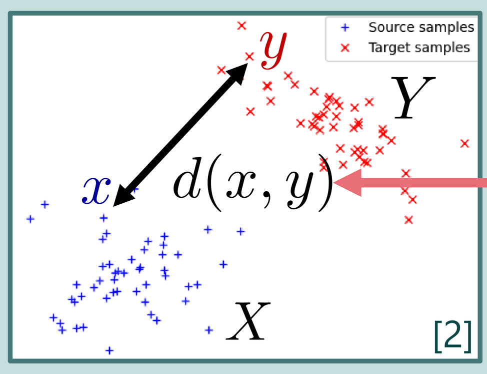
Metric Learning on Distributions & Elements with supervised Optimal Transport

Motivation

Optimal Transport (OT)

Distance Measure between Distributions based on the cost of an optimal mapping **(Transport Plan)**

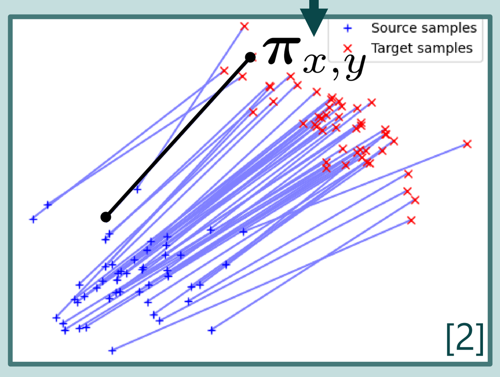
Wasserstein (EMD)



$$W(X,Y) = \min_{\pi} \sum_{x,y} d(x,y)\,\pi_{x,y}$$

Ground Metric (or cost)

- **critically influences Optimal Transport**
- usually pre-defined (e.g. Euclidean, cosine)



[1]

Methods

Global Ground Metric Learning (GGML)

Learn Global Metric as Ground Metric based on Distribution classes

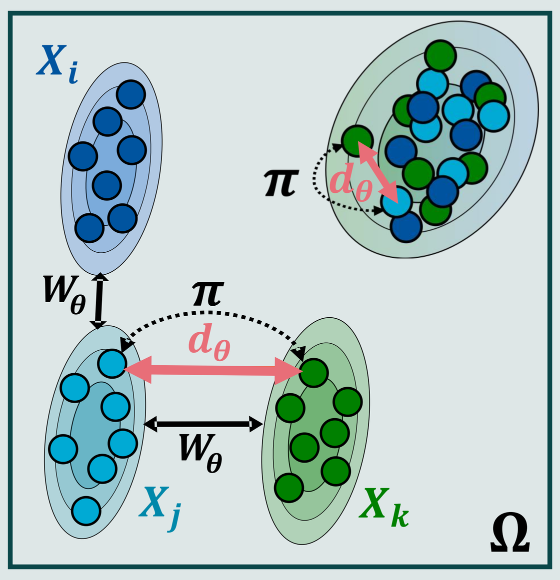
Distributions with classes

$\{X_i \ldots\}\ \{X_k \ldots\}$ class 1, class 2

Relative Relationships (Triplets)

$\tilde{\mathcal{T}} = \{(\ldots), \ldots\}$

$$W_\theta(X,Y) = \min_{\pi} \sum_{x,y} d_\theta(x,y)\,\pi_{x,y}$$



Loss Function (Biconvex Optimization)

$$\mathcal{L}_{\alpha,\lambda}(\theta, X, \tilde{\mathcal{T}}) = \sum_{t \in \tilde{\mathcal{T}}} \mathcal{L}_\alpha(\theta, X, t) + \lambda R(\theta)$$

where $\mathcal{L}_\alpha(\theta, X, (i,j,k)) =$

$$\max\left(W_\theta(X_i, X_j) - W_\theta(X_j, X_k) + \alpha, 0\right)$$

learn to separate Relative Relationships by margin α

Global Metric Learning

Learn metric between labeled elements

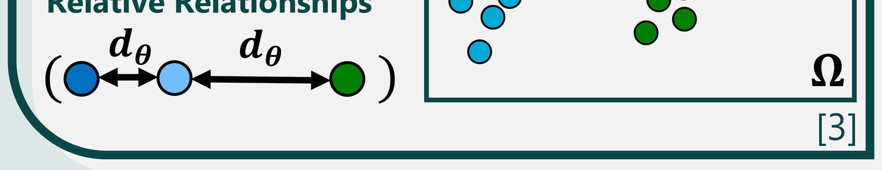
Elements with classes

$\{x_i \ldots\}\ \{\ldots\}$ class 1, class 2

Parameterized Metric

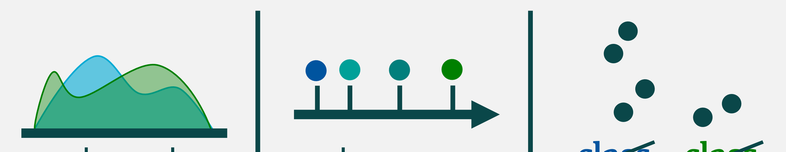
$d_\theta(x,y)\colon \Omega \times \Omega \to \mathbb{R}_{\geq 0}$

Relative Relationships



Ground Metric Learning

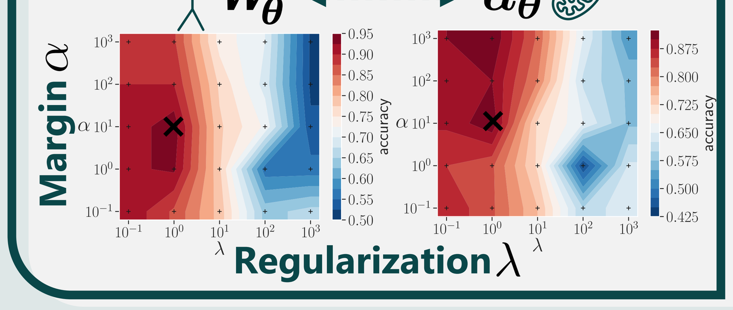
Limitations of existing approaches



shared supports [4]	known timesteps [5]	unsupervised [6]

Hyperparameters

W_θ ⟷ d_θ



[7] Mahalanobis Distance in GGML

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)} = \|W x_i - W x_j\|$$

$$\widetilde{W}^T \widetilde{W} \approx W^T W = M$$

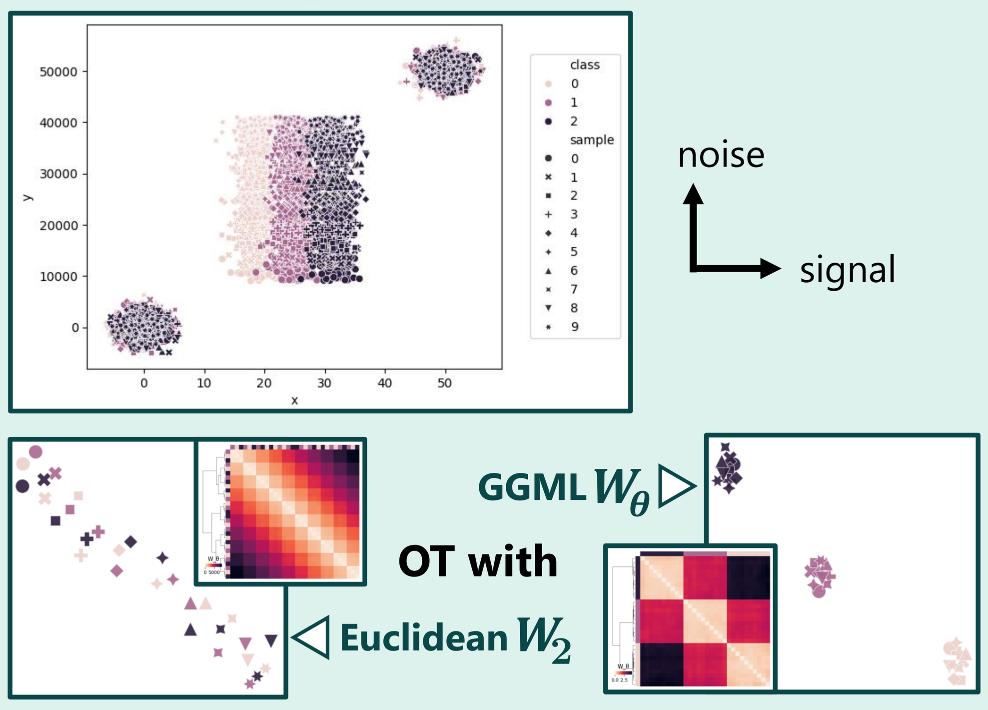
low-dim. subspace
underlying class relations

learn projection into linear subspace as θ

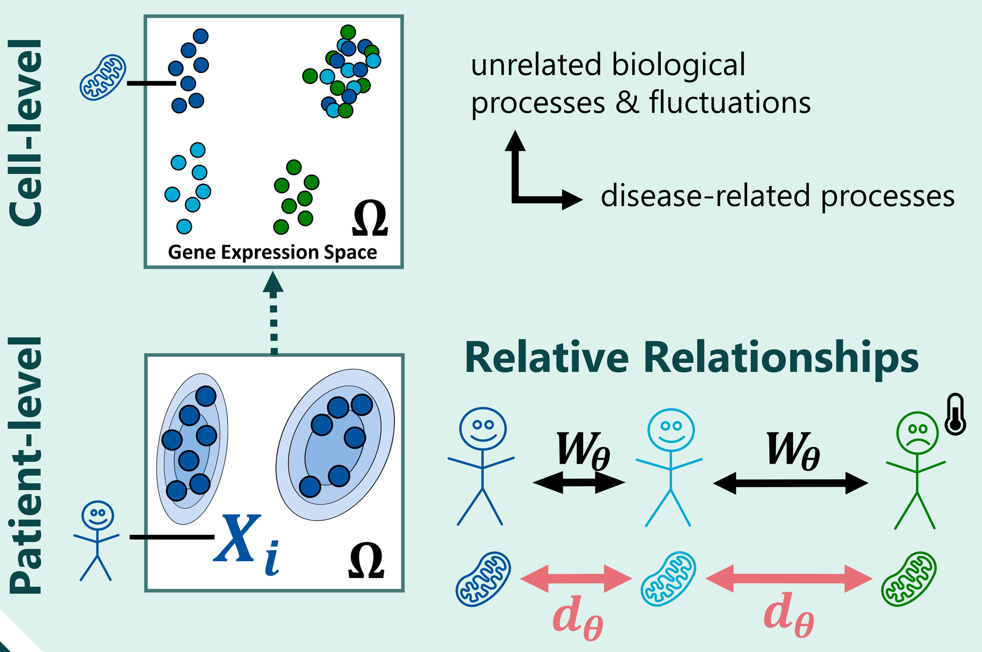
Applications

2D Example

Distributions from 3 classes

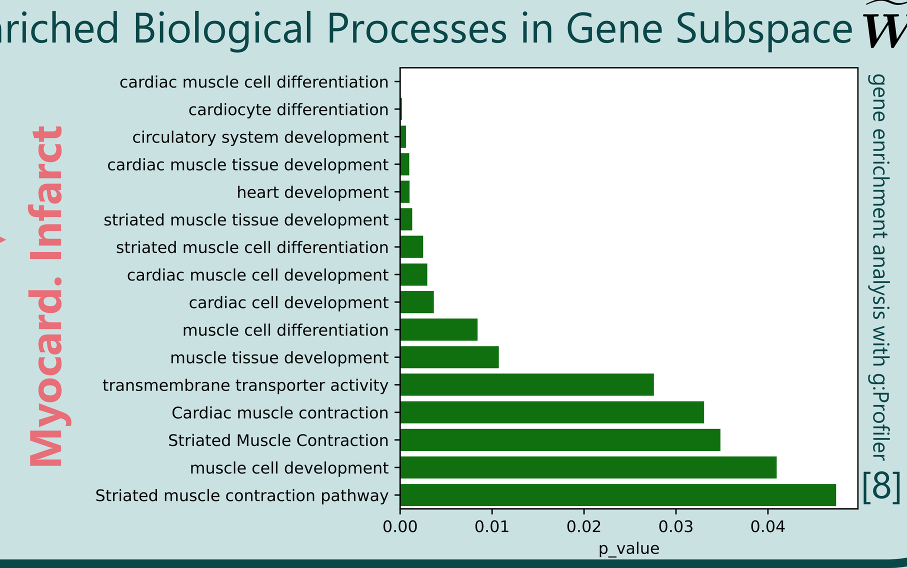


Single-cell RNA-seq

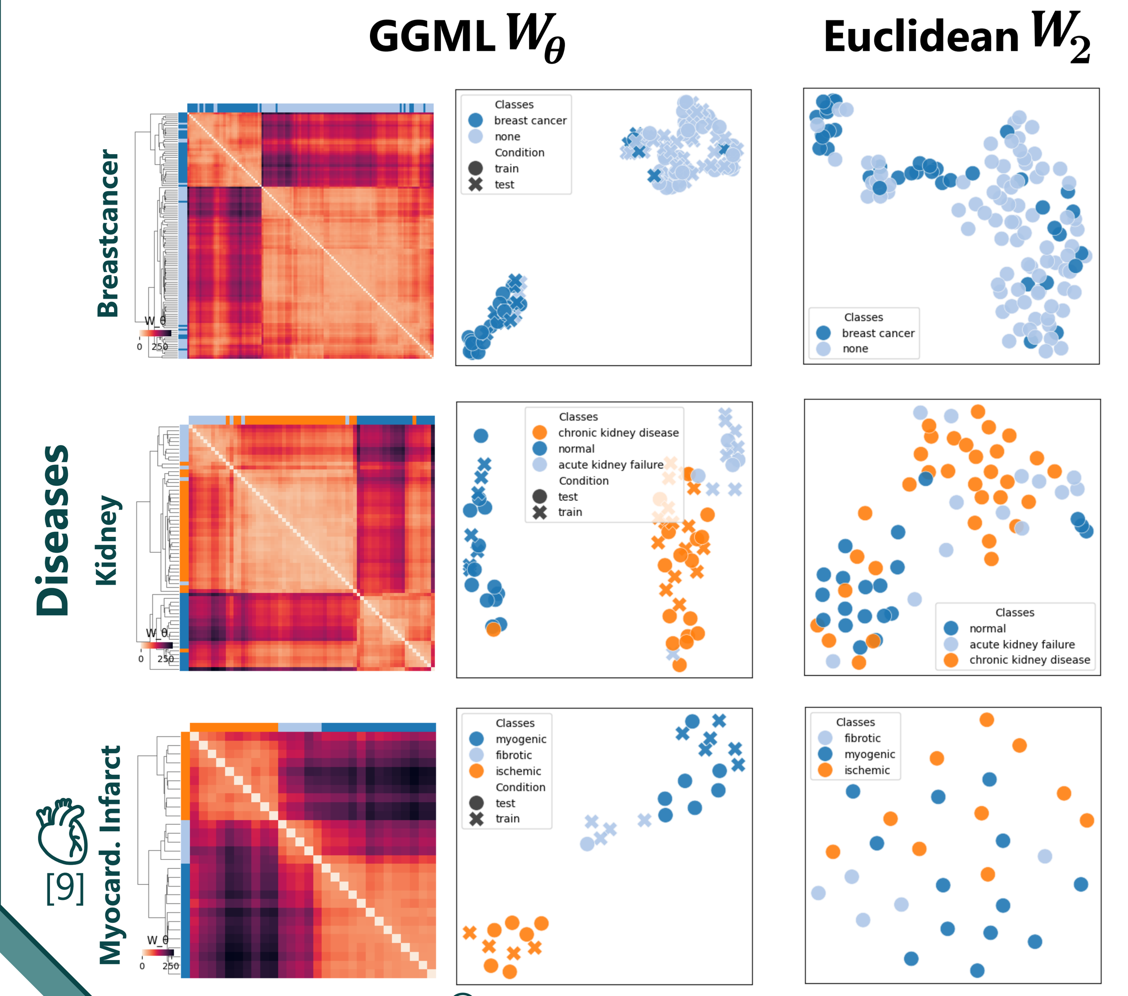


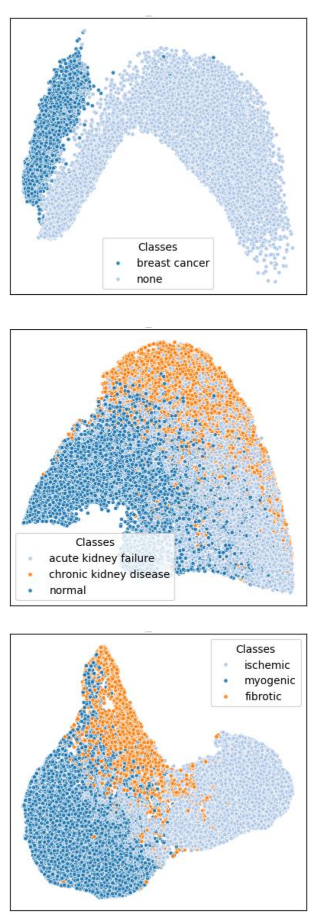
Biological Interpretability

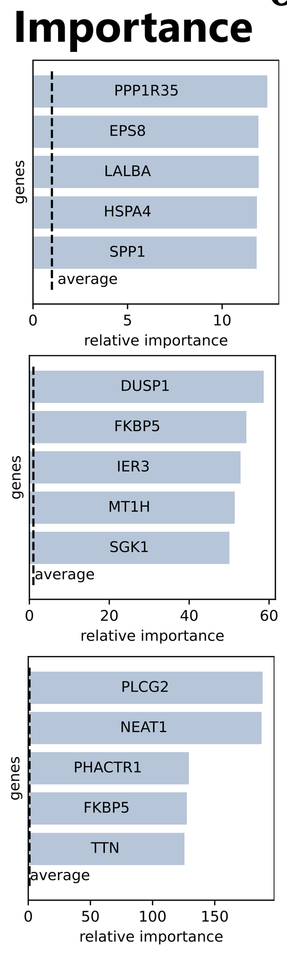
Enriched Biological Processes in Gene Subspace $\widetilde{W}$



Results



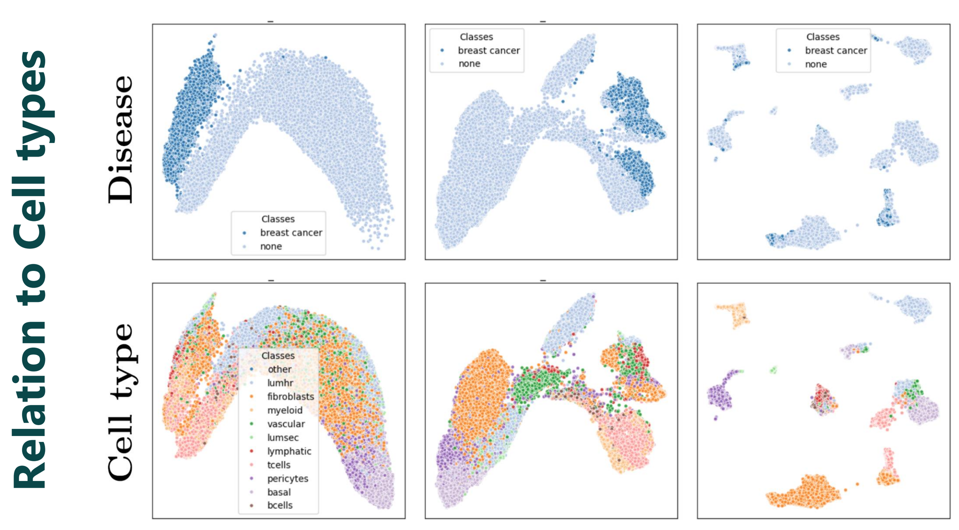




KNN Classification

Method	Synth$_{2D}$	Synth$_{200D}$	Kidney	Brst.Canc.	Myocard.	Synth$_{2D}$	Synth$_{200D}$	Kidney	Brst.Canc.	Myocard.
	patient-level					*cell-level*				
Eucl.	0.24±0.08	0.39±0.12	0.52±0.10	0.77±0.03	0.49±0.03	0.32±0.01	0.45±0.01	0.45±0.11	0.79±0.04	0.48±0.10
Manh.	0.24±0.08	0.39±0.11	0.57±0.08	0.79±0.03	0.85±0.03	0.33±0.01	0.41±0.01	0.48±0.07	0.79±0.04	0.56±0.08
Cos.	0.43±0.07	0.46±0.10	0.54±0.07	0.79±0.03	0.53±0.06	0.36±0.01	0.35±0.01	0.46±0.10	0.79±0.04	0.53±0.12
LMNN	0.22±0.07	0.29±0.11	*OOM*	*OOM*	*OOM*	0.38±0.01	0.45±0.01	*OOM*	*OOM*	*OOM*
LFDA	0.46±0.06	0.47±0.09	0.86±0.11	0.82±0.07	0.88±0.11	0.40±0.01	0.37±0.01	0.52±0.06	0.67±0.03	0.88±0.07
NCA	0.35±0.11	0.25±0.11	*OOT*	*OOT*	0.81±0.06	0.37±0.00	0.46±0.02	*OOT*	*OOT*	0.78±0.08
ITML	0.51±0.09	0.43±0.08	0.55±0.10	0.76±0.04	0.79±0.04	0.41±0.01	0.36±0.01	0.37±0.06	0.77±0.04	0.54±0.07
GGML	**0.96±0.04**	**0.95±0.11**	**0.94±0.02**	**0.91±0.04**	**0.92±0.08**	**0.53±0.01**	**0.53±0.01**	**0.74±0.00**	**0.81±0.03**	**0.94±0.00**

Patient-level — Cell-level



Euclidean d_2 is used in computational genomics to define celltypes.

Computational Network Science
RWTH AACHEN UNIVERSITY
SPONSORED BY THE
Federal Ministry of Education and Research

References

[1] Peyré, Gabriel, and Marco Cuturi. "Computational optimal transport: With applications to data science." (2019)
[2] Flamary, Rémi, et al. "Pot: Python optimal transport." (2021)
[3] Kulis, Brian. "Metric learning: A survey." (2013)
[4] Cuturi, Marco, and David Avis. "Ground metric learning." (2014)
[5] Scarvelis, Christopher, and Justin Solomon. "Riemannian metric learning via optimal transport." (2023)
[6] Huizing, Geert-Jan, et al. "Unsupervised ground metric learning using wasserstein singular vectors." (2022)
[7] Davis, Jason V., et al. "Information-theoretic metric learning." (2007)
[8] Raudvere, Uku, et al. "g: Profiler: a web server for functional enrichment analysis and conversions of gene lists." (2019)
[9] Kuppe et al., "Spatial multi-omic map of human myocardial infarction." (2022)

Damin Kühn
RWTH Aachen
kuehn@cs.rwth-aachen.de
daminkuehn.de