Global Ground Metric Learning with Applications to scRNA data

Damin Kühn **Michael T. Schaub**
Department of Computer Science
RWTH Aachen University, Germany

Abstract

Optimal transport (OT) provides a robust framework for comparing probability distributions. Its effectiveness is significantly influenced by the choice of the underlying ground metric. Traditionally, the ground metric has either been (i) predefined, e.g., as a Euclidean metric, or (ii) learned in a supervised way, by utilizing labeled data to learn a suitable ground metric for enhanced task-specific performance. Yet, predefined metrics typically cannot account for the inherent structure and varying significance of different features in the data, and existing supervised ground metric learning methods often fail to generalize across multiple classes or are limited to distributions with shared supports. To address this issue, this paper introduces a novel approach for learning metrics for arbitrary distributions over a shared metric space. Our method provides a distance between individual points (samples) like a global metric, but requires only class labels on a distribution-level for training. The resulting learned global ground metric enables more accurate OT distances, which can significantly improve clustering and classification tasks. Further, we can create task-specific shared embeddings for elements (samples) from different distributions, including unseen data.

1 INTRODUCTION

Optimal transport (OT) has reemerged as a powerful mathematical framework for comparing probability distributions. Recently, it has found applications in analyzing data across various domains like computer vision [Bonneel and Digne, 2023], natural language processing [Sato et al., 2022] and computational genomics [Schiebinger et al., 2019, Joodaki et al., 2024].

Our work is motivated by patient-level single cell RNA (scRNA) analysis, where the available data consists of the expression patterns of individual cells from different patients. Specifically, we consider a setting in which we have multiple patients, each represented by a set of cells, where each cell is characterized by a high-dimensional gene expression vector. Hence, these gene expression vectors of cells from the same patient form a distribution over the gene space, i.e., each patient is represented by a distribution over the gene space. Our objective is to compare these distributions to identify similarities and differences between patients. Intuitively, we can use OT to probabilistically map two such (empirical) distributions (cells from two patients) onto each other while minimizing the overall cost of this transport. The associated cost of this optimal transport is then a Wasserstein distance W, which quantifies the distance (dissimilarity) between the distributions of the two patients. However, for this computation we need to choose an underlying Ground Metric d, a distance between the gene expression vectors of the cells. Clearly, the chosen ground metric will critically influence the resulting Wasserstein distance W.

Yet, most current applications of OT manually choose a predefined and fixed ground metric such as the Euclidean metric. Such simple ground metrics are however rarely tailored to the task at hand, and can thus lead to suboptimal results, as they do not account for the inherent structure and varying significance of different features in the data. Ground Metric Learning (Ground ML) aims at directly learning the distance between sampled elements based on prior information about the Wasserstein distance between their distributions. A promising application of Ground ML in scRNA data is to leverage different disease states of patients as prior information for their Wasserstein distances. Under this

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

setup, Ground ML learns relevant distances between the expression vectors of single cells that optimally separate disease states on a patient-level. However, early work in Ground ML is constrained to pairs of distributions or requires the distributions to have shared support, which limits its applicability [Wang and Guibas, 2012, Cuturi and Avis, 2014, Huang et al., 2016, Huizing et al., 2022].

In this paper, we introduce a general framework to ground metric learning for classes of arbitrary distributions over a shared space, called Global Ground Metric Learning (GGML), that circumvents such problems of prior works. Theoretically, GGML can learn arbitrary differentiable metrics d_θ between sampled elements based solely on the class labels of their distributions. We demonstrate this with a low rank approximation of a popular learnable global metric, the Mahalanobis distance. Notably, our framework enables data analysis on both, the distribution- and the element-level corresponding to patients respective cells. Used as a ground metric in OT, the learned metric can improve performance and interpretability of the resulting OT distances in down-stream applications. Used as a global metric between elements of the distributions, the learned metric offers similar benefits for analysis on the element-level. This includes applications such as embedding, clustering, classification and feature importance. We validate and benchmark our approach using both synthetic and real-world single cell genomics datasets spanning various diseases.

2 BACKGROUND

Notation. Vectors $\boldsymbol{x}$ and matrices $\boldsymbol{M}$ are denoted in bold, sets $\mathcal{T}$ calligraphically. We use $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to denote a metric $d : \Omega^2 \mapsto \mathbb{R}_{\geq 0}$ between two sampled elements from the space Ω (gene expression space).

Mahalanobis distance. The Mahalanobis distance is a parameterized, linear metric [Davis et al., 2007]. It is defined as:

$$d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)} \quad (1)$$

$$= \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{x}_j\| \quad (2)$$

$$s.t.\ \boldsymbol{W}^T\boldsymbol{W} = (\boldsymbol{Q}\Lambda^{\frac{1}{2}})(\boldsymbol{Q}\Lambda^{\frac{1}{2}})^T = \boldsymbol{Q}\Lambda\boldsymbol{Q}^T = \boldsymbol{M}$$

where $\boldsymbol{M}$ is a symmetric, positive semi-definite (psd) matrix that can be learned. As every real symmetric psd matrix $\boldsymbol{M}$ can be represented as a product $\boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W}$, e.g., using the spectral decomposition, the Mahalanobis distance can be equivalently expressed as performing a linear mapping ($\boldsymbol{W}$) into a transformed space in which the Euclidean distance better approximates the desired distances.

We remark that the Mahalanobis distance is convex and Lipschitz continuous [Zantedeschi et al., 2016]. Further, the gradient of d_M with respect to row $\boldsymbol{W}_r$ of $\boldsymbol{W}$ is given as:

$$\frac{\partial d_M}{\partial \boldsymbol{W}_r} = \boldsymbol{W}_r^T(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j) \quad (3)$$

which allows to efficiently learn an optimal $\boldsymbol{W}$ (resp. $\boldsymbol{M}$) using gradient descent over some differentiable cost function $L(\boldsymbol{W})$ [Hocke and Martinetz, 2014].

Global Metric Learning. Global Metric Learning is a framework to learn a parameterized distance metric $d_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for some ground truth distances $d^*(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The Mahalanobis distance is a popular choice for such a metric d_θ, where the parameters are $\theta = \boldsymbol{M}$ or equivalently $\boldsymbol{W}$. As ground truth distances are rarely available, most Global ML methods rely on the availability of a set of pairwise similar points $\mathcal{P}_{\not\approx}$ and pairwise dissimilar points $\mathcal{P}_{\approx}$ to learn the metric. Global ML then typically amounts to solving an optimization problem of the form (see also section 5):

$$\min_\theta \sum_{(i,j)\in\mathcal{P}_{\approx}} d_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) \quad (4)$$

$$s.t.\ \forall(\boldsymbol{x}_j, \boldsymbol{x}_k) \in \mathcal{P}_{\not\approx} :\ d_\theta(\boldsymbol{x}_j, \boldsymbol{x}_k) \geq 1$$

where the learned metric d_θ minimizes the distance between pairs of similar points $(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{P}_{\approx}$ while satisfying a distance margin between dissimilar points $(\boldsymbol{x}_j, \boldsymbol{x}_k) \in \mathcal{P}_{\not\approx}$.

Triplet Learning. Triplet learning (also known in terms of triplet loss in deep learning) was orig-

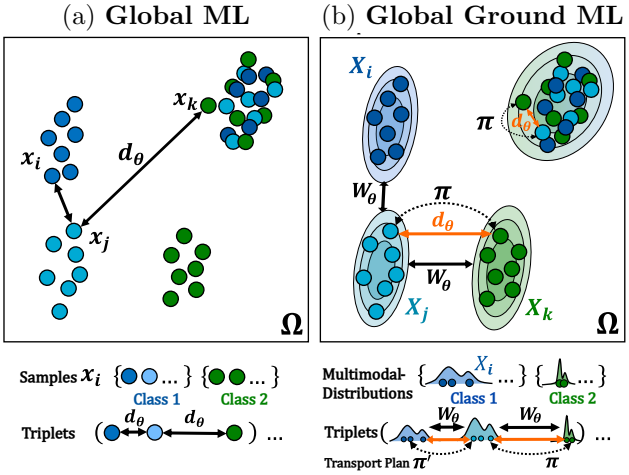


Figure 1: (a) Global ML aims to learn from relative relationships between elements of different classes, without considering multimodal-distributions in a class. (b) Global Ground ML learns from triplet relationships of distributions of different classes, enabling to learn a metric that: globally captures relationships in different modes and, used as a ground metric, improves Wasserstein distances between distributions.

inally introduced in the context of (global) metric learning under the term Relative Comparisons [Schultz and Joachims, 2003]. The idea is to use triplets $\mathcal{T} = \{(i,j,k) |\, d^*(\boldsymbol{x}_j, \boldsymbol{x}_k) - d^*(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 1\}$ that capture the relative relationships between distributions. More specifically, $\mathcal{T}$ contains exactly all triplets (i,j,k) where the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is smaller than between $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ by at least 1. The corresponding metric learning (optimization) problem can then be formulated as:

$$\min_{\theta} \frac{1}{2} \| \theta \|^2 \tag{5}$$
$$s.t. \, \forall (i,j,k) \in \mathcal{T} : \; d_\theta(\boldsymbol{x}_j, \boldsymbol{x}_k) - d_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 1$$

which aims at finding a regularized θ that satisfies the constraints on the relative relationships.

Wasserstein distance (EMD). Given two (empirical) distributions X, Y, the Wasserstein distance, also called Earth Movers Distance (EMD), is given as:

$$W(X,Y) = \min_{\pi} \sum_{x,y} d(x,y) \boldsymbol{\pi}_{x,y} \tag{6}$$

where $\boldsymbol{\pi}$ is an admissible coupling (or transport plan) which maps elements $x \sim X$ to $y \sim Y$ [Peyré et al., 2019]. Here d is typically an underlying (a priori) defined ground metric such as the Euclidean, Manhattan and Cosine distance.

Ground Metric Learning. We consider the notion of Ground Metric Learning as learning a distance function $d_\theta(x,y) : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$. For a discussion of different approaches please refer to section 5.

Let $d_\theta(x,y)$ be a parameterized metric between two data points x, y that is partially differentiable with respect to θ. Using d_θ as the ground metric yields a parameterized Wasserstein distance:

$$W_\theta(X,Y) = \min_{\pi} \sum_{x,y} d_\theta(x,y) \boldsymbol{\pi}_{x,y} \tag{7}$$

The ground metric learning problem now is to learn d_θ such that $W_\theta(X,Y)$ approximates a given ground truth distance $W^*(X,Y)$. Specifically we want to find the parameters for the ground metric d_θ that minimizes the prediction error $|W_\theta(X,Y) - W^*(X,Y)|$. We observe that even for two distributions and ground truth distances available, learning the ground metric now becomes a nested optimization problem. The problem is biconvex with respect to the optimal transport plan π between X, Y and the parameters θ of the underlying ground metric d_θ.

3 GLOBAL GROUND METRIC LEARNING

Let $X_1, ..., X_n$ be probability distributions over the space Ω, each labeled with a corresponding class $c_1, ..., c_n \in \{1, ..., k\}$. The general aim is to learn a Wasserstein distance between these distributions that separates the classes. As described in Equation 7, the underlying ground metric d_θ between elements $\boldsymbol{x}_i \sim X_i, \boldsymbol{x}_j \sim X_j$ can be adjusted. Hence, we learn an underlying ground metric that that assigns distances between individual data points like a global metric, but only requires distance information on the distribution level for training like a ground metric.

To learn d_θ with Ground ML using a triplet loss, we would like to construct a set of triplets $\mathcal{T} = \{(i,j,k) |\, W^*(X_j, X_k) - W^*(X_i, X_j) \geq \alpha\}$ containing relative distance relationships. However, as typically no distance information is available but only class labels, we approximate $\mathcal{T}$ by $\tilde{\mathcal{T}} = \{(i,j,k) | c_i = c_j \wedge c_j \neq c_k\}$. This set $\tilde{\mathcal{T}}$ contains exactly the triplets (i,j,k) where the empirical distributions X_i, X_j belong to the same class while $X_j, X_k k$ do not. Intuitively, we learn a ground metric such that distributions from the same class are closer while distributions from different classes are further apart.

Assumption 1. *Distributions X_i, X_j from the same class $c_i = c_j$ are closer than distributions X_j, X_k from different classes $c_j \neq c_k$. Specifically, there exists a margin $\alpha \in \mathbb{R}_{>0}$ between the distances $W^*(X_i, X_j)$ of distributions from the same class and the distance $W^*(X_j, X_k)$ of distributions from different classes.*

Under Assumption, it holds that 1, $\tilde{\mathcal{T}} \subseteq \mathcal{T}$. However, as the size of $\tilde{\mathcal{T}}$ still scales with $\mathcal{O}(n^3)$, we introduce a neighbor parameter t that controls how many neighbors of distribution j are considered to form triplets (i,j,k). See section 5, for related approaches. For each j, we take the Cartesian product of t neighbors from the same class, indexed by i, and t neighbors from the $C-1$ other classes, indexed by k. This significantly improves the scalability as the number of classes C typically does not grow with the number of data points n. Under this assumption, the set of triplets $\tilde{\mathcal{T}}_t$ on which we optimize the ground metric has size $|\tilde{\mathcal{T}}_t| = nt^2(C-1)$, i.e., scales linearly in $\mathcal{O}(n)$.

We now aim at learning a metric d_θ such that for all triplets $(i,j,k) \in \tilde{\mathcal{T}}_t$ it holds that $W_\theta(X_j, X_k) - W_\theta(X_i, X_j) \geq \alpha$, or equivalently $W_\theta(X_i, X_j) - W_\theta(X_j, X_k) + \alpha \leq 0$. We write $W_\theta \approx_\alpha W^*$ if this condition is fullfilled. To allow for errors, we further relax this hard constraint and instead formulate a minimization problem over all triplets. We thus define the (unregularized) loss function to learn global ground

metrics from distributions with class labels as:

$$\mathcal{L}_\alpha(\theta, X, \tilde{\mathcal{T}}_t) = \sum_{t \in \tilde{\mathcal{T}}_t} \mathcal{L}_\alpha(\theta, X, t) \quad (8)$$

$$\text{where } \mathcal{L}_\alpha(\theta, X, (i,j,k)) = \max\left(W_\theta(X_i, X_j) - W_\theta(X_j, X_k) + \alpha, 0\right) \quad (9)$$

To balance the influence of the unbounded $-W_\theta(X_j, X_k)$ term on the objective function, we use a ReLU activation function to bound the margin of each triplet relationship in Equation 9. Intuitively, the loss corresponds to the sum of errors of all triplet relationships separated by a margin less than α.

Theorem 1. *The GGML loss is 0 if, and only if, the global ground metric d_θ in W_θ approximates the ground truth distances W^* with margin α.*

$$\mathcal{L}_\alpha(\theta, X, \tilde{\mathcal{T}}_t) = 0 \;\; \textit{iff} \;\; W_\theta \approx_\alpha W^*$$

This theorem states the desirable property that the minimal loss is only 0 if the ground metric d_θ in W_θ separates relative relationships between classes by at least margin α.

If there exist $\theta' \neq \theta$ such that $d_\theta = d_{\theta'}$, the optimization of the loss (8) can become ill-posed. This is certainly the case for the Mahalanobis distance (Equation 2) as uniqueness of $\boldsymbol{W}$ not guaranteed in the matrix decomposition. To facilitate unique optimal solutions, we thus introduce the regularized loss function:

$$\mathcal{L}_{\alpha,\lambda}(\theta, X, \tilde{\mathcal{T}}_t) = \sum_{t \in \tilde{\mathcal{T}}_t} \mathcal{L}_\alpha(\theta, X, t) + \lambda R(\theta) \quad (10)$$

where $R(\theta)$ acts as a (differentiable) regularization term on θ. Parameter $\lambda \in \mathbb{R}_{\geq 0}$ controls the regularization strength. In our numerical evaluation on scRNA data, we regularize with the Frobenius norm to avoid overfitting and improve generalizability (Ridge Regression). To increase the sparsity of learned θ a L1 norm can be used instead (Lasso Regression) which we demonstrate on synthetic data. For different purposes other regularization schemes may be used as well.

Learning θ by optimizing $\mathcal{L}(\theta, X, \tilde{\mathcal{T}}_t)$ corresponds to a finite sum problem over $\mathcal{T}_t$. By construction d_θ, and hence W_θ, is partially differentiable with respect to θ which enables optimization using gradient descent. Despite our approach facilitating a linear complexity of the triplet set $\mathcal{T}_t$, this optimization can become difficult to solve for large datasets. To address this, we perform Stochastic Gradient Descent (SGD) over subsets of $\mathcal{T}_t$ as mini batches. Unlike previous SGD approaches in Ground ML [Wang and Guibas, 2012], we do not alternate between optimizing the transport plan and the ground metric. Instead, we directly compute the full gradient over the nested optimization problem which is guaranteed to exist.

3.1 Low Rank Mahalanobis Distance as Global Ground Metric

We illustrate the applicability of our approach using a Mahalanobis distance (Equation 2). More specifically, we aim to learn a Mahalanobis distance as our ground metric d_θ with $\theta = \boldsymbol{W}$. Interestingly when using the Mahalanobis distance as a Global Ground Metric in our framework, we can give upper bounds on the partial triplet loss and the total loss

Theorem 2 [Triplet Loss Bound]**.** *For all distributions X and triplets $(i,j,k) \in \tilde{\mathcal{T}}_t$, there exist a θ such that $\mathcal{L}(\theta, X, (i,j,k))$ is at most α.*

For all distributions X and sets of triplets $\tilde{\mathcal{T}}_t$, the minimal loss $\mathcal{L}(\theta, X, \tilde{\mathcal{T}}_t)$ is at most $\alpha|\tilde{\mathcal{T}}_t|$.

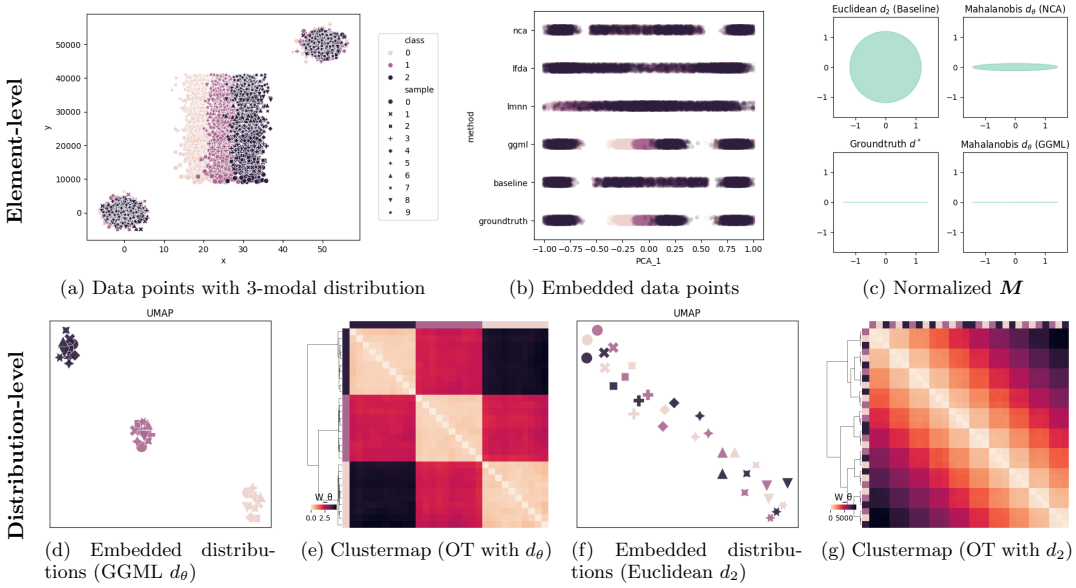
The number of parameters of a Mahalanobis distance scales quadratically with feature dimensions which prohibits its application to large datasets. To facilitate efficient computations on large datasets, we propose an approximation of the Mahalanobis matrix $\boldsymbol{M}$ by some low rank matrix factorization such that $\widetilde{\boldsymbol{W}}^T\widetilde{\boldsymbol{W}} \approx \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{M}$, where $\widetilde{\boldsymbol{W}}$ is of rank $k \ll n$. For low rank, symmetric $\boldsymbol{M}$ of rank k this approximation is of course exact. In practice, we observe that the approximation is accurate for low rank $\boldsymbol{M}$ and becomes less accurate for higher rank.

Intuitively, the existence of a low rank Mahalanobis (covariance) matrix $\boldsymbol{M}$ corresponds to the data being supported in some lower-dimensional subspace. Our low rank method enables us to efficiently learn such subspaces that capture class relations. Other approaches to address high dimensionality (e.g., PCA) reduce the feature space to some lower-dimensional subspace that does not necessarily capture such class relationships. More specifically, methods like PCA aim at capturing all aspects of the data which might include a lot of variance shared between classes (i.e. unrelated signals, noise). Using GGML to learn a low rank Mahalanobis distance thus enables us to differentiate class-related variations in the data from noise and class-unrelated signals.

4 APPLICATIONS TO SYNTHETIC AND REAL-WORLD SCRNA-SEQ DATA

To evaluate our novel framework, we perform several tasks on synthetic and real-world scRNA-seq datasets. In this section, we describe the setup and evaluation of different tasks using distances d resp. $\boldsymbol{W}$ learned by different competing methods, introduced in section 5.

Classification serves as a benchmark to evaluate how effectively the learned distances from various



(a) Data points with 3-modal distribution (b) Embedded data points (c) Normalized $\boldsymbol{M}$

(d) Embedded distributions (GGML d_θ) (e) Clustermap (OT with d_θ) (f) Embedded distributions (Euclidean d_2) (g) Clustermap (OT with d_2)

Figure 2: (a) Multimodal distributions with class labels as synthetic data. (b) Embeddings of elements show that only GGML can recover the ground truth where class-related variance is only given by one axis, shown by the (c) learned covariance matrix. (d) Clustermaps and (e) Embeddings between distributions show that GGML learns to differentiate classes under Optimal Transport, which a (f,g) Euclidean ground metric can not.

metric learning methods reflect the relationships between classes. We perform classification on both the single cell- and patient-level using a weighted k-NN [Guo et al., 2003]. This classifier assigns class labels to data points based on the weighted majority class of its k-closest neighbors under d resp. W. To evaluate generalizability, we measure the prediction accuracy over 10 test-train splits with respective half of the data points. We are withholding 20% of data points in each split for the hyperparameter tuning in subsection 4.4. Test-train splits are done under a group shuffle split such that cells from a specific patient only occur in a given train or test set.

Clustering of data points and distributions using agglomerative hierarchical clustering [Murtagh and Contreras, 2012]. This method builds hierarchical clusters by iteratively merging existing clusters based on the learned distances. It is evaluated using common clustering metrics, see appendix A.2

Embedding into a low dimensional space where the euclidean distance approximates the learned d resp. W. We show embeddings on the distribution-level corresponding to relationships between patients under a Wasserstein distance W using d as a ground metric. The corresponding single cell-level embeddings show the relationships between cells with d as a global metric. We use UMAP to compute 2D embeddings of the learned distances for visualization purposes [McInnes et al., 2018].

Feature Importance by interpreting rows of $\widetilde{\boldsymbol{W}}$ as distinct class-related processes. To get gene importance values over all class-related processes, we reconstruct the full Mahalanobis matrix as $\boldsymbol{M} \approx \widetilde{\boldsymbol{W}}^T\widetilde{\boldsymbol{W}}$. Diagonal entries correspond to the features relative importance in differentiating the classes. Off-diagonal entries correspond to learned relative importance of interactions (i.e. correlations). For scRNA data, this feature importance can be interpreted as the relative importance of expressed genes in explaining disease states. Here, we qualitatively describe the identified genes in the literature for the respective disease. An exemplary gene enrichment analysis can be found in Figure 7.

4.1 Synthetic Datasets

To validate our approach in a controlled setting, we generate synthetic data with reliable ground truth. Elements $\boldsymbol{x_i}$ in this dataset are distributed according to three-modal distributions X_i from three classes. As shown in Figure 2, the three classes are only distinguishable in one of the modes (center) and along one axis. The other two modes are identically distributed and correspond to some unrelated process. The first synthetic Dataset (Synth 2D) shown in Figure 2(a), contains two dimensional data for visualization purposes. While varying scales on the axes are used to demonstrate the influence of noise in the two dimensional setting, it directly translates to problems arising from small noise levels in large dimensional data. We demonstrate this on the second synthetic dataset with

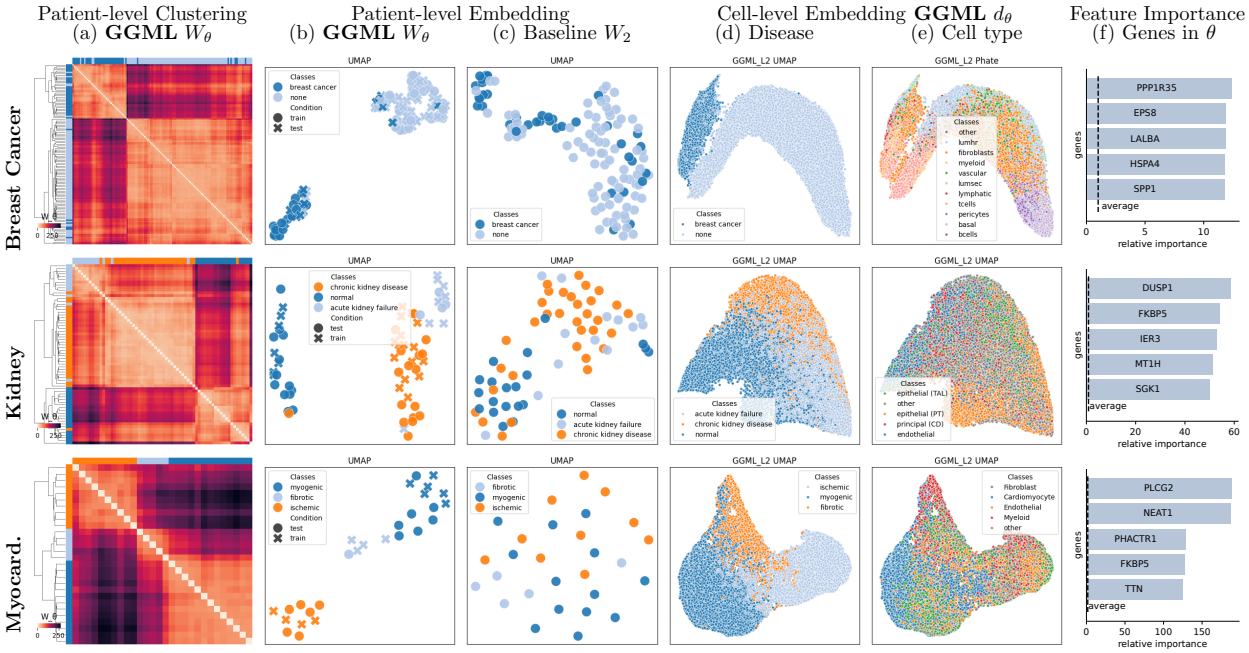


Figure 3: Embeddings of Patients and (single) Cells for scRNA-seq data from different diseases using d_θ learned by Global Ground Metric Learning (GGML) and Euclidean d_2 as baseline. To highlight the capabilities of GGML to generalize to unseen data the shown plots are produced by only learning on half of the data points as indicated. Relative weights of θ can be directly interpreted as gene importance in distinguishing disease stages.

200 identically scaled dimensions. Hyperparameters are tuned with a grid search evaluated on 20% validation splits as shown in Figure 5.

4.2 scRNA-seq Datasets

Single-cell RNA sequencing data contains measured gene expression vectors $\boldsymbol{x}$ of single cells from different patients. Disease (i.e. class) labels c_i are generally provided as pathological annotations of the respective tissue and correspond to classes on the patient-level. This setup fits naturally in our framework with sampled cells $\boldsymbol{x_i} \in \Omega$ from a multimodal patient-specific distribution of cells X_i for some patient i. Similar to the synthetic data, the assumption is that disease stages are not distinguishable for cells (samples) from cell types (modes) that are not disease-relevant, presenting a realistic and challenging scenario for testing our approach. We perform our experimental evaluation on datasets from three distinct diseases: Breast Cancer [Sikkema et al., 2023], Kidney Disease [Lake et al., 2023] and Myocardial Infarction [Kuppe et al., 2022]. This diverse collection of large datasets allows us to benchmark our framework across a variety of biological contexts and disease states. The datasets contain 31 to 132 tissue samples with at total of 191k to 714k cells from which we sample 1000 cells per tissue. The datasets included 28,975 to 33,145 genes per cell. We selected only those genes with above-average variance, resulting in a reduced dimensionality of 7,734 to 8,433 genes.

4.3 Results

We present the results of the evaluation on the synthetic datasets in Figure 2 and the scRNA datasets in Figure 3. The classification results are found in Table 1. To reproduce the presented results or use GGML on other data, the code is provided on GitHub[1].

Classification Table 1 presents classification results of various methods using a kNN classifier on 10 test-train splits of the patient-level and cell-level data. For patient-level data, the GGML method consistently outperforms other methods such as Euclidean (Eucl.), Manhattan (Manh.), Cosine (Cos.), LMNN, LFDA, NCA, and ITML across different datasets. It achieves high accuracies on synthetic data (2D: 0.96±0.04, 200D: 0.95±0.11) and real-world genomics datasets from different diseases (Kidney: 0.94±0.02, Breastcancer: 0.91±0.04, Myocard: 0.92±0.08). On the cell-level, GGML again shows superior performance on all datasets (Synth: 0.53±0.01, Kidney: 0.74±0.00, Breastcancer: 0.81±0.03, Myocard: 0.94±0.00). As expected, differentiating disease states at the cell-level is a more challenging task as not all cells are involved or affected by the disease. We indicate that certain

[1] github.com/DaminK/GlobalGround-MetricLearning

Table 1: kNN Classification accuracy on patient- and (single) cell-level for competing metric learning methods. Accuracy is given as mean±variance over 10 test-train splits. *OOM/T* stands for *Out of Memory/Time*.

	Method	Synth_{2D}	Synth_{200D}	Kidney	Brst.Canc.	Myocard.	Synth_{2D}	Synth_{200D}	Kidney	Brst.Canc.	Myocard.
patient-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**

cell-level

methods were unable to process large-scale scRNA-seq data within the available memory limits (1TB) or computational time constraints (64 cores, 8 hours). This underscores the need for efficient algorithms capable of handling high-dimensional biological data.

Clustering of patients using the learned W from GGML are shown for the 2D synthetic dataset in Figure 2(e), and the sccRNA datasets in Figure 3(a). The shown cluster maps clearly differentiate disease states at the patient level for Kidney, Breast Cancer, and Myocardial Infarction. Notably, the clusters meaningfully capture all patients, even though GGML was only trained on half of the patients as indicated in Figure 3(b). Similar to classification, GGML demonstrates superior performance compared other methods in these clustering tasks based on standard clustering metrics. Detailed results on the clustering performance can be found in the appendix subsection A.2.

Embedding of the learned Wasserstein distances between the distributions (patients) shows well separated classes in the synthetic (Figure 2d) and scRNA data (Figure 3b). To highlight the capabilities of GGML to generalize to unseen data, shown embeddings are produced by only learning on half of the data points as indicated. For the baseline of using the euclidean d_2 as a predefined ground metric, the embeddings (Figure 2f,3c) show no clear separations of classes or disease states. Embeddings between elements shows that GGML learns to differentiate the only differentiable mode (center) in the synthetic data (Figure 2b) unlike all other competing methods. In the synthetic data, GGML is the only method that captures the classes-related variances along a single axis shown by the first Principal Components of learned spaces from different methods. In the cell-level of the scRNA data, we only show embeddings of cells that are classified with 90% accuracy over the train-test splits as a proxy for being involved in the disease progression. Baseline embeddings of all cells, including competing methods, can be found in the appendix A.1.

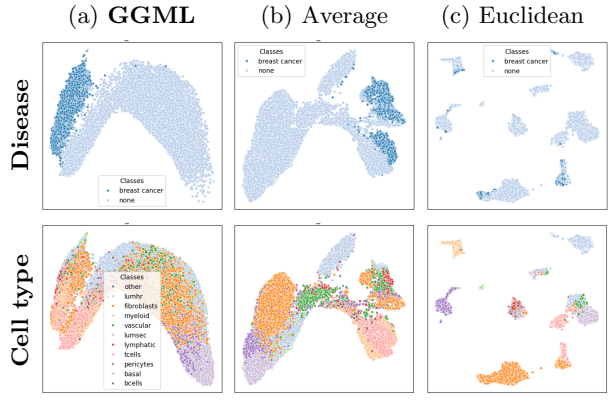


Figure 4: GGML(a) learns a Mahalanobis distance that differentiates cells by disease state. Cell types are defined by the Euclidean distance(c) which does not capture disease states. The average of both distances(b) differentiates cells by disease state and type.

Relating Disease State and Cell Types Analyzing scRNA data is commonly done by clustering cells into distinct cell types that are assumed to serve different functions in the tissue. These clusterings often use the Euclidean distance over all genes or their principal components as measure of cell similarity. Note that the Euclidean distance is a Mahalanobis distance with identity matrix. To relate different Mahalanobis distances over the same cells, we aim at computing averages as weighted linear combinations of the cell distances. Specifically, computing such averages enables us to fully interpolate between the GGML Mahalanobis distance that captures disease states and the Euclidean distance that captures cell types. Figure 4(b) shows that the average distance captures properties from both distances differentiating the disease state and cell type. It shows two large clusters of different cell types with disease states associated to breast cancer. This highlights its capabilities as a tool for researchers in computational biology where relating novel findings such a disease states to cell types can be crucial in understanding disease progressions.

Feature Importance For the low dimensional synthetic data, we can represent the learned Mahalanobis (covariance) matrix of different methods as 2D ellipses Figure 2(c). GGML recovers the ground

truth where only variance across the first axis explains the class differences on a distribution-level. For scRNA data, we show the 5 most important genes in differentiating each disease in Figure 3(f). The identified genes contain known markers for the respective diseases or are associated with related biological processes. In the disease progression of breast cancer, genes PPP1R35 and HSPA4 are associated with lymph node metastases [Mamoor, 2021, Gu et al., 2019], while EPS8 regulates cell migration and proliferation [Chen et al., 2015]. SPP1 has a role in angiogenesis and fibroblast activation in the tumor micro-environment [Butti et al., 2021]. In the kidney dataset, the identified genes impact both chronic kidney disease (CKD) and acute kidney injury (AKI). DUSP1 and FKBP5 were found to protect against AKI and CKD caused by ischemia [Shi et al., 2023] respective inflammatory responses [Xu and Wang, 2022]. IER3 affects the same signalling pathways related to inflammation [Arlt and Schäfer, 2011]. Elevated levels of SGK1 expression have been demonstrated to promote hypertrophy and fibrosis, leading to kidney damage in a mouse model [Sierra-Ramos et al., 2021]. In the disease progression of myocardial infarction, PHACTR1 is related to affecting arterial compliance which can be understood as flexibility of the vascular system [Wood et al., 2023]. NEAT1 is related to early onset of myocardial infarction involved in multiple pathways related to immune and inflammatory responses [Gast et al., 2019], with PLCG2 being related to similar processes causing immune dysregulation [Welzel et al., 2022]. FKBP5, a gene already found to be relevant in kidney disease, is also associated with increasing risk for myocardial infarction contributing to the same signal pathway [Zannas et al., 2019]. These findings highlight the overlap of individual biological processes and disease mechanisms whose varying interactions manifest into different diseases in different tissues. This first exemplary analysis of learned gene importance's highlights the interpretabilty of the learned projections as disease related subspaces.

4.4 Influence of Hyperparameters

As introduced in section 3, the learning process of GGML is mainly influenced by hyperparameters α (margin) and λ (regularization strength). Hyperparameters are tuned with a grid search on 20% validation splits from the 10 test-train-validation splits in the classification task. Figure 5 shows the respective classification performance for the high-dimensional synthetic dataset and the scRNA dataset on myocardial infarction. For the synthetic data with few dimensions a small L1 regularization strength ($\lambda < 1$) with moderate margins ($\alpha = 10$) achieves optimal performance. For the real-world scRNA data with many correlated

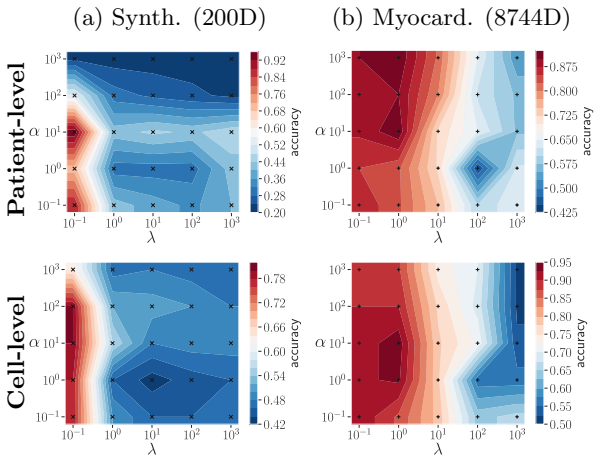


Figure 5: Grid search over hyperparameters α (margin) and λ (regularization strength) on: (a) synthetic data and (b) myocardial infarction scRNA data. The classification accuracy on validation splits shows unique local maxima with smooth surfaces in the considered ranges which enables an efficient tuning.

genes, optimization with a L1 norm becomes unstable. On the myocardial infarction dataset optimal results are achieved for a moderate L2 regularization ($\lambda = 1$). As expected, too strong regularization leads to underfitting in all considered datasets. In the noisy real-world datasets with large variances within classes (i.e. disease states) large margins α can lead to a better generalizability, and thus a higher classification accuracy in the validation split.

The hyperparameters k and t are strongly related to computational optimizations. k specifies the maximal rank of the learned decomposed matrix $\widetilde{\boldsymbol{W}}$ of the Mahalanobis matrix. t specifies the number of neighbors from each class that are considered to form triplets of relative relationships. Setting these hyperparameters correspond to a trade-off between computational complexity and quality of approximation. In the datasets presented, a 10-fold reduction in computation time is achieved while maintaining high classification performance. A more detailed ablation study w.r.t. computation times can be found in the appendix 3.

5 RELATED WORK

This section differentiates our proposed framework from existing methods in Global and Ground Metric Learning. Global Metric Learning learns a single distance metric tuned to a particular task [Yang and Jin, 2006, Kulis et al., 2013]. The metric is globally applicable to any two points as opposed to Local Metric Learning where multiple local metrics are learned with varying notions of locality

[Dong et al., 2019]. In this paper, we consider approaches that learn a Mahalanobis distances between data points from (multiple) class labels. Neighborhood Components Analysis (NCA)[Goldberger et al., 2004] and Large Margin Nearest Neighbor (LMNN) [Weinberger and Saul, 2009] maximize the margin between the data points from a class and differently labeled ones. While these methods optimize over a decomposed matrix (see Equation 2), Information Theoretic Metric Learning (ITML) [Davis et al., 2007] is regularizing with the log-determinant divergence to enforce positive semi-definiteness. All of these global approaches fail to handle heterogeneous data from multimodal distributions. Existing work addresses this through learning multiple local metrics for each mode or other notions of locality [Wang et al., 2012, Dong et al., 2019]. In this paper, we aim at learning a single metric that globally describes heterogeneous data.

To achieve this, we consider a different methodological approach by learning global metrics as ground metrics, rather than increasing the models' complexity by means of Local or Deep Metric Learning. While both approaches are promising future directions in ground metric learning, we believe that current OT methods can be significantly improved by ground metric learning of simple metrics such as the Mahalanobis distance as demonstrated in this work. A promising Deep Metric Learning to consider as part of future work is Neuronal Optimal Transport (NOT) which trains Input Convex Neural Networks to compute transport plans for an underlying ground metric [Korotin et al., 2022]. However, it is not a ground metric learning approach as it requires a fixed (weak) ground metric as a prior. Nonetheless, NOT has found promising applications on scRNA data [Bunne et al., 2023] which might further benefit from extending this approach with Ground Metric Learning.

The intially proposed Ground Metric Learning framework [Cuturi and Avis, 2014] learns a metric distance matrix between supports. It is not a learned metric function in a sense that it can compare two unseen data points. Rather, it learns specific distance values between supports of distributions that satisfy the metric property. A similar formulation was earlier introduced as supervised EMD in the context of computer vision [Wang and Guibas, 2012]. Another approach proposes the use of Singular Vectors to learn such distance values in an unsupervised manner [Huizing et al., 2022]. While these approaches do not require prior knowledge about the relationships between supports, they also can't leverage such information given by the supports position in some underlying space. Furthermore, by directly learning distances between supports such supervised approaches can not be applied to datasets without shared supports. For such datasets the resulting metric distance matrix contains n^2 values for n distinct supports which becomes under-determined and prone to overfitting. This is clearly the case for patient-level scRNA data where no identical cells are sampled for different patients which leads to disjoint supports. The Supervised Word Mover's Distance learns a Mahalanobis distance of word embeddings using an approach like NCA [Huang et al., 2016]. However, it also assumes shared supports in the form of histograms over a fixed vocabulary. Other recently proposed Ground ML methods learn geodesics to interpolate intermediate steps of trajectories [Scarvelis and Solomon, 2022, Kapusniak et al., 2024]. While these approaches yield promising results, they only apply to datasets where time scales are known which is generally not the case for diseases progressions in scRNA datasets.

6 CONCLUSION

We introduced a global ground metric learning framework, GGML, and demonstrated its strong performance and robustness using large synthetic and scRNA datasets from diverse diseases. The effectiveness of GGML in both classification and clustering of disease states underscores its capability to handle high-dimensional heterogeneous data. As the learned metrics generalize to unseen data, GGML has the potential to enhance performance in many existing Optimal Transport applications while minimizing the risk of overfitting.

Notably, GGML learns relationships on the distribution and element-level (sample points). For scRNA data, this means that GGML can accurately capture disease states at the level of patients (distribution) and cell-level (individual sample points) by implicitly learning relevant directions (subspaces) within the gene space along which distances best discriminate between classes. Using a low rank approximation, our method efficiently learn a low dimensional subspace with axes corresponding to distinct gene activation patterns. Weights in the learned subspace can be directly interpreted as gene importance, identifying disease related genes in all considered diseases. To illustrate this, we have provided an exemplary analysis of feature importance, and found results matching existing literature on the respective diseases. However, our framework is not limited to scRNA data: GGML has the potential to generate new insights in various other domains in which optimal transport is used with heterogeneous data.

Acknowledgments

This work was funded as part of the Graphs4Patients Consortia by the BMBF (Federal Ministry of Education and Research) and the Ministry of Culture and Science of North Rhine-Westphalia (NRW Rückkehrprogramm).

References

[Arlt and Schäfer, 2011] Arlt, A. and Schäfer, H. (2011). Role of the immediate early response 3 (ier3) gene in cellular stress response, inflammation and tumorigenesis. *European journal of cell biology*, 90(6-7):545–552.

[Bonneel and Digne, 2023] Bonneel, N. and Digne, J. (2023). A survey of optimal transport for computer graphics and computer vision. In *Computer Graphics Forum*, volume 42, pages 439–460. Wiley Online Library.

[Bunne et al., 2023] Bunne, C., Stark, S. G., Gut, G., Del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. (2023). Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768.

[Butti et al., 2021] Butti, R., Nimma, R., Kundu, G., Bulbule, A., Kumar, T. V., Gunasekaran, V. P., Tomar, D., Kumar, D., Mane, A., Gill, S. S., et al. (2021). Tumor-derived osteopontin drives the resident fibroblast to myofibroblast differentiation through twist1 to promote breast cancer progression. *Oncogene*, 40(11):2002–2017.

[Chen et al., 2015] Chen, C., Liang, Z., Huang, W., Li, X., Zhou, F., Hu, X., Han, M., Ding, X., and Xiang, S. (2015). Eps8 regulates cellular proliferation and migration of breast cancer. *International journal of oncology*, 46(1):205–214.

[Cuturi and Avis, 2014] Cuturi, M. and Avis, D. (2014). Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564.

[Davis et al., 2007] Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216.

[De Vazelhes et al., 2020] De Vazelhes, W., Carey, C., Tang, Y., Vauquier, N., and Bellet, A. (2020). metric-learn: Metric learning algorithms in python. *Journal of Machine Learning Research*, 21(138):1–6.

[Dong et al., 2019] Dong, M., Wang, Y., Yang, X., and Xue, J.-H. (2019). Learning local metrics and influential regions for classification. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1522–1529.

[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.

[Gast et al., 2019] Gast, M., Rauch, B. H., Haghikia, A., Nakagawa, S., Haas, J., Stroux, A., Schmidt, D., Schumann, P., Weiss, S., Jensen, L., et al. (2019). Long noncoding rna neat1 modulates immune cell functions and is suppressed in early onset myocardial infarction patients. *Cardiovascular research*, 115(13):1886–1906.

[Goldberger et al., 2004] Goldberger, J., Hinton, G. E., Roweis, S., and Salakhutdinov, R. R. (2004). Neighbourhood components analysis. *Advances in neural information processing systems*, 17.

[Gu et al., 2019] Gu, Y., Liu, Y., Fu, L., Zhai, L., Zhu, J., Han, Y., Jiang, Y., Zhang, Y., Zhang, P., Jiang, Z., et al. (2019). Tumor-educated b cells selectively promote breast cancer lymph node metastasis by hspa4-targeting igg. *Nature medicine*, 25(2):312–322.

[Guo et al., 2003] Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer.

[Harbrecht et al., 2012] Harbrecht, H., Peters, M., and Schneider, R. (2012). On the low-rank approximation by the pivoted cholesky decomposition. *Applied numerical mathematics*, 62(4):428–440.

[Hocke and Martinetz, 2014] Hocke, J. and Martinetz, T. (2014). Global metric learning by gradient descent. In *Artificial Neural Networks and Machine Learning–ICANN 2014: 24th International Conference on Artificial Neural Networks, Hamburg, Germany, September 15-19, 2014. Proceedings 24*, pages 129–135. Springer.

[Huang et al., 2016] Huang, G., Guo, C., Kusner, M. J., Sun, Y., Sha, F., and Weinberger, K. Q. (2016). Supervised word mover's distance. *Advances in neural information processing systems*, 29.

[Huizing et al., 2022] Huizing, G.-J., Cantini, L., and Peyré, G. (2022). Unsupervised ground metric learning using wasserstein singular vectors. In *International Conference on Machine Learning*, pages 9429–9443. PMLR.

[Joodaki et al., 2024] Joodaki, M., Shaigan, M., Parra, V., Bülow, R. D., Kuppe, C., Hölscher, D. L., Cheng, M., Nagai, J. S., Goedertier, M., Bouteldja, N., et al. (2024). Detection of patient-level distances from single cell genomics and pathomics data with optimal transport (pilot). *Molecular systems biology*, 20(2):57–74.

[Kapusniak et al., 2024] Kapusniak, K., Potaptchik, P., Reu, T., Zhang, L., Tong, A., Bronstein, M., Bose, A. J., and Di Giovanni, F. (2024). Metric flow matching for smooth interpolations on the data manifold. *arXiv preprint arXiv:2405.14780*.

[Kingma, 2014] Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Korotin et al., 2022] Korotin, A., Selikhanovych, D., and Burnaev, E. (2022). Neural optimal transport. *arXiv preprint arXiv:2201.12220*.

[Kulis et al., 2013] Kulis, B. et al. (2013). Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364.

[Kuppe et al., 2022] Kuppe, C., Ramirez Flores, R. O., Li, Z., Hayat, S., Levinson, R. T., Liao, X., Hannani, M. T., Tanevski, J., Wünnemann, F., Nagai, J. S., et al. (2022). Spatial multiomic map of human myocardial infarction. *Nature*, 608(7924):766–777.

[Lake et al., 2023] Lake, B. B., Menon, R., Winfree, S., Hu, Q., Melo Ferreira, R., Kalhor, K., Barwinska, D., Otto, E. A., Ferkowicz, M., Diep, D., et al. (2023). An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, 619(7970):585–594.

[Mamoor, 2021] Mamoor, S. (2021). Ppp1r35 is differentially expressed in the lymph node metastases of patients with breast cancer.

[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*.

[Murtagh and Contreras, 2012] Murtagh, F. and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97.

[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.

[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.

[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.

[Raudvere et al., 2019] Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198.

[Sato et al., 2022] Sato, R., Yamada, M., and Kashima, H. (2022). Re-evaluating word mover's distance. In *International Conference on Machine Learning*, pages 19231–19249. PMLR.

[Scarvelis and Solomon, 2022] Scarvelis, C. and Solomon, J. (2022). Riemannian metric learning via optimal transport. *arXiv preprint arXiv:2205.09244*.

[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.

[Schultz and Joachims, 2003] Schultz, M. and Joachims, T. (2003). Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16.

[Shi et al., 2023] Shi, L., Zha, H., Pan, Z., Wang, J., Xia, Y., Li, H., Huang, H., Yue, R., Song, Z., and Zhu, J. (2023). Dusp1 protects against ischemic acute kidney injury through stabilizing mtdna via interaction with jnk. *Cell Death & Disease*, 14(11):724.

[Sierra-Ramos et al., 2021] Sierra-Ramos, C., Velazquez-Garcia, S., Keskus, A. G., Vastola-Mascolo, A., Rodríguez-Rodríguez, A. E., Luis-Lima, S., Hernández, G., Navarro-González, J. F., Porrini, E., Konu, O., et al. (2021). Increased sgk1 activity potentiates mineralocorticoid/nacl-induced kidney injury. *American Journal of Physiology-Renal Physiology*, 320(4):F628–F643.

[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.

[Wang and Guibas, 2012] Wang, F. and Guibas, L. J. (2012). Supervised earth mover's distance learning and its computer vision applications. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 442–455. Springer.

[Wang et al., 2012] Wang, J., Kalousis, A., and Woznica, A. (2012). Parametric local metric learning for nearest neighbor classification. *Advances in neural information processing systems*, 25.

[Weinberger and Saul, 2009] Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2).

[Welzel et al., 2022] Welzel, T., Oefelein, L., Holzer, U., Müller, A., Menden, B., Haack, T. B., Groβ, M., and Kuemmerle-Deschner, J. B. (2022). Variant in the plcg2 gene may cause a phenotypic overlap of aplaid/plaid: case series and literature review. *Journal of clinical medicine*, 11(15):4369.

[Wood et al., 2023] Wood, A., Antonopoulos, A., Chuaiphichai, S., Kyriakou, T., Diaz, R., Al Hussaini, A., Marsh, A.-M., Sian, M., Meisuria, M., McCann, G., et al. (2023). Phactr1 modulates vascular compliance but not endothelial function: a translational study. *Cardiovascular research*, 119(2):599–610.

[Xu and Wang, 2022] Xu, H. and Wang, Z. (2022). Microrna-23a-3p ameliorates acute kidney injury by targeting fkbp5 and nf-κb signaling in sepsis. *Cytokine*, 155:155898.

[Yang and Jin, 2006] Yang, L. and Jin, R. (2006). Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2(2):4.

[Zannas et al., 2019] Zannas, A. S., Jia, M., Hafner, K., Baumert, J., Wiechmann, T., Pape, J. C., Arloth, J., Ködel, M., Martinelli, S., Roitman, M., et al. (2019). Epigenetic upregulation of fkbp5 by aging and stress contributes to nf-κb–driven inflammation and cardiovascular risk. *Proceedings of the National Academy of Sciences*, 116(23):11370–11379.

[Zantedeschi et al., 2016] Zantedeschi, V., Emonet, R., and Sebban, M. (2016). Lipschitz continuity of mahalanobis distances and bilinear forms. *arXiv preprint arXiv:1604.01376*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**, see section 2, section 3 and section C.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**, reducing the cubic scaling of optimized terms to linear section 3, and a low rank approximation of the learned Mahalanobis distance in subsection 3.1. Results on computational improvements can be found in Table 3.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **Yes**, in section B.
   (c) Clear explanations of any assumptions. **Yes**, in section 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes**, see github.com/DaminK/GlobalGround-MetricLearning.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**, in section 4 and section C
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**, in section 4 and section A
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**, in section C

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes**, used packages, notably Python Optimal Transport (POT) and PyTorch, are cited.
   (b) The license information of the assets, if applicable. **Yes**, MIT for POT and BSD-3 for PyTorch.
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable**.
   (d) Information about consent from data providers/curators. **Yes**, we used published and publicly available scRNA datasets from CEllxGENE.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable**.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**.

SUPPLEMENTARY MATERIALS

A RESULTS

A.1 Embeddings

Figure 6 shows the patient-level and cell-level embeddings of all competing methods on the breastcancer scRNA dataset. Methods are trained as before on a train split of the data. For visualization purposes the learned (and unsupervised) metrics are shown on all patients. We also show the embedding of all single cells as a baseline. Figure 3 only showed cells that were correctly classified with at least 90% across test-splits. This indicates whether cells are distinguishable between disease stages and served as proxy for being involved in the different disease stages. NCA and LMNN did not terminate on this dataset due to time or memory constraints as seen in Table 2.

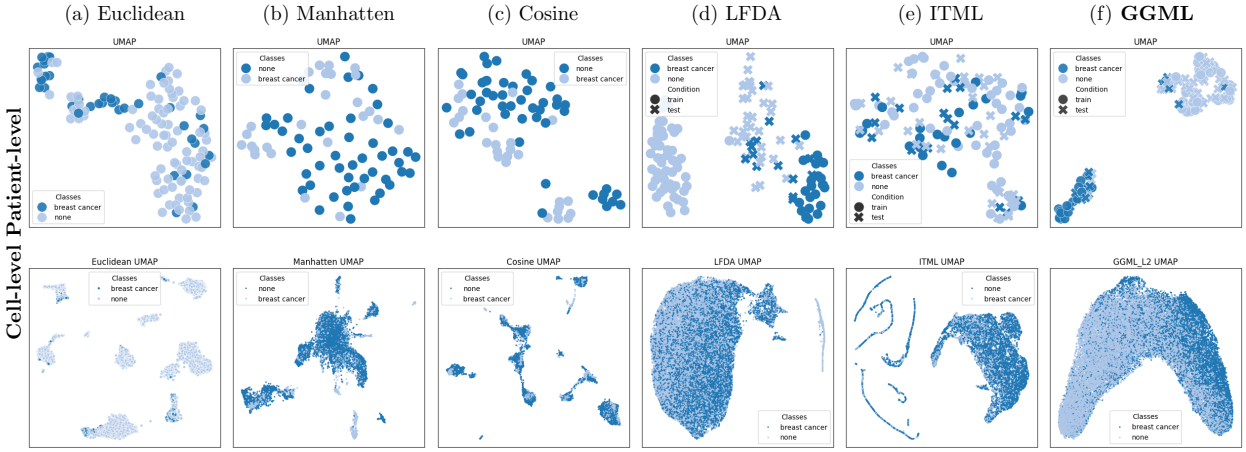


Figure 6: Embeddings of Patients and (single) Cells for scRNA-seq data from different diseases using d_θ learned by Global Ground Metric Learning (GGML) and Euclidean d_2 as baseline. To highlight the capabilities of GGML to generalize to unseen data the shown plots are produced by only learning on half of the data points as indicated. Relative weights of θ can be directly interpreted as gene importance in distinguishing disease stages.

Table 2: Clustering performance of all considered metrics on the patient- and cell-level, evaluated against patient disease-states as proxy for an unknown ground truth. GGML consistently outperforms other methods, as measured by Mutual Information (MI), Adjusted Rand Index (ARI), and Variation of Information (VI).

		Synth_{2D}			Synth_{200D}			Kidney			Brst.Canc.			Myocard.		
Method		MI	ARI	VI	MI	ARI	VI	MI	ARI	VI	MI	ARI	VI	MI	ARI	VI
Euclidean	*patient-level*	0.00	-0.07	2.19	0.00	-0.07	2.13	0.07	0.02	1.11	0.01	-0.02	0.68	0.11	0.02	1.21
Manhatten		0.00	-0.07	2.13	0.00	-0.07	2.19	0.07	0.02	1.11	0.03	0.13	0.93	0.22	0.17	1.00
Cosine		0.03	-0.03	1.97	0.07	-0.00	1.24	0.07	0.02	1.11	0.01	-0.02	0.68	0.07	0.00	1.19
LMNN		0.01	-0.07	2.17	0.01	-0.06	2.17		*OOM*			*OOM*			*OOM*	
LFDA		0.03	-0.05	2.07	0.08	-0.01	1.40	0.05	-0.01	1.39	0.06	0.13	0.68	0.64	0.53	0.75
NCA		0.01	-0.06	2.15	0.04	-0.03	2.10		*OOT*			*OOT*		0.87	0.77	0.37
ITML		0.18	0.02	1.68	0.22	0.11	1.38	0.06	0.01	1.27	0.01	-0.02	0.68	0.09	0.03	1.15
GGML		**1.10**	**1.00**	**0.00**	**1.10**	**1.00**	**0.00**	**0.65**	**0.66**	**0.60**	**0.21**	**0.49**	0.79	**0.91**	**0.92**	**0.20**
Euclidean	*cell-level*	0.00	-0.00	1.78	0.00	-0.00	1.79	0.18	0.00	3.17	**0.11**	-0.00	2.87	0.29	0.11	2.83
Manhatten		0.00	-0.00	2.18	0.00	-0.00	1.79	0.15	0.01	3.12	0.06	0.02	2.33	0.27	0.11	2.25
Cosine		0.05	0.02	2.03	0.34	0.01	3.05	0.18	0.00	3.09	**0.11**	0.03	2.91	0.33	0.09	3.40
LMNN		0.01	0.01	2.15	0.00	0.00	1.77		*OOM*			*OOM*			*OOM*	
LFDA		0.06	0.03	1.87	0.00	-0.00	2.40	0.04	-0.01	2.74	0.04	**0.06**	2.47	0.51	0.64	0.71
NCA		0.00	0.00	2.16	0.10	0.06	1.93		*OOT*			*OOT*		0.23	0.04	1.50
ITML		0.15	0.09	1.82	0.01	0.00	2.46	0.04	0.00	2.47	0.02	0.04	2.27	0.07	0.01	2.77
GGML		**0.77**	**0.75**	**0.65**	**0.68**	**0.64**	**0.83**	**0.24**	**0.22**	**1.67**	0.06	**0.06**	**1.61**	**0.63**	**0.71**	**0.81**

A.2 Clustering

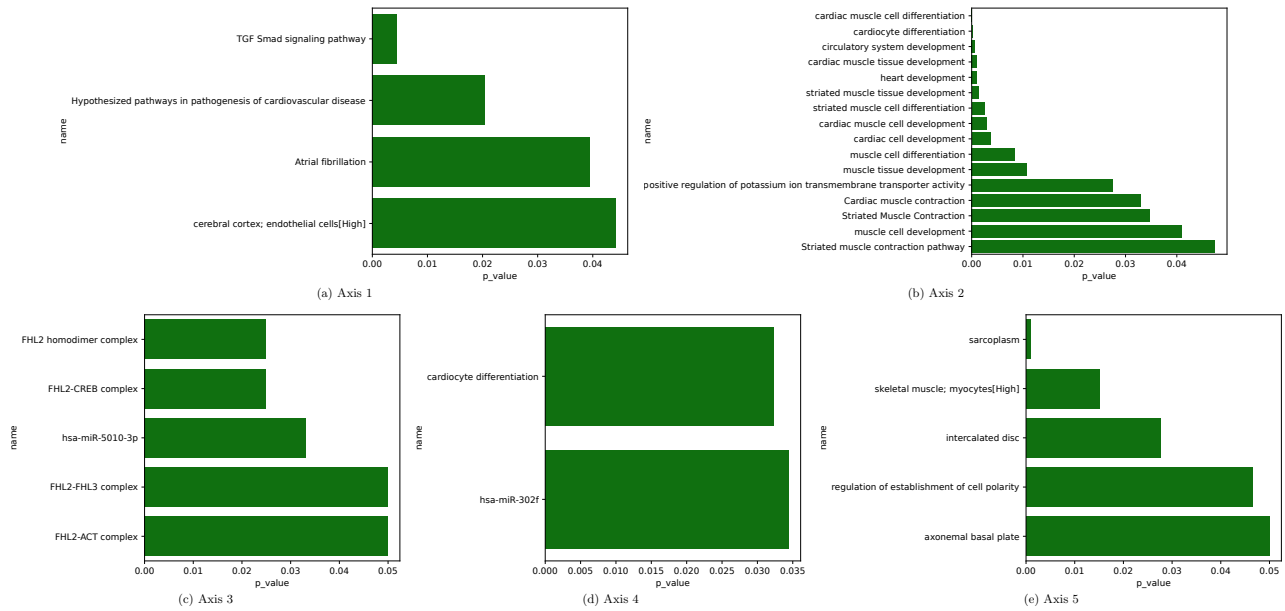
This section contains detailed clustering results on the presented datasets shown in Table 2. Patient-level clusters corresponded to the different disease states of the tissue sample. On a single cell-level, cells may form disease-related sub-clusters while unrelated cells might mix into additional clusters. To account for this, we do not enforce a fixed number of clusters for agglomerative hierarchical clustering at the cell level. Instead, we aggregate clusters until reaching a threshold based on the median distance of all pairwise global distances in the dataset. For synthetic data, the number of clusters is predetermined by construction at both levels. To establish a baseline with reliable ground truth, the two non-differentiable modes (corners in Figure 2(a)) are excluded and only the modes that can be differentiated are clustered. The results shown her use parameters $\alpha = 1, \lambda = 1$ for the synthetic data and $\alpha = 100, \lambda = 100$ for the scRNA data and do not reflect the hyperparameter tuning for the classification.

The clustering is evaluated using the Mutual Information (MI), Adjusted Rand Index (ARI), and Variational Information (VI). The MI quantifies the shared information between true labels U and cluster V assignments by computing $MI(U;V) = \sum_{i,j} P(i,j) \log \frac{P(i,j)}{P(i)P(j)}$, where $P(i,j)$ is the joint probability distribution of the clusters. ARI measures the ratio of pairs of samples that are correctly classified, adjusted for random chance. It is given as $ARI = \frac{RI - E[RI]}{max(RI) - E[RI]}$, where $E[RI]$ is the expected value of the Rand Index (RI). The optimal ARI value is 1 for perfect clusterings w.r.t labels. VI can be calculated as $VI(U,V) = H(U) + H(V) - 2MI(U;V)$, where H is the entropy of the resp. clustering. VI is a distance measure between clusterings where 0 corresponds to perfect clusterings.

The results in Table 2 demonstrate that GGML outperforms other methods across all datasets at the patient level and achieves better overall performance at the cell level. Cell-level clustering is more challenging for all methods, as expected, due to the inclusion of unrelated cell types and the formation of sub-clusters among disease-related cell types. These sub-clusters are not well captured by the clustering metrics when evaluated against patient disease states, which we used as a proxy due to the lack of ground truth distances for cell-level disease states.

A.3 Biological Processes in learned Subspaces

The individual rows of $\widetilde{W}$ correspond to different axis of the learned subspace. The axes can be directly interpreted as distinct genes linked to different disease-related processes. We present an exemplary analysis of the learned subspace for the myocardial infarction dataset, highlighting numerous relevant processes. Notably, all results were derived using only single-cell gene expression data and patient-level disease labels for training GGML.



(a) Axis 1 (b) Axis 2 (c) Axis 3 (d) Axis 4 (e) Axis 5

Figure 7: Significantly enriched biological processes for the individual axes of the learned subspace in the myocardial infarction dataset. The processes highlight distinct and relevant pathways, demonstrating GGML's ability to uncover disease-related gene expression changes using only patient disease states.

To identify biological processes from the most important genes within each axis, we performed gene enrichment analysis. Identified genes were mapped against existing databases of functionally related genes and their corresponding processes using g:Profiler [Raudvere et al., 2019]. Figure 7 shows the significantly enriched processes for each axis.

Each axis contains distinct and interpretable processes relevant to myocardial infarction. For instance, axis 2 captures various heart muscle cell differentiation and development processes that drive disease progression, while axis 1 includes a process that's already hypothesized to drive cardiovascular diseases. These findings highlight GGML's promising capabilities and interpretability in identifying disease-related gene expression changes and associated biological processes.

B Proofs

Theorem 1. *The GGML loss is 0 if, and only if, the global ground metric d_θ in W_θ approximates the ground truth distances W^* with margin α.*

$$\mathcal{L}_\alpha(\theta, X, \tilde{\mathcal{T}}_t) = 0 \text{ iff } W_\theta \approx_\alpha W^*$$

We show that both sides are equivalent for $\alpha \in \mathbb{R}_{\geq 0}$.

$$\begin{aligned}
& W_\theta \approx_\alpha W^* \\
\Leftrightarrow \quad & \forall (i,j,k) \in \tilde{\mathcal{T}}_t : W_\theta(X_j, X_k) - W_\theta(X_i, X_j) \geq \alpha \\
\Leftrightarrow \quad & \forall (i,j,k) \in \tilde{\mathcal{T}}_t : \\
& W_\theta(X_i, X_j) - W_\theta(X_j, X_k) + \alpha \leq 0 \\
\Leftrightarrow \quad & \forall (i,j,k) \in \tilde{\mathcal{T}}_t : \\
& \max\left(W_\theta(X_i, X_j) - W_\theta(X_j, X_k) + \alpha, 0\right) = 0 \\
\Leftrightarrow \quad & \forall (i,j,k) \in \tilde{\mathcal{T}}_t : \mathcal{L}_\alpha(\theta, X, (i,j,k)) = 0 \\
\Leftrightarrow \quad & \sum_{t \in \tilde{\mathcal{T}}_t} \mathcal{L}_\alpha(\theta, X, t) = \mathcal{L}_\alpha(\theta, X, \tilde{\mathcal{T}}_t) = 0
\end{aligned}$$

The last equivalence holds true as $\mathcal{L}_\alpha(\theta, X, t) \geq 0$.

Theorem 2 [Triplet Loss Bound]. *For all distributions X and triplets $(i,j,k) \in \tilde{\mathcal{T}}_t$, there exist a θ such that $\mathcal{L}(\theta, X, (i,j,k))$ is at most α.*

For all distributions X and sets of triplets $\tilde{\mathcal{T}}_t$, the minimal loss $\mathcal{L}(\theta, X, \tilde{\mathcal{T}}_t)$ is at most $\alpha|\tilde{\mathcal{T}}_t|$.

Let $d_{\mathbb{0}} : \Omega^2 \mapsto 0$ be the zero function that maps pairs of elements $x, y \in \Omega$ to a distance of 0. With $d_{\mathbb{0}}$ as ground metric, it holds that $W_{\mathbb{0}}(X_i, X_j) = 0$ for any distributions X_i, X_j as $d_{\mathbb{0}}(x,y) = 0$ for all elements x, y. Clearly, it also holds for all triplets $\forall (i,j,k) \in \tilde{\mathcal{T}}_t$ that $W_{\mathbb{0}}(X_i, X_j) - W_{\mathbb{0}}(X_j, X_k) = 0$.

Consider that the zero matrix $\mathbb{0}$ is psd and thus a valid covariance matrix of the Mahalanobis distance. Hence, the zero function $d_{\mathbb{0}}$ is in the hypothesis space of learnable Mahalanobis ground metrics. It follows that:

$$\begin{aligned} & (i,j,k) \in \tilde{\mathcal{T}}_t : \min_\theta W_\theta(X_i, X_j) - W_\theta(X_j, X_k) \leq 0 \\ \implies \quad & (i,j,k) \in \tilde{\mathcal{T}}_t : \min_\theta \mathcal{L}_\alpha(\theta, X, (i,j,k)) \leq \alpha \end{aligned}$$

with the equality being satisfied for $\theta = \mathbb{0}$.

Given that $\exists \theta : \mathcal{L}(\theta, X, t) \leq \alpha$ and using the same reasoning as above, it follows:

$$\min_\theta \sum_{t \in \tilde{\mathcal{T}}_t} \mathcal{L}(\theta, X, t) \leq \min_\theta \mathcal{L}(\theta, X, t)\, |\tilde{\mathcal{T}}_t| \leq \alpha\, |\tilde{\mathcal{T}}_t|$$

with equality given for $\theta = \mathbb{0}$.

C Implementation & Computation

C.1 Neighborhood Parameters

In the context of using a kNN classifier with metric learning approaches, multiple neighborhood parameters arise for which we want to provide some clarification. For global metric learning, it refers to the neighbors of a data point, respective distribution for ground metric learning. We have introduced a neighborhood parameter t for our global ground metric learning on the distribution level determining on how many neighbors are used to train on. Refer to section 3 for details on how this parameter is used to construct the triplet sets containing relative relationships between distributions. Due to training on only half of the data and the presence of small classes in scRNA disease states, we have set $t = 3$. For the synthetic data, where classes are larger, we use $t = 5$. In the context of kNN classification, the k-closest neighbors predict the label of a data point or distribution. We used $k = 5$ for patient-level classification and $k = 100$ cell-level classification due to the significant differences in number of data points and expected heterogeneity. Note that in all other contexts we refer to the rank of the decomposed Mahalanobis matrix with k.

C.2 Packages

We use Python Optimal Transport [Flamary et al., 2021] and PyTorch [Paszke et al., 2019] to compute a differentiable Wasserstein distance w.r.t. to the parameter of the underlying ground metric. ADAM [Kingma, 2014] is used to optimize the loss function. Competing metric learning methods are taken from metric-learn [De Vazelhes et al., 2020]. Computation of fixed metrics and the evaluation of the classification and clustering benchmark is done with Scikit-learn [Pedregosa et al., 2011]. Gene enrichment is performed with g:Profiler [Raudvere et al., 2019].

C.3 Computing Infrastructure & Training Details

All results were computed on an internal computing node running Linux, equipped with 64 cores (AMD EPYC 7763 at 3.5 GHz) and 1 TB RAM. We train GGML using the ADAM optimizer with a learning rate of 0.01. The training is conducted on minibatches that contain 128 triplets. Training is stopped after 30 to 50 iterations depending on the size of the dataset. Each epoch takes approximately 130 seconds (Synth_{200D}) to 950 seconds (Breastcancer) to compute for rank $k = 5$ and $t = 3$ neighbors, as indicated in Table 3. The Breastcancer dataset is the largest considered dataset with 131 patients and 714k cells which are subsampled to 131k cells in our experiments.

Table 3: Ablation Study on computational improvements using a low rank approximation and a fixed numbers of neighbors, measured by average training time per epoch (in seconds). Parameters used in the results section of the paper are indicated in bold.

Panel A: Synth_{200D}

Neigh. \ Rank.	**5**	10	50	100	200(full)
1	12.9	14.5	15.8	17.0	22.3
3	130.4	132.5	143.3	156.0	195.8
5	348.6	361.3	365.2	417.3	525.4
7	646.0	687.6	789.3	866.7	1040.5
9(all)	1049.1	1119.8	1249.2	1446.2	1703.2

Panel B: Breastcancer

Neigh. \ Rank.	**5**	10	100	500	1000	8433(full)
1	148.9	145.1	189.0	276.2	572.0	OOM
2	390.4	473.8	666.2	1256.9	3009.9	OOM
3	955.4	1035.8	1548.8	2610.7	5877.5	OOM
4	1776.1	1924.6	2563.5	4665.4	10425.0	OOM
5	2587.5	3097.7	3519.5	7353.1	19159.3	OOM

In the benchmarking pipeline, we run competing method subsequently so that methods have access to the same resources. We suspended the training of a metric learning methods after 8 hours of computation ("Out of time") or trying to assigning more than 1TB of RAM ("Out of memory"). To improve computation time and make the benchmark more comparable, we perform a low-rank Approximation with a pivoted Cholesky decomposition [Harbrecht et al., 2012] for competing methods who can only learn full rank Mahalanobis matrices M. As not all competing methods utilize multi-threading in their provided implementations, we refrained from discussing computation times as an indicator of the methods computational efficiency.