Appendices

A Proofs

The propositions throughout the paper are stated in the specific context of the label-free alignment of the sets $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$. However, they are more general. Here, we restate them in a broader context with more details. We remark that the numbering of the statements corresponds to the numbering in the paper. We add several lemmas (numbered separately) that are used in the proofs.

We remark that part of our theoretical analysis holds for any well-defined Riemannian manifold. Specifically, propositions 1, 3, 4, 5, 6 and 8. However, propositions 2, 7, 9 and 10 are specific for the Lorentz model in hyperbolic space. Concretely, Prop. 2 provides a compact closed-form expression for the Riemannian translation, recasting it as a standard mean alignment in a linear vector space but with appropriate (nonstandard) scales. Prop. 7 and 9 are based on the mapping from the tangent space to the manifold and then to a Euclidean space under the particular constraint of Lorentzian orthogonality, and therefore, they are specific to the Lorentz model. Lastly, Prop. 10, which shows that if the discrepancy between the two sets is derived by translation, scaling and rotation, then our method can perfectly align the sets, is based on the rotation component (Prop. 7), and therefore, also specific to the Lorentz model.

A.1 Riemannian translation

Proposition 1. *Let $\boldsymbol{x}$ be the Riemannian mean of a set $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{L}^d\}_{i=1}^n$, given by $\boldsymbol{x} = m(\mathcal{X}) = \arg\min_{\boldsymbol{x}\in\mathbb{L}^d} \sum_{i=1}^n d^2_{\mathbb{L}^d}(\boldsymbol{x}, \boldsymbol{x}_i)$, and let $\boldsymbol{y} \in \mathbb{L}^d$. The map $\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}$ shifts the mean of the set $\mathcal{X}$ from $\boldsymbol{x}$ to $\boldsymbol{y}$, i.e., it satisfies*

$$\boldsymbol{y} = m(\{\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x}_i)\}_{i=1}^n). \tag{17}$$

Proof.

$$\begin{aligned} m(\{\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x}_i)\}_{i=1}^n) &= \arg\min_{\boldsymbol{z}\in\mathbb{L}^d} \sum_{i=1}^n d^2_{\mathbb{L}^d}(\boldsymbol{z}, \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x}_i)) \\ &= \arg\min_{\boldsymbol{z}\in\mathbb{L}^d} \sum_{i=1}^n d^2_{\mathbb{L}^d}(\boldsymbol{z}, \mathrm{Exp}_{\boldsymbol{y}}(\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)))) \\ &= \boldsymbol{y}, \end{aligned}$$

due to the fact that PT preserves the metric tensor. □

Lemma 1. *Given any two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{L}^d$, the function $f : \mathbb{L}^d \times \mathbb{L}^d \to \mathbb{R}$ defined by*

$$f(\boldsymbol{x}, \boldsymbol{y}) = -\boldsymbol{x}(1) + (-2\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{y} + 1)\boldsymbol{y}(1) \tag{18}$$

is strictly positive, where $\boldsymbol{H} = [-1, \boldsymbol{0}^\top; \boldsymbol{0}, \boldsymbol{I}_d]$.

Proof. We break the proof into two cases: (i) $\boldsymbol{y}(1) > \frac{\boldsymbol{x}(1)}{3}$ and (ii) $1 \leq \boldsymbol{y}(1) \leq \frac{\boldsymbol{x}(1)}{3}$. We remark that the first entry of $\boldsymbol{x} \in \mathbb{L}^d$ in the upper sheet of the hyperboloid model is lower bounded by 1. Showing that the function f is strictly positive is straight-forward in case $\boldsymbol{y}(1) > \frac{\boldsymbol{x}(1)}{3}$, because $\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{y}$ is upper bounded by -1:

$$f(\boldsymbol{x}, \boldsymbol{y}) = -\boldsymbol{x}(1) + (-2\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{y} + 1)\boldsymbol{y}(1) > -\boldsymbol{x}(1) + (-2\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{y} + 1)\frac{\boldsymbol{x}(1)}{3} > 0.$$

In case $1 \leq \boldsymbol{y}(1) \leq \frac{\boldsymbol{x}(1)}{3}$, by expanding the expressions, we have

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{y}) &= -\boldsymbol{x}(1) + (-2\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{y} + 1)\boldsymbol{y}(1) \\ &= -\boldsymbol{x}(1) + 2\boldsymbol{x}(1)\boldsymbol{y}^2(1) - 2\boldsymbol{y}(1)\sum_{i\geq 2} \boldsymbol{x}(i)\boldsymbol{y}(i) + \boldsymbol{y}(1). \end{aligned}$$

Using the Cauchy-Schwarz inequality gives

$$f(\boldsymbol{x},\boldsymbol{y}) \geq -\boldsymbol{x}(1) + 2\boldsymbol{x}(1)\boldsymbol{y}^2(1) + \boldsymbol{y}(1) - 2\boldsymbol{y}(1)\left(\sqrt{\sum_{i\geq 2}\boldsymbol{x}^2(i)}\sqrt{\sum_{i\geq 2}\boldsymbol{y}^2(i)}\right).$$

Since $\left(-\boldsymbol{x}(1) + 2\boldsymbol{x}(1)\boldsymbol{y}^2(1) + \boldsymbol{y}(1)\right)^2 > 4\boldsymbol{y}^2(1)(\boldsymbol{x}^2(1)-1)(\boldsymbol{y}^2(1)-1)$ and $-\boldsymbol{x}(1)+2\boldsymbol{x}(1)\boldsymbol{y}^2(1)+\boldsymbol{y}(1) \geq 0$, we have

$$f(\boldsymbol{x},\boldsymbol{y}) \geq -\boldsymbol{x}(1) + 2\boldsymbol{x}(1)\boldsymbol{y}^2(1) + \boldsymbol{y}(1) - 2\boldsymbol{y}(1)\left(\sqrt{\boldsymbol{x}^2(1)-1}\sqrt{\boldsymbol{y}^2(1)-1}\right) > 0.$$

□

Lemma 2. *Given any three points $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \mathbb{L}^d$, the function $f : \mathbb{L}^d \times \mathbb{L}^d \times \mathbb{L}^d \to \mathbb{R}$ defined by*

$$f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{z} - (-2\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{y} + 1)\boldsymbol{y}^\top \boldsymbol{H}\boldsymbol{z} \tag{19}$$

is strictly positive, where $\boldsymbol{H} = [-1, \boldsymbol{0}^\top; \boldsymbol{0}, \boldsymbol{I}_d]$.

Proof. Let $L(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) + \lambda g(\boldsymbol{z})$, where $g(\boldsymbol{z}) = \boldsymbol{z}^\top \boldsymbol{H}\boldsymbol{z} + 1$. Solving

$$\frac{\partial L(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{\partial \boldsymbol{z}} = \boldsymbol{0}^\top$$

yields

$$\boldsymbol{z}^* = \frac{1}{2\lambda}\left(-\boldsymbol{x} + (-2\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{y} + 1)\boldsymbol{y}\right).$$

Based on Lemma 1 and $\boldsymbol{z}(1) > 0\ \forall \boldsymbol{z} \in \mathbb{L}^d$, we have $\lambda > 0$. Moreover, demanding $g(\boldsymbol{z}^*) = 0$ implies that $\lambda = \sqrt{\frac{1-\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{y}}{2}}$. Therefore, by denoting $\alpha = -\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{y}$, we can show that:

$$\begin{aligned} f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}^*) &= \frac{1}{2\lambda}(\boldsymbol{x} - (2\alpha+1)\boldsymbol{y})^\top \boldsymbol{H}(-\boldsymbol{x} + (2\alpha+1)\boldsymbol{y}) \\ &= \frac{1}{2\lambda}(1 + 4\alpha^2 + 4\alpha + 1 - 4\alpha^2 - 2\alpha) \\ &= \frac{1}{2\lambda}(2 + 2\alpha) > 0. \end{aligned}$$

□

Proposition 2. *Let $\boldsymbol{x}$ be the Riemannian mean of a set $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$, and let $\boldsymbol{y} \in \mathbb{L}^d$. The map $\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}$ can be recast as:*

$$\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{x}_i) = \boldsymbol{x}_i - \beta(\boldsymbol{x}_i|\boldsymbol{y},\boldsymbol{x})\boldsymbol{x} + \gamma(\boldsymbol{x}_i|\boldsymbol{y},\boldsymbol{x})\boldsymbol{y}, \tag{20}$$

for all $\boldsymbol{x}_i \in \mathcal{X}$, where

$$0 < \beta(\boldsymbol{x}_i|\boldsymbol{y},\boldsymbol{x}) = -\left\langle \frac{\boldsymbol{x}+\boldsymbol{y}}{\alpha+1}, \boldsymbol{x}_i \right\rangle_{\mathcal{L}},\ 0 < \gamma(\boldsymbol{x}_i|\boldsymbol{y},\boldsymbol{x}) = \left\langle \frac{\boldsymbol{y} - (2\alpha+1)\boldsymbol{x}}{\alpha+1}, \boldsymbol{x}_i \right\rangle_{\mathcal{L}}$$

and $0 < \alpha = -\langle \boldsymbol{x},\boldsymbol{y}\rangle_{\mathcal{L}}$.

Proof. We first derive the compact closed-form expression of the map $\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}$, and then, show that the functions β and γ are positive.

i. The map $\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}$ can be explicitly expressed by using the definition of the Exponential map and the Logarithmic map

$$\begin{aligned} &\mathrm{Exp}_{\boldsymbol{y}}(\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i))) \\ &= \mathrm{Exp}_{\boldsymbol{y}}\left(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i) + \frac{\langle \boldsymbol{y}, \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)\rangle_{\mathcal{L}}}{\alpha+1}(\boldsymbol{x}+\boldsymbol{y})\right) \end{aligned}$$

$$= \cosh\left(\left|\left|\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)\right|\right|_{\mathcal{L}}\right)\boldsymbol{y} + \sinh\left(\left|\left|\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)\right|\right|_{\mathcal{L}}\right)\frac{\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i) + \frac{\langle \boldsymbol{y}, \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)\rangle_{\mathcal{L}}}{\alpha+1}(\boldsymbol{x}+\boldsymbol{y})}{\left|\left|\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)\right|\right|_{\mathcal{L}}}.$$

Since the Lorentzian norm of vectors on the tangent space can be written as $||\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)||_{\mathcal{L}} = \cosh^{-1}(-\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle_{\mathcal{L}})$, we have

$$\begin{aligned}
&\mathrm{Exp}_{\boldsymbol{y}}(\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i))) \\
&= -\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle_{\mathcal{L}}\boldsymbol{y} + \frac{\sqrt{(-\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle_{\mathcal{L}})^2 - 1}}{\cosh^{-1}(-\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle_{\mathcal{L}})}\left(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i) + \frac{\langle \boldsymbol{y}, \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)\rangle_{\mathcal{L}}}{\alpha+1}(\boldsymbol{x}+\boldsymbol{y})\right) \\
&= -\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle_{\mathcal{L}}\boldsymbol{y} + \boldsymbol{x}_i + \langle \boldsymbol{x}, \boldsymbol{x}_i\rangle_{\mathcal{L}}\boldsymbol{x} + \frac{\langle \boldsymbol{y}, \boldsymbol{x}_i\rangle_{\mathcal{L}} - \alpha\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle_{\mathcal{L}}}{\alpha+1}(\boldsymbol{x}+\boldsymbol{y}) \\
&= \boldsymbol{x}_i - \beta(\boldsymbol{x}_i|\boldsymbol{y}, \boldsymbol{x})\boldsymbol{x} + \gamma(\boldsymbol{x}_i|\boldsymbol{y}, \boldsymbol{x})\boldsymbol{y}.
\end{aligned}$$

ii. We can show that $\beta(\boldsymbol{x}_i|\boldsymbol{y}, \boldsymbol{x}) > 0$ by

$$\beta(\boldsymbol{x}_i|\boldsymbol{y}, \boldsymbol{x}) = -\left\langle \frac{\boldsymbol{x}+\boldsymbol{y}}{\alpha+1}, \boldsymbol{x}_i \right\rangle_{\mathcal{L}} = -\frac{1}{\alpha+1}(\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle_{\mathcal{L}} + \langle \boldsymbol{y}, \boldsymbol{x}_i\rangle_{\mathcal{L}}).$$

Since $\alpha > 0$ and the Lorentzian inner product for any two points in $\mathbb{L}^d$ is smaller than or equal to -1, we obtain $\beta(\boldsymbol{x}_i|\boldsymbol{y}, \boldsymbol{x}) > 0$.

iii. The function $\gamma(\boldsymbol{x}_i|\boldsymbol{y}, \boldsymbol{x})$ is positive due to Lemma 1 and 2.

□

Proposition 3. *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{L}^d$. The map $\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}$ preserves distances (i.e., it is an isometry):*

$$d_{\mathbb{L}^d}(\boldsymbol{z}_1, \boldsymbol{z}_2) = d_{\mathbb{L}^d}(\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}_1), \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}_2)), \tag{21}$$

for any two points $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{L}^d$.

Proof. Showing that the map $\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}$ is an isometry is equivalent to showing that the Lorentzian inner product is preserved due to $d_{\mathbb{L}^d}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \cosh^{-1}(-\langle \boldsymbol{z}_1, \boldsymbol{z}_2\rangle_{\mathcal{L}})$. By Prop. 2, we have

$$\begin{aligned}
&\langle \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}_1), \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}_2)\rangle_{\mathcal{L}} \\
&= \langle \tau\boldsymbol{y} + \boldsymbol{z}_1 - \tau\boldsymbol{x}, \tau'\boldsymbol{y} + \boldsymbol{z}_2 - \tau'\boldsymbol{x}\rangle_{\mathcal{L}} + \left\langle \frac{\langle \boldsymbol{y} - \alpha\boldsymbol{x}, \boldsymbol{z}_1\rangle_{\mathcal{L}}}{\alpha+1}(\boldsymbol{x}+\boldsymbol{y}), \tau'\boldsymbol{y} + \boldsymbol{z}_2 - \tau'\boldsymbol{x} \right\rangle_{\mathcal{L}} + \\
&\left\langle \tau\boldsymbol{y} + \boldsymbol{z}_1 - \tau\boldsymbol{x}, \frac{\langle \boldsymbol{y} - \alpha\boldsymbol{x}, \boldsymbol{z}_2\rangle_{\mathcal{L}}}{\alpha+1}(\boldsymbol{x}+\boldsymbol{y}) \right\rangle_{\mathcal{L}} + \left\langle \frac{\langle \boldsymbol{y} - \alpha\boldsymbol{x}, \boldsymbol{z}_1\rangle_{\mathcal{L}}}{\alpha+1}(\boldsymbol{x}+\boldsymbol{y}), \frac{\langle \boldsymbol{y} - \alpha\boldsymbol{x}, \boldsymbol{z}_2\rangle_{\mathcal{L}}}{\alpha+1}(\boldsymbol{x}+\boldsymbol{y}) \right\rangle_{\mathcal{L}}.
\end{aligned}$$

Denoting $\alpha = -\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\mathcal{L}}$, $\tau = -\langle \boldsymbol{x}, \boldsymbol{z}_1\rangle_{\mathcal{L}}$, $\tau' = -\langle \boldsymbol{x}, \boldsymbol{z}_2\rangle_{\mathcal{L}}$, $\xi = -\langle \boldsymbol{y}, \boldsymbol{z}_1\rangle_{\mathcal{L}}$, and $\xi' = -\langle \boldsymbol{y}, \boldsymbol{z}_2\rangle_{\mathcal{L}}$, we obtain

$$\begin{aligned}
&\langle \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}_1), \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}_2)\rangle_{\mathcal{L}} \\
&= \langle \boldsymbol{z}_1, \boldsymbol{z}_2\rangle_{\mathcal{L}} - \tau'\xi + \tau'\tau - \tau\xi' + \tau\tau' - 2\tau\tau' + 2\alpha\tau\tau' + \\
&\quad \frac{-\xi + \alpha\tau}{\alpha+1}(-\tau' - \xi') + \frac{-\xi' + \alpha\tau'}{\alpha+1}(-\tau - \xi) - 2\frac{(-\xi + \alpha\tau)(-\xi' + \alpha\tau')}{\alpha+1} \\
&= \langle \boldsymbol{z}_1, \boldsymbol{z}_2\rangle_{\mathcal{L}}.
\end{aligned}$$

□

Proposition 4. *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{L}^d$. The map $\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}$ preserves geodesic velocities:*

$$\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{v}_0) = \boldsymbol{v}_1, \tag{22}$$

where $\mathbb{L}^d \ni \boldsymbol{v}_0 = \mathrm{Exp}_{\boldsymbol{x}}(\psi'(0)) = \boldsymbol{y}$ and $\mathbb{L}^d \ni \boldsymbol{v}_1 = \mathrm{Exp}_{\boldsymbol{y}}(\psi'(1))$ are the counterparts of the geodesic velocities of the geodesic path $\psi(t)$ connecting $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{L}^d$.

Proof. Firstly, we can explicitly write

$$\psi'(t) = \cosh^{-1}(\alpha)\sinh(\cosh^{-1}(\alpha)t)\boldsymbol{x} + \cosh(\cosh^{-1}(\alpha)t)\ \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{y}),$$

where $\alpha = -\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\mathcal{L}}$. Then, for $t = 0$,

$$\psi'(0) = \cosh^{-1}(\alpha)\sinh(\cosh^{-1}(\alpha)0)\boldsymbol{x} + \cosh(\cosh^{-1}(\alpha)0)\ \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{y}) = \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{y}).$$

From the above equation and the definition of the parallel transport operator and the Logarithmic map, we have

$$\begin{aligned}\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\psi'(0)) &= \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{y}) + \frac{\langle \boldsymbol{y}, \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{y})\rangle_{\mathcal{L}}}{\alpha+1}(\boldsymbol{y}+\boldsymbol{x}) \\ &= \frac{\cosh^{-1}(\alpha)}{\sqrt{\alpha^2-1}}(\boldsymbol{y}-\alpha\boldsymbol{x}) + \frac{\cosh^{-1}(\alpha)}{\sqrt{\alpha^2-1}}\frac{\alpha^2-1}{\alpha+1}(\boldsymbol{y}+\boldsymbol{x}).\end{aligned}$$

Re-organizing the above equation gives

$$\begin{aligned}\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\psi'(0)) &= \alpha\ \frac{\cosh^{-1}(\alpha)}{\sqrt{\alpha^2-1}}\boldsymbol{y} - \frac{\cosh^{-1}(\alpha)}{\sqrt{\alpha^2-1}}\boldsymbol{x} \\ &= \frac{\cosh^{-1}(\alpha)}{\sqrt{\alpha^2-1}}(\alpha^2-1)\boldsymbol{x} + \alpha\ \frac{\cosh^{-1}(\alpha)}{\sqrt{\alpha^2-1}}(\boldsymbol{y}-\alpha\boldsymbol{x}) \\ &= \frac{\cosh^{-1}(\alpha)}{\sqrt{\alpha^2-1}}(\alpha^2-1)\boldsymbol{x} + \alpha\ \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{y}) \\ &= \cosh^{-1}(\alpha)\sinh(\cosh^{-1}(\alpha))\boldsymbol{x} + \alpha\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{y}) \\ &= \psi'(1).\end{aligned}$$

Combining the above expressions gives

$$\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{v}_0) = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\mathrm{Exp}_{\boldsymbol{x}}(\psi'(0))) = \mathrm{Exp}_{\boldsymbol{y}}(\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\psi'(0))) = \mathrm{Exp}_{\boldsymbol{y}}(\psi'(1)).$$

□

Proposition 5. *Consider two subsets $\mathcal{A}, \mathcal{B} \subset \mathcal{X} \subset \mathbb{L}^d$ and their translations $\widetilde{\mathcal{A}} = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\mathcal{A}), \widetilde{\mathcal{B}} = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\mathcal{B}) \subset \widetilde{\mathcal{X}} \subset \mathbb{L}^d$, where $\boldsymbol{x}$ is the Riemannian mean of the set $\mathcal{X}$ and $\boldsymbol{y} = m(\widetilde{\mathcal{X}})$. Let $\boldsymbol{a}_1 = m(\mathcal{A}), \boldsymbol{b}_1 = m(\mathcal{B}), \boldsymbol{a}_2 = m(\widetilde{\mathcal{A}})$, and $\boldsymbol{b}_2 = m(\widetilde{\mathcal{B}})$ be the Riemannian means of the subsets. Then,*

$$\Gamma_{\boldsymbol{x}\to\boldsymbol{y}} \circ \Gamma_{\boldsymbol{b}_1\to\boldsymbol{a}_1} = \Gamma_{\boldsymbol{b}_2\to\boldsymbol{a}_2} \circ \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}. \tag{23}$$

Proof. Let $\boldsymbol{z}$ be a point in the set $\mathcal{B}$. Based on Prop. 2, we have

$$\Gamma_{\boldsymbol{b}_2\to\boldsymbol{a}_2} \circ \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) - \beta\left(\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z})|\boldsymbol{a}_2, \boldsymbol{b}_2\right)\boldsymbol{b}_2 + \gamma\left(\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z})|\boldsymbol{a}_2, \boldsymbol{b}_2\right)\boldsymbol{a}_2.$$

By Prop. 1, we get that $\boldsymbol{a}_2 = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{a}_1)$ and $\boldsymbol{b}_2 = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{b}_1)$. Combining it with Prop. 3, we have $\beta\left(\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z})|\boldsymbol{a}_2, \boldsymbol{b}_2\right) = \beta(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)$ and $\gamma\left(\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z})|\boldsymbol{a}_2, \boldsymbol{b}_2\right) = \gamma(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)$. Therefore, we obtain

$$\Gamma_{\boldsymbol{b}_2\to\boldsymbol{a}_2} \circ \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) - \beta(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{b}_2 + \gamma(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{a}_2.$$

By Prop. 1, we have that $\boldsymbol{a}_2 = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{a}_1)$ and $\boldsymbol{b}_2 = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{b}_1)$. Using the closed-form expression of the RT (Prop. 2), we get

$$\begin{aligned}&\Gamma_{\boldsymbol{b}_2\to\boldsymbol{a}_2} \circ \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) \\ &= \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) - \beta(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\left(\boldsymbol{b}_1 - \beta(\boldsymbol{b}_1|\boldsymbol{y}, \boldsymbol{x})\boldsymbol{x} + \gamma(\boldsymbol{b}_1|\boldsymbol{y}, \boldsymbol{x})\boldsymbol{y}\right) + \\ &\quad \gamma(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\left(\boldsymbol{a}_1 - \beta(\boldsymbol{a}_1|\boldsymbol{y}, \boldsymbol{x})\boldsymbol{x} + \gamma(\boldsymbol{a}_1|\boldsymbol{y}, \boldsymbol{x})\boldsymbol{y}\right).\end{aligned}$$

Using the closed-form expression (Prop. 2) again gives $\Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) = \boldsymbol{z} - \beta(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})\boldsymbol{x} + \gamma(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x})\boldsymbol{y}$. In addition, because the functions β and γ, defined in Prop. 2, are linear functions and homogeneous functions of degree 1, we have

$$\begin{aligned}\Gamma_{\boldsymbol{b}_2\to\boldsymbol{a}_2} \circ \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) = \boldsymbol{z} &- \beta(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{b}_1 + \gamma(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{a}_1 \\ &- \beta\left(\boldsymbol{z} - \beta(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{b}_1 + \gamma(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{a}_1|\boldsymbol{y}, \boldsymbol{x}\right)\boldsymbol{x} \\ &+ \gamma\left(\boldsymbol{z} - \beta(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{b}_1 + \gamma(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{a}_1|\boldsymbol{y}, \boldsymbol{x}\right)\boldsymbol{y}.\end{aligned}$$

Now, Prop. 2 also entails that $\Gamma_{\boldsymbol{b}_1 \to \boldsymbol{a}_1}(\boldsymbol{z}) = \boldsymbol{z} - \beta(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{b}_1 + \gamma(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{a}_1$. Plugging it into the derivation concludes the proof:

$$\begin{aligned}\Gamma_{\boldsymbol{b}_2 \to \boldsymbol{a}_2} \circ \Gamma_{\boldsymbol{x} \to \boldsymbol{y}}(\boldsymbol{z}) &= \Gamma_{\boldsymbol{x} \to \boldsymbol{y}} \left(\boldsymbol{z} - \beta(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{b}_1 + \gamma(\boldsymbol{z}|\boldsymbol{a}_1, \boldsymbol{b}_1)\boldsymbol{a}_1\right) \\ &= \Gamma_{\boldsymbol{x} \to \boldsymbol{y}} \circ \Gamma_{\boldsymbol{b}_1 \to \boldsymbol{a}_1}(\boldsymbol{z}).\end{aligned}$$

□

We remark that considering the transport along the geodesic path is important. For example, an alternative approach could be to first center one set to the origin, and then re-center it to the mean of the other set. In Euclidean spaces, these two options are equivalent, however, they are significantly different when Riemannian geometry is considered, as we demonstrate next.

Consider the same setting as in Prop. 5 that includes two subsets $\mathcal{A}, \mathcal{B} \subset \mathbb{L}^d$ and their translations $\widetilde{\mathcal{A}} = \Gamma_{\boldsymbol{x} \to \boldsymbol{y}}(\mathcal{A}), \widetilde{\mathcal{B}} = \Gamma_{\boldsymbol{x} \to \boldsymbol{y}}(\mathcal{B}) \subset \mathbb{L}^d$. Let $\boldsymbol{a}_1 = m(\mathcal{A}), \boldsymbol{b}_1 = m(\mathcal{B}), \boldsymbol{a}_2 = m(\widetilde{\mathcal{A}})$, and $\boldsymbol{b}_2 = m(\widetilde{\mathcal{B}})$ be the Riemannian means of the subsets, and recall the origin point $\boldsymbol{\mu}_0 \in \mathbb{L}^d$. Replacing the transport along the geodesic path from $\boldsymbol{x}$ to $\boldsymbol{y}$ with a transport from $\boldsymbol{x}$ to the origin $\boldsymbol{\mu}_0$ (along the geodesic path) and then from $\boldsymbol{\mu}_0$ to $\boldsymbol{y}$ does not necessarily admit the commuting property of Proposition 5: $(\Gamma_{\boldsymbol{x} \to \boldsymbol{\mu}_0} \circ \Gamma_{\boldsymbol{\mu}_0 \to \boldsymbol{y}}) \circ \Gamma_{\boldsymbol{b}_1 \to \boldsymbol{a}_1} \neq \Gamma_{\boldsymbol{b}_2 \to \boldsymbol{a}_2} \circ (\Gamma_{\boldsymbol{x} \to \boldsymbol{\mu}_0} \circ \Gamma_{\boldsymbol{\mu}_0 \to \boldsymbol{y}})$. We remark that only when $\boldsymbol{\mu}_0$ is on the geodesic path $\psi(t)$ from $\boldsymbol{x}$ to $\boldsymbol{y}$ for any $t \in \mathbb{R}$, the commuting property holds.

A.2 Riemannian scaling

Lemma 3. *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{L}^d$ and $\boldsymbol{v} = \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{y}) \in \mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$. Let $\phi_{\boldsymbol{v}}$ be the geodesic path from $\boldsymbol{x}$ to $\boldsymbol{y}$ with the initial velocity $\boldsymbol{v}$. Then, for any $t \geq 0$*

$$d_{\mathbb{L}^d}\big(\boldsymbol{x}, \phi_{\boldsymbol{v}}(t)\big) = t\, d_{\mathbb{L}^d}(\boldsymbol{x}, \boldsymbol{y}). \tag{24}$$

Proof. By the definition of the geodesic distance, we have

$$\begin{aligned}d_{\mathbb{L}^d}\left(\boldsymbol{x}, \phi_{\boldsymbol{v}}(t)\right) &= \cosh^{-1}\left(-\langle \boldsymbol{x}, \phi_{\boldsymbol{v}}(t)\rangle_{\mathcal{L}}\right) \\ &= \cosh^{-1}\left(-\left\langle \boldsymbol{x}, \cosh(||\boldsymbol{v}||_{\mathcal{L}} t)\boldsymbol{x} + \sinh(||\boldsymbol{v}||_{\mathcal{L}} t)\frac{\boldsymbol{v}}{||\boldsymbol{v}||_{\mathcal{L}}}\right\rangle_{\mathcal{L}}\right).\end{aligned}$$

Using $\langle \boldsymbol{x}, \boldsymbol{v}\rangle_{\mathcal{L}} = 0$ (due to the Lorentzian orthogonality) and $\langle \boldsymbol{x}, \boldsymbol{x}\rangle_{\mathcal{L}} = -1$ (by the definition of the hyperboloid manifold space) yields

$$\begin{aligned}d_{\mathbb{L}^d}\left(\boldsymbol{x}, \phi_{\boldsymbol{v}}(t)\right) &= \cosh^{-1}\left(-\left\langle \boldsymbol{x}, \cosh(\cosh^{-1}(-\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\mathcal{L}})t)\boldsymbol{x}\right\rangle_{\mathcal{L}}\right) \\ &= \cosh^{-1}(\cosh(\cosh^{-1}(-\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\mathcal{L}})t)) \\ &= t\cosh^{-1}(-\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\mathcal{L}}) \\ &= t\, d_{\mathbb{L}^d}(\boldsymbol{x}, \boldsymbol{y}).\end{aligned}$$

□

Proposition 6. *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be the Riemannian means of the sets $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N_x}$ and $\mathcal{Y} = \{\boldsymbol{y}_i\}_{i=1}^{N_y}$, respectively, and let $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ be their Riemannian dispersions. Then, we have*

$$r(\{\Upsilon_{\boldsymbol{y}}^s(\boldsymbol{y}_i)\}_{i=1}^{N_y}) = d_{\mathcal{X}}, \tag{25}$$

where $s = \sqrt{\frac{d_{\mathcal{X}}}{d_{\mathcal{Y}}}}$.

Proof. Let ϕ_i be the geodesic path from $\boldsymbol{y}$ to $\boldsymbol{y}_i$ such that $\phi_i(0) = \boldsymbol{y}$ and $\phi_i(1) = \boldsymbol{y}_i$. Then, by Lemma 3, we obtain

$$\begin{aligned}r(\{\Upsilon_{\boldsymbol{y}}^s(\boldsymbol{y}_i)\}_{i=1}^{N_y}) &= \frac{1}{N_y}\sum_{i=1}^{N_y} d^2_{\mathbb{L}^d}\left(\boldsymbol{y}, \phi_i(s)\right) \\ &= \frac{1}{N_y}\frac{d_{\mathcal{X}}}{d_{\mathcal{Y}}}\sum_{i=1}^{N_y} d^2_{\mathbb{L}^d}\left(\boldsymbol{y}, \boldsymbol{y}_i\right) \\ &= d_{\mathcal{X}}.\end{aligned}$$

□

The following Lemma 4 does not appear in the paper, but it provides additional insight to the Riemannian scaling. See Appendix E for related comparisons with HPA and Euclidean Procrustes analysis on the tangent space.

Lemma 4. *Let $\boldsymbol{x}$ and $d_{\mathcal{X}}$ be the Riemannian mean and dispersion of a set $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N_x}$. The Riemannian scaling $\Upsilon_{\boldsymbol{x}}^{s}$ is equivalent to the modulation on the tangent space $\mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$ with the same scaling factor. That is,*

$$\operatorname{Log}_{\boldsymbol{x}}(\phi_i(s)) = s \operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i), \tag{26}$$

where ϕ_i is the geodesic path from $\boldsymbol{x}$ to $\boldsymbol{x}_i$ such that $\phi_i(0) = \boldsymbol{x}$ and $\phi_i(1) = \boldsymbol{x}_i$.

Proof. Let $\kappa = -\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle_{\mathcal{L}}$ and $\widetilde{\kappa} = -\langle \boldsymbol{x}, \phi_i(s) \rangle_{\mathcal{L}}$. By the definition of the Logarithmic map, we have

$$\operatorname{Log}_{\boldsymbol{x}}(\phi_i(s)) = \frac{\cosh^{-1}(\widetilde{\kappa})}{\sqrt{\widetilde{\kappa}^2 - 1}} (\cosh(||\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)||_{\mathcal{L}} s)\boldsymbol{x} + \sinh(||\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)||_{\mathcal{L}} s) \frac{\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)}{||\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)||_{\mathcal{L}}} - \widetilde{\kappa}\boldsymbol{x}).$$

Since $||\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)||_{\mathcal{L}} = \cosh^{-1}(\kappa)$, and $\widetilde{\kappa} = \cosh(s \cosh^{-1}(\kappa))$ by Lemma 3, we have

$$\operatorname{Log}_{\boldsymbol{x}}(\phi_i(s)) = \frac{s \cosh^{-1}(\kappa)}{\sinh(\cosh^{-1}(\cosh(s \cosh^{-1}(\kappa))))} \left(\sinh(s \cosh^{-1}(\kappa)) \frac{\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)}{\cosh^{-1}(\kappa)} \right).$$

Re-organizing the above equation, we get

$$\begin{aligned} \operatorname{Log}_{\boldsymbol{x}}(\phi_i(s)) &= \frac{s \cosh^{-1}(\kappa)}{\sinh(s \cosh^{-1}(\kappa))} \left(\sinh(s \cosh^{-1}(\kappa)) \frac{\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i)}{\cosh^{-1}(\kappa)} \right) \\ &= s \operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{x}_i). \end{aligned}$$

□

A.3 Riemannian wrapped rotation

Proposition 7. *Given a point $\boldsymbol{x} \in \mathbb{L}^d$ and a rotation matrix $\boldsymbol{U} \in \mathbb{O}(d)$, the wrapped rotation $\Theta_{\boldsymbol{x}}^{\boldsymbol{U}}$ is bijective and its inverse is given by*

$$(\Theta_{\boldsymbol{x}}^{\boldsymbol{U}})^{-1} = \Theta_{\boldsymbol{x}}^{\boldsymbol{U}^\top}. \tag{27}$$

Proof. Let $\boldsymbol{z}$ be a point in $\mathbb{L}^d$. We have

$$\begin{aligned} (\Theta_{\boldsymbol{x}}^{\boldsymbol{U}})^{-1}(\Theta_{\boldsymbol{x}}^{\boldsymbol{U}}(\boldsymbol{z})) &= \Theta_{\boldsymbol{x}}^{\boldsymbol{U}^\top}(\Theta_{\boldsymbol{x}}^{\boldsymbol{U}}(\boldsymbol{z})) \\ &= \operatorname{Exp}_{\boldsymbol{x}}(\mathcal{P}_{\boldsymbol{x}}^{-1}(\boldsymbol{U}(\mathcal{P}_{\boldsymbol{x}}(\operatorname{Log}_{\boldsymbol{x}}(\operatorname{Exp}_{\boldsymbol{x}}(\mathcal{P}_{\boldsymbol{x}}^{-1}(\boldsymbol{U}^\top(\mathcal{P}_{\boldsymbol{x}}(\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{z}))))))))))) \\ &= \operatorname{Exp}_{\boldsymbol{x}}(\mathcal{P}_{\boldsymbol{x}}^{-1}(\boldsymbol{U}(\mathcal{P}_{\boldsymbol{x}}(\mathcal{P}_{\boldsymbol{x}}^{-1}(\boldsymbol{U}^\top(\mathcal{P}_{\boldsymbol{x}}(\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{z}))))))))). \end{aligned}$$

Now, since the mapping function $\mathcal{P}$ is bijective, we have

$$\begin{aligned} (\Theta_{\boldsymbol{x}}^{\boldsymbol{U}})^{-1}(\Theta_{\boldsymbol{x}}^{\boldsymbol{U}}(\boldsymbol{z})) &= \operatorname{Exp}_{\boldsymbol{x}}(\mathcal{P}_{\boldsymbol{x}}^{-1}(\boldsymbol{U}(\boldsymbol{U}^\top(\mathcal{P}_{\boldsymbol{x}}(\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{z})))))) \\ &= \operatorname{Exp}_{\boldsymbol{x}}(\mathcal{P}_{\boldsymbol{x}}^{-1}(\mathcal{P}_{\boldsymbol{x}}(\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{z})))) \\ &= \operatorname{Exp}_{\boldsymbol{x}}(\operatorname{Log}_{\boldsymbol{x}}(\boldsymbol{z})) \\ &= \boldsymbol{z}. \end{aligned}$$

□

A.4 Analysis

Proposition 8. *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{L}^d$ be any two points on the manifold, and let $s \in \mathbb{R}^+$ be a scaling factor. The Riemannian translation and the Riemannian scaling commute w.r.t. $\boldsymbol{x}$ and $\boldsymbol{y}$:*

$$\Upsilon_{\boldsymbol{y}}^{s} \circ \Gamma_{\boldsymbol{x} \to \boldsymbol{y}} = \Gamma_{\boldsymbol{x} \to \boldsymbol{y}} \circ \Upsilon_{\boldsymbol{x}}^{s}. \tag{28}$$

Proof. Let $\boldsymbol{z} \in \mathbb{L}^d$, and denote $\boldsymbol{z}' = \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z})$. In addition, let $\varphi_{\boldsymbol{x}}$ be the geodesic path from $\boldsymbol{x}$ to $\boldsymbol{z}$ such that $\varphi_{\boldsymbol{x}}(0) = \boldsymbol{x}$ and $\varphi_{\boldsymbol{x}}(1) = \boldsymbol{z}$. We have

$$\begin{aligned}\Upsilon^s_{\boldsymbol{y}} \circ \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) &= \mathrm{Exp}_{\boldsymbol{y}}(s\ \mathrm{Log}_{\boldsymbol{y}}(\boldsymbol{z}')) \\ &= \mathrm{Exp}_{\boldsymbol{y}}(s\ \mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z}))).\end{aligned}$$

Since Lemma 4 implies that the Riemannian scaling (with the same scaling factor) is equivalent to modulation on the tangent spaces, and the parallel transport operator is a homogeneous function of degree 1, we get

$$\begin{aligned}\Upsilon^s_{\boldsymbol{y}} \circ \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{z}) &= \mathrm{Exp}_{\boldsymbol{y}}(\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(s\ \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z}))) \\ &= \mathrm{Exp}_{\boldsymbol{y}}(\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\mathrm{Log}_{\boldsymbol{x}}(\varphi_{\boldsymbol{x}}(s)))) \\ &= \Gamma_{\boldsymbol{x}\to\boldsymbol{y}} \circ \Upsilon^s_{\boldsymbol{x}}(\boldsymbol{z}).\end{aligned}$$

□

Proposition 9. *Let $\boldsymbol{x} \in \mathbb{L}^d$ be a point, $s \in \mathbb{R}^+$ be a scaling factor, and $\boldsymbol{U} \in \mathbb{O}(d)$ be a rotation matrix. The Riemannian scaling $\Upsilon^s_{\boldsymbol{x}}$ and the wrapped rotation $\Theta^{\boldsymbol{U}}_{\boldsymbol{x}}$ commute:*

$$\Upsilon^s_{\boldsymbol{x}} \circ \Theta^{\boldsymbol{U}}_{\boldsymbol{x}} = \Theta^{\boldsymbol{U}}_{\boldsymbol{x}} \circ \Upsilon^s_{\boldsymbol{x}}. \tag{29}$$

Proof. Let $\boldsymbol{z} \in \mathbb{L}^d$. Because the mapping function $\mathcal{P}$ and its inverse are homogeneous functions of degree 1, i.e., $s\mathcal{P}_{\boldsymbol{x}}(\boldsymbol{y}) = \mathcal{P}_{\boldsymbol{x}}(s\boldsymbol{y})$ and $s\mathcal{P}^{-1}_{\boldsymbol{x}}(\boldsymbol{y}) = \mathcal{P}^{-1}_{\boldsymbol{x}}(s\boldsymbol{y})$, we have

$$\begin{aligned}\Upsilon^s_{\boldsymbol{x}}(\Theta^{\boldsymbol{U}}_{\boldsymbol{x}}(\boldsymbol{z})) &= \mathrm{Exp}_{\boldsymbol{x}}(s\ \mathcal{P}^{-1}_{\boldsymbol{x}}(\boldsymbol{U}^\top(\mathcal{P}_{\boldsymbol{x}}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z})))))\\ &= \mathrm{Exp}_{\boldsymbol{x}}(\mathcal{P}^{-1}_{\boldsymbol{x}}(s\ \boldsymbol{U}^\top(\mathcal{P}_{\boldsymbol{x}}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z})))))\\ &= \mathrm{Exp}_{\boldsymbol{x}}(\mathcal{P}^{-1}_{\boldsymbol{x}}(\boldsymbol{U}^\top(s\ \mathcal{P}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z})))))\\ &= \mathrm{Exp}_{\boldsymbol{x}}(\mathcal{P}^{-1}_{\boldsymbol{x}}(\boldsymbol{U}^\top(\mathcal{P}(s\ \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z}))))).\end{aligned}$$

Noting that the scaling factor does not change the matrix consisting of left singular vectors results in

$$\Upsilon^s_{\boldsymbol{x}}(\Theta^{\boldsymbol{U}}_{\boldsymbol{x}}(\boldsymbol{z})) = \Theta^{\boldsymbol{U}}_{\boldsymbol{x}}(\Upsilon^s_{\boldsymbol{x}}(\boldsymbol{z})).$$

□

Proposition 10. *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{L}^d$ be any two hyperbolic vectors on the manifold, $s \in \mathbb{R}^+$ be a scaling factor, and $\boldsymbol{U} \in \mathbb{O}(d)$ be a rotation matrix. Let $\eta : \mathbb{L}^d \to \mathbb{L}^d$ be a map, given by $\eta = \Theta^{\boldsymbol{U}}_{\boldsymbol{y}} \circ \Upsilon^s_{\boldsymbol{y}} \circ \Gamma_{\boldsymbol{x}\to\boldsymbol{y}}$. For any point $\boldsymbol{z} \in \mathbb{L}^d$, if $\widetilde{\boldsymbol{z}} = \eta(\boldsymbol{z})$, then there exists a rotation matrix $\boldsymbol{U}' \in \mathbb{O}(d)$ such that*

$$\boldsymbol{z} = (\Theta^{\boldsymbol{U}'}_{\boldsymbol{x}} \circ \Upsilon^{\frac{1}{s}}_{\boldsymbol{x}} \circ \Gamma_{\boldsymbol{y}\to\boldsymbol{x}})(\widetilde{\boldsymbol{z}}). \tag{30}$$

Proof. Because the parallel transport operator, the mapping function $\mathcal{P}$ and its inverse are homogeneous functions of degree 1, we have

$$\begin{aligned}&(\Theta^{\boldsymbol{U}'}_{\boldsymbol{x}} \circ \Upsilon^{\frac{1}{s}}_{\boldsymbol{x}} \circ \Gamma_{\boldsymbol{y}\to\boldsymbol{x}})(\widetilde{\boldsymbol{z}})\\ &= \mathrm{Exp}_{\boldsymbol{x}}(\mathcal{P}^{-1}_{\boldsymbol{x}}((\boldsymbol{U}')^\top(\mathcal{P}_{\boldsymbol{x}}(\frac{1}{s}\ \mathrm{Log}_{\boldsymbol{x}}(\mathrm{Exp}_{\boldsymbol{x}}(\mathrm{PT}_{\boldsymbol{y}\to\boldsymbol{x}}(\mathcal{P}^{-1}_{\boldsymbol{y}}(\boldsymbol{U}^\top(\mathcal{P}_{\boldsymbol{y}}(\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(s\ \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z}))))))))))))\\ &= \mathrm{Exp}_{\boldsymbol{x}}(\mathcal{P}^{-1}_{\boldsymbol{x}}((\boldsymbol{U}')^\top(\mathcal{P}_{\boldsymbol{x}}(\mathrm{Log}_{\boldsymbol{x}}(\mathrm{Exp}_{\boldsymbol{x}}(\mathrm{PT}_{\boldsymbol{y}\to\boldsymbol{x}}(\mathcal{P}^{-1}_{\boldsymbol{y}}(\boldsymbol{U}^\top(\mathcal{P}_{\boldsymbol{y}}(\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z}))))))))))))\\ &= \mathrm{Exp}_{\boldsymbol{x}}(\mathcal{P}^{-1}_{\boldsymbol{x}}((\boldsymbol{U}')^\top(\mathcal{P}_{\boldsymbol{x}}(\mathrm{PT}_{\boldsymbol{y}\to\boldsymbol{x}}(\mathcal{P}^{-1}_{\boldsymbol{y}}(\boldsymbol{U}^\top(\mathcal{P}_{\boldsymbol{y}}(\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z}))))))))))\\ &= \mathrm{Exp}_{\boldsymbol{x}}(\mathrm{PT}_{\boldsymbol{y}\to\boldsymbol{x}}(\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{z}))))\\ &= \boldsymbol{z}\end{aligned}$$

□

B Illustrations

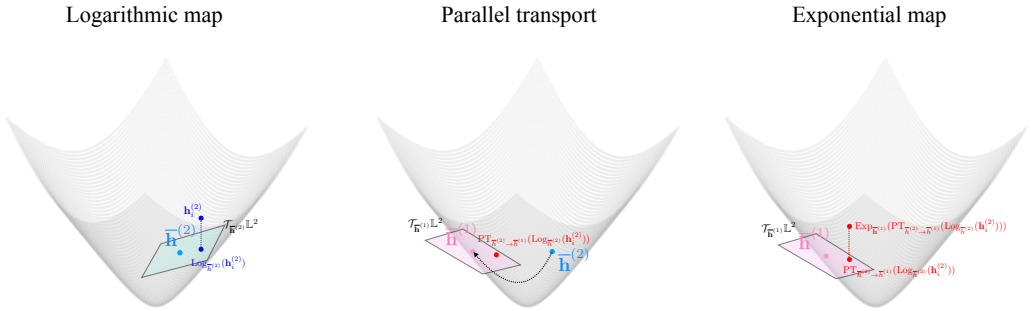

Figure B.1: Illustration of the Riemannian translation $\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{h}_i^{(2)})$. The function $\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{h}_i^{(2)})$ consists of three Riemannian operations in $\mathbb{L}^d$: (*left*) the Logarithmic map applied to $\boldsymbol{h}_i^{(2)}$ at $\overline{\boldsymbol{h}}^{(2)}$, (*middle*) parallel transport applied to $\text{Log}_{\overline{\boldsymbol{h}}^{(2)}}(\boldsymbol{h}_i^{(2)})$ from $\mathcal{T}_{\overline{\boldsymbol{h}}^{(2)}}\mathbb{L}^d$ to $\mathcal{T}_{\overline{\boldsymbol{h}}^{(1)}}\mathbb{L}^d$ along the geodesic path from $\overline{\boldsymbol{h}}^{(2)}$ to $\overline{\boldsymbol{h}}^{(1)}$, and (*right*) the Exponential map projecting the transported point $\text{PT}_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\text{Log}_{\overline{\boldsymbol{h}}^{(2)}}(\boldsymbol{h}_i^{(2)}))$ back to the manifold $\mathbb{L}^d$ at $\overline{\boldsymbol{h}}^{(1)}$.

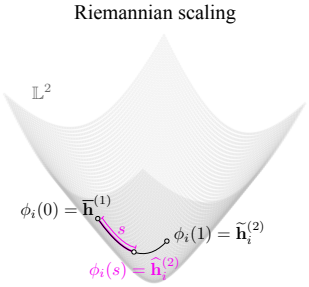

Figure B.2: Illustration of the Riemannian scaling $\Upsilon^s_{\overline{\boldsymbol{h}}^{(1)}}$. The function $\Upsilon^s_{\overline{\boldsymbol{h}}^{(1)}}$ scales the translated point $\widetilde{\boldsymbol{h}}_i^{(2)}$ along the geodesic path $\phi_i(t)$ between $\phi_i(0) = \overline{\boldsymbol{h}}^{(1)}$ and $\phi_i(1) = \widetilde{\boldsymbol{h}}_i^{(2)}$.

C More details on the experimental study

C.1 Datasets

The datasets we consider are described below. They are all publicly available and completely anonymized (without any personally identifiable content).

1. **Breast cancer gene expression data.** Two public breast cancer (BC) gene expression datasets are considered: METABRIC[2] [8] and TCGA[3] [26]. Prior to pre-processing, the data entries (samples and/or features) with `nan` values are removed since dealing with missing data is not in the scope of this work. Then, following common practice, the genes with the largest variance in both datasets are selected. Here, we present the results based on 200 genes. We report that repeating the experiments with different number of genes does not significantly affect the results. Note that, in contrast to [37], the choice of genes is done in an unsupervised manner. Finally, we apply a standard pre-processing using z-score normalization to the data samples.

[2] `https://www.cbioportal.org/study/summary?id=brca_metabric`
[3] `https://www.cbioportal.org/study/summary?id=brca_tcga_pub`

Algorithm 2 Hyperbolic OT-DA with weighted Fréchet mean (HOT-F)

Input: Source set $\mathcal{H}^{(s)} = \{\boldsymbol{h}_i^{(s)}\}_{i=1}^{N_s}$ and target set $\mathcal{H}^{(t)} = \{\boldsymbol{h}_i^{(t)}\}_{i=1}^{N_t}$

Output: Transported set $\underline{\mathcal{H}} = \{\underline{\boldsymbol{h}}_i\}_{i=1}^{N_s}$

1: compute $\boldsymbol{p}_s = \frac{1}{N_S}\mathbf{1}_{N_S}$ and $\boldsymbol{p}_t = \frac{1}{N_t}\mathbf{1}_{N_t}$

2: compute the optimal transport cost matrix $\boldsymbol{M}$, where $\boldsymbol{M}(i,j) = d^2_{\mathbb{L}^d}(\boldsymbol{h}_i^{(s)}, \boldsymbol{h}_j^{(t)})$

3: compute the transport plan $\boldsymbol{R}$ using Sinkhorn optimal transport [9] by

$$\min_{\boldsymbol{R}\in\Omega} \langle \boldsymbol{R}, \boldsymbol{M}\rangle - \frac{1}{\vartheta} g(\boldsymbol{R}), \tag{31}$$

where $\Omega = \{\boldsymbol{R} \in \mathbb{R}^{N_s \times N_t} | \boldsymbol{R}\mathbf{1} = \boldsymbol{c}_s \text{ and } \boldsymbol{R}^\top \mathbf{1} = \boldsymbol{c}_t\}$, $\boldsymbol{c}_s$ and $\boldsymbol{c}_t$ are the discrete densities, and $g(\boldsymbol{R}) = \sum_{i,j} \boldsymbol{R}(i,j)\log(\boldsymbol{R}(i,j))$ is the entropic regularization

4: **for** all $i \in \{1, 2, 3, \ldots, N_s\}$ **do**

5: compute the weighted Fréchet mean

$$\underline{\boldsymbol{h}}_i = \arg\min_{\boldsymbol{x}\in\mathbb{L}^d} \sum_{j=1}^{N_t} \boldsymbol{R}(i,j) d^2_{\mathbb{L}^d}(\boldsymbol{x}, \boldsymbol{h}_j^{(t)}) \tag{32}$$

6: **end for**

2. **Lung cancer gene expression data.** Three lung adenocarcinoma (cancer) datasets[4] [21] are unitized, consisting of 2553 genes that were reported as reliable features for discovering three lung adenocarcinoma subtypes. Similarly to the BC datasets, a standard pre-processing of z-score normalization is applied.
3. **CyTOF data.** We consider eight batches of measurements: two patients, two conditions (before and after treatment), and two different days. Each batch consists of $1800 - 5000$ cells. We use already denoised data available in this link [5] provided by the authors of [48].

C.2 Competing methods

1. **HOT-F.** We follow the algorithm proposed in [54] for the manifold of symmetric and positive-definite matrices and adapt it to Lorentz model $\mathbb{L}^d$ as follows. We solve the optimal transport (OT) problem with ground distances that equal the geodesic distances based on the Riemannian geometry of $\mathbb{L}^d$. In addition, we use the weighted Fréchet mean in the hyperbolic space to translate the resulting transport map to a discrete point-to-point map. We present the entire HOT-F algorithm in Algorithm 2.
2. **HOT-L & HOT-ME.** HOT-L and HOT-ME stand for two variants of hyperbolic optimal transport (with W-linear map and with mapping estimation), which were presented in [22]. Both algorithms solve an OT problem in the Poincaré ball [41]. Therefore here, we first transform the hyperbolic vectors $\{\boldsymbol{x}_i \in \mathbb{L}^d\}$ in the Lorentz model into the Poincaré model [42] using the function $\mathcal{K} : \mathbb{L}^d \to \mathbb{B}^d$ given by

$$\mathcal{K}(\boldsymbol{x}_i) = \frac{[\boldsymbol{x}_i(2), \boldsymbol{x}_i(3), \ldots, \boldsymbol{x}_i(d+1)]^\top}{1 + \boldsymbol{x}_i(1)},$$

where $\mathbb{B}^d = \{\boldsymbol{q} \in \mathbb{R}^d \big| \; ||\boldsymbol{q}|| < 1\}$. Then, HOT-L and HOT-ME are applied to the transformed Poincaré samples.

C.3 Implementation details

Our code generating the experimental results as well as the simulated ones is included in the supplemental material. All the experiments were performed on NVIDIA RTX 1080 Ti GPU. A fixed random seed (9512) was used in all the experiments in Section 4 and Appendix D.

[4] https://ascopubs.org/doi/suppl/10.1200/JCO.2005.05.1748

[5] https://github.com/ushaham/BatchEffectRemoval

Hyperbolic Lorentz representation. For the hyperbolic embedding of the bioinformatics datasets, we used the code in this link[6] from [42], where the manifold is set to be `lorentz` with learning rate 10^{-3}, epoch 1000, train threads 2, and batch size 20. The computation of the Fréchet mean is implemented according to the efficient algorithm proposed in [33].

More details on HPA. As stated in Section 3, the Riemannian translation, the Riemannian scaling, and the wrapped rotation are functions whose domain and range are in $\mathbb{L}^d$, and all three functions comprise the Logarithmic map and the Exponential map, projecting the sample from the manifold to the tangent space and back, respectively. When applying the three operations (translation, scaling, and rotation) in cascade, we obtain two subsequent Exponential map and Logarithmic map, which cancel out. Therefore, based on the properties in Section 3 and Appendix A, we can consider the function $\theta_k : \mathbb{L}^d \to \mathbb{L}^d$, given by

$$\theta_k(\boldsymbol{h}_i^{(k)}) = \mathrm{Exp}_{\overline{\boldsymbol{h}}}(\mathcal{P}_{\overline{\boldsymbol{h}}}^{-1}(\boldsymbol{U}^\top(\mathcal{P}_{\overline{\boldsymbol{h}}}(\mathrm{PT}_{\overline{\boldsymbol{h}}^{(k)}\to\overline{\boldsymbol{h}}}\sqrt{1/d^{(k)}}(\mathrm{Log}_{\overline{\boldsymbol{h}}^{(2)}}(\boldsymbol{h}_i^{(k)}))) - \boldsymbol{p}^{(k)}) + \boldsymbol{p}^{(k)})),$$

where $\boldsymbol{U} = \boldsymbol{U}^{(1)}(\boldsymbol{U}^{(k)})^\top$, which consists of only one Logarithmic mapping and one Exponential mapping.

More details on the competing methods. POT library [12] is used in the implementation of HOT-F. In the experiments, the weight ϑ of the entropic regularization in Eq. (31) is set to 1. For the implementation of HOT-L and HOT-ME, we use the code in this link[7] provided by the author of [22] (learning the OT map in the Poincaré model). We run the algorithms using the default parameters except the number of hidden unit used in HOT-ME, which is set to 500.

Remarks on the extension to multiple sets.

- In the *Riemannian translation*, the reason we propose to align all the datasets to the middle point $\overline{\boldsymbol{h}}$ (the center of mass of the Riemannian means of the sets) is twofold. First, it circumvents the choice of a reference set. Our empirical study show that the choice of the particular reference set can dramatically affect the alignment results, for example, in cases where the chosen reference set is positioned near the boundary of the manifold. Second, PT accumulates distortion along the transport path. By parallel transporting all the sets to a middle point, we guarantee that the total transport distance is minimized.
- In the *Riemannian wrapped rotation*, the choice of the reference set may be improved by choosing the set with the smallest sum of rotations w.r.t. the other sets. In addition, the point to which we align all the means need to be far from the boundary to minimize distortions caused by the hyperbolic metric. Indeed, $\overline{\boldsymbol{h}}$ lies in the convex hull of the means $\{\overline{\boldsymbol{h}}^{(k)}\}_{k=1}^K$, and therefore, is naturally away from the boundary.

More details on the evaluation metrics. We use the following metrics to assess the alignment quality in the hyperbolic space.

1. **Discrepancy.** Fully described in the paper in Section 4.
2. **k-NN.** To test the adequacy of the representation of the data in hyperbolic space, we test the k-NN classification obtained based on each batch separately. We term this classification as S-Baseline, and it is performed using a leave-one-sample-out cross-validation. The evaluation of the alignment quality using k-NN classification is performed with a leave-one-batch-out cross-validation.
3. **MMD.** Given two sets $\mathcal{Z}^{(1)} = \{\boldsymbol{z}_i^{(1)} \in \mathbb{L}^d\}_{i=1}^n$ and $\mathcal{Z}^{(2)} = \{\boldsymbol{z}_i^{(2)} \in \mathbb{L}^d\}_{i=1}^m$, the MMD in hyperbolic space is computed based on the geodesic distance as follows:

$$\begin{aligned}&\mathrm{MMD}^2(\mathcal{Z}^{(1)}, \mathcal{Z}^{(2)})\\ &= \frac{1}{n^2}\sum_{\boldsymbol{z}_i,\boldsymbol{z}_j\in\mathcal{Z}^{(1)}} k(\boldsymbol{z}_i,\boldsymbol{z}_j) + \frac{1}{m^2}\sum_{\boldsymbol{z}_i,\boldsymbol{z}_j\in\mathcal{Z}^{(2)}} k(\boldsymbol{z}_i,\boldsymbol{z}_j) - \frac{2}{nm}\sum_{\boldsymbol{z}_i\in\mathcal{Z}^{(1)},\boldsymbol{z}_j\in\mathcal{Z}^{(2)}} k(\boldsymbol{z}_i,\boldsymbol{z}_j)\end{aligned}$$

[6] https://github.com/facebookresearch/poincare-embeddings
[7] https://github.com/ahoyosid/hyperbolic_alignment

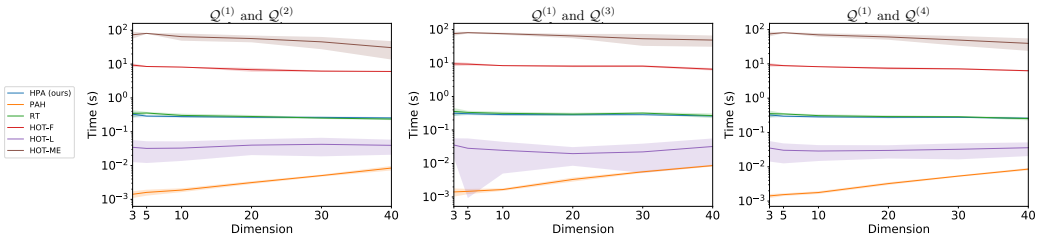

Figure D.1: Runtime of aligning the synthetic data from Sec. 4.1.

Table 3: The k-NN AUC-ROC for ablation studies.

Dataset	Baseline	HPA	RT	RS	RWR	RT+RS	RT+RWR	RS+RWR
BC	0.5254 ± 0.0091	$\mathbf{0.7410 \pm 0.0354}$	0.6001 ± 0.0658	0.5013 ± 0.0032	0.5184 ± 0.0027	0.6143 ± 0.0633	0.7111 ± 0.0219	0.5338 ± 0.0132
LC	0.5521 ± 0.0991	$\mathbf{0.8316 \pm 0.0904}$	0.5318 ± 0.0370	0.5435 ± 0.0138	0.4578 ± 0.0382	0.5123 ± 0.0042	0.7729 ± 0.0913	0.5398 ± 0.0821
P1 BT	0.6646 ± 0.1556	$\mathbf{0.9401 \pm 0.0068}$	0.9020 ± 0.0086	0.7028 ± 0.0993	0.6897 ± 0.0049	0.8842 ± 0.0053	0.9127 ± 0.0032	0.5498 ± 0.0931
P1 AT	0.7656 ± 0.1564	$\mathbf{0.9329 \pm 0.0011}$	0.8270 ± 0.0880	0.7358 ± 0.0873	0.7732 ± 0.1392	0.8347 ± 0.0763	0.8923 ± 0.0286	0.7322 ± 0.0121
P2 BT	0.6971 ± 0.1335	$\mathbf{0.9329 \pm 0.0186}$	0.8830 ± 0.0142	0.6210 ± 0.0152	0.0723 ± 0.0313	0.8741 ± 0.0194	0.9228 ± 0.0029	0.7013 ± 0.0182
P2 AT	0.5688 ± 0.0688	$\mathbf{0.8453 \pm 0.0798}$	0.7190 ± 0.0439	0.5248 ± 0.0086	0.5390 ± 0.0402	0.7851 ± 0.0227	0.8223 ± 0.0138	0.5739 ± 0.0812

where the kernel k is defined by

$$k(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp\left(- \frac{d^2_{\mathbb{L}^d}(\boldsymbol{z}_i, \boldsymbol{z}_j)}{\epsilon} \right).$$

Following common practice, we set $\epsilon = 10 \times \mu$, where μ is the median of the pairwise distances. We note that such a choice yields a positive definite kernel k.

D Additional experimental results

D.1 Runtime analysis of simulations

To complement the accuracy results presented in Fig. 2, we present in Fig. D.1 the runtime of HPA and the other methods. We remark that PAH is the fastest, but its performance is comparable to the performance of using RT alone, and it is inferior to the results of HPA. In addition, HOT-L is faster than HPA, because there exists a closed-form solution for the Gaussian transport [22]. Nonetheless, as presented in Section 3, all the tested OT-based methods are outperformed by HPA.

D.2 k-NN AUC-ROC for different values of k in batch correction tasks

The k-NN results reported in Table 1 were obtained by using a different value of k for each method (the value that yields the best result was chosen). To complete the picture, Fig. D.2 presents the k-NN performance of all the methods as a function of k. Same as before, the classification is performed using a leave-one-batch-out cross-validation. We remark that the performance of the P2 AT alignment task drops dramatically with k because the number of PMA/ionomycin stimulated PBMCs and the number of non-stimulated PBMCs are imbalanced.

D.3 Ablation study

In Table 3 and Table 4, we present ablation study results, where we compare the performance of HPA to the performance obtained by each of the three components (Riemannian translation (RT), Riemannian scaling (RS), and Riemannian wrapped rotation (RWR)), and their combinations. First, we see that each component has a contribution and the combination of all three components yields the best classification results. We also see that the Riemannian translation arguably plays the most important role among the three components. In addition, we see that the RT is critical, since applying the scaling and/or the rotation without the RT results in poor performance.

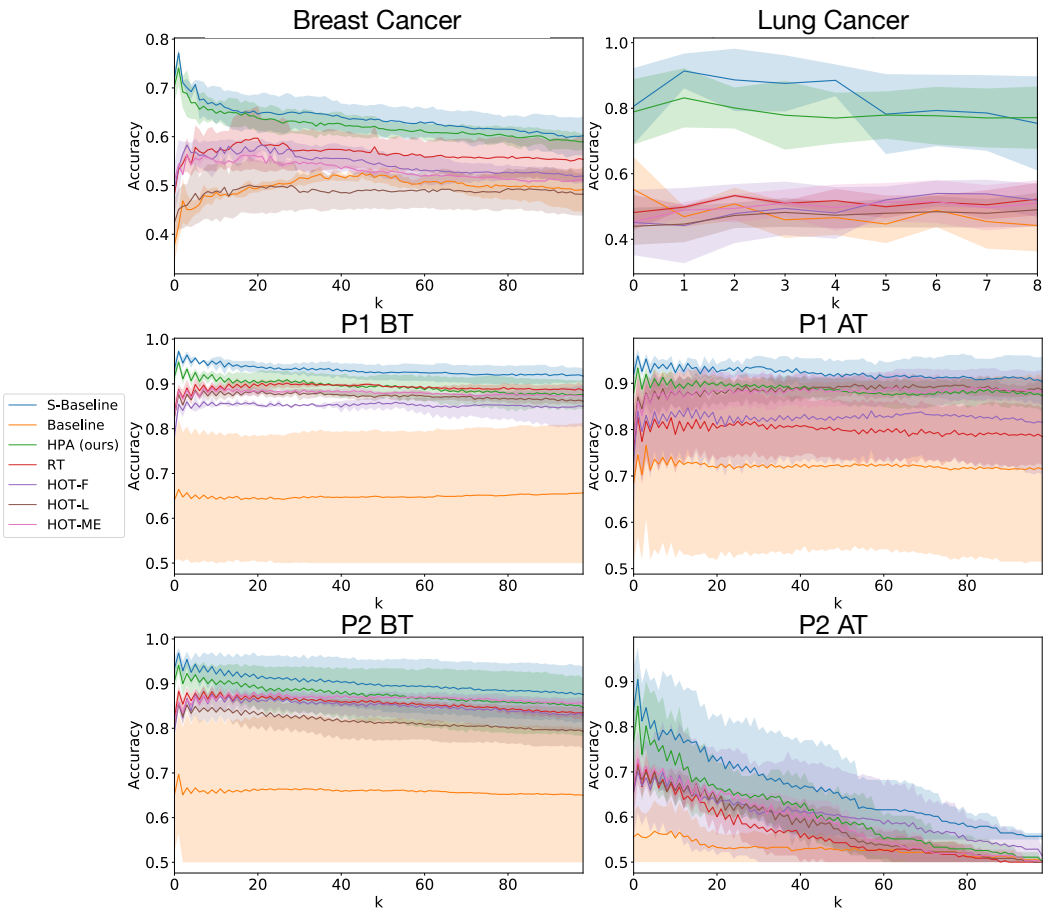

Figure D.2: k-NN AUC-ROC for various k values.

Table 4: The MMD values for ablation studies.

Dataset	Baseline	HPA	RT	RS	RWR	RT+RS	RT+RWR	RS+RWR
BC	0.2089 ± 0.0027	0.0013 ± 0.0004	0.0072 ± 0.0011	0.1930 ± 0.0018	0.2231 ± 0.0098	0.0066 ± 0.0008	0.0022 ± 0.0008	0.1820 ± 0.0031
ST&UM	0.1072 ± 0.0051	0.0162 ± 0.0048	0.0250 ± 0.0049	0.1003 ± 0.0043	0.1189 ± 0.0076	0.0231 ± 0.0023	0.0018 ± 0.0019	0.1311 ± 0.0097
ST&D-F	0.3213 ± 0.0152	0.0122 ± 0.0042	0.0150 ± 0.0106	0.3358 ± 0.0231	0.3281 ± 0.0185	0.0142 ± 0.0087	0.0124 ± 0.0030	0.3092 ± 0.0073
UM&D-M	0.0790 ± 0.0071	0.0168 ± 0.0090	0.0168 ± 0.0032	0.0813 ± 0.0063	0.0890 ± 0.0082	0.0172 ± 0.0083	0.0165 ± 0.0059	0.0831 ± 0.0011
P1 BT	0.0638 ± 0.0024	0.0012 ± 0.0002	0.0020 ± 0.0002	0.0688 ± 0.0049	0.0898 ± 0.0039	0.0018 ± 0.0007	0.0013 ± 0.0003	0.0644 ± 0.0031
P1 AT	0.0598 ± 0.0014	0.0006 ± 0.0001	0.0015 ± 0.0001	0.0605 ± 0.0051	0.0923 ± 0.0030	0.0014 ± 0.0001	0.0008 ± 0.0002	0.0543 ± 0.0022
P2 BT	0.0424 ± 0.0021	0.0012 ± 0.0001	0.0015 ± 0.0001	0.0404 ± 0.0066	0.0591 ± 0.0012	0.0014 ± 0.0004	0.0013 ± 0.0001	0.0561 ± 0.0096
P2 AT	0.0758 ± 0.0053	0.0011 ± 0.0002	0.0013 ± 0.0002	0.0733 ± 0.0041	0.0744 ± 0.0079	0.0013 ± 0.0003	0.0011 ± 0.002	0.0731 ± 0.0055

D.4 Alternative implementation of rotation

The tangent space is define by $\mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d := \{\boldsymbol{v} \in \mathbb{R}^{d+1} | \langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathcal{L}} = 0\}$, where $\boldsymbol{x} \in \mathbb{L}^d$, $\langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathcal{L}} = \boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{v}$ is the Lorentzian inner product, and $\boldsymbol{H} \in \mathbb{R}^{(d+1)\times(d+1)}$ is the hyperbolic metric tensor, defined by $\boldsymbol{H} = [-1, \boldsymbol{0}^\top; \boldsymbol{0}, \boldsymbol{I}_d]$.

This tangent space $\mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d \subset \mathbb{R}^{d+1}$ is isometric to the Euclidean vector space $\mathbb{R}^d$, but it is not a Euclidean vector space with the standard inner product due to the Lorentzian orthogonality constraint. Therefore, implementing the hyperbolic rotation as in [51] or by applying SVD to the tangent space directly (without applying first the isometry) is not appropriate, because the resulting rotated points might violate the orthogonality constraint, and as a result, might not be in $\mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$. That is, if $\boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{v} = 0$, then $\boldsymbol{x}^\top \boldsymbol{H} (\boldsymbol{R}_h \boldsymbol{v})$ and $\boldsymbol{x}^\top \boldsymbol{H} (\boldsymbol{V} \boldsymbol{v})$ might not equal 0, where $\boldsymbol{R}_h \in \mathbb{R}^{(d+1)\times(d+1)}$ is the rotation map in [51] and $\boldsymbol{V} \in \mathbb{O}(d+1)$ is the standard rotation matrix in $\mathbb{R}^{d+1}$. Both rotations also do not preserve the Riemannian mean $\overline{\boldsymbol{h}}$, in case the mean does not coincide with the origin in $\mathbb{L}^d$.

Table 5: The k-NN AUC-ROC for different rotation strategies.

Dataset	Baseline	RT+RS	HPA	Iso($\mathcal{K}$)
BC	0.5254 ± 0.0091	0.6143 ± 0.0633	$\mathbf{0.7410 \pm 0.0354}$	0.7033 ± 0.0210
LC	0.5521 ± 0.0991	0.5123 ± 0.0042	$\mathbf{0.8316 \pm 0.0904}$	0.7558 ± 0.0833
P1 BT	0.6646 ± 0.1556	0.8842 ± 0.0053	0.9401 ± 0.0068	$\mathbf{0.9533 \pm 0.0041}$
P1 AT	0.7656 ± 0.1564	0.8347 ± 0.0763	0.9329 ± 0.0011	$\mathbf{0.9333 \pm 0.0012}$
P2 BT	0.6971 ± 0.1335	0.8741 ± 0.0194	$\mathbf{0.9329 \pm 0.0186}$	0.8914 ± 0.0103
P2 AT	0.5688 ± 0.0688	0.7851 ± 0.0227	$\mathbf{0.8453 \pm 0.0798}$	0.8103 ± 0.0141

One possibility to implement the Riemannian rotation is to find the isometric mapping from the tangent space to the Euclidean vector space. Concretely, there exists an invertible map $\mathcal{K} : \mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d \to \mathbb{R}^d$ such that for any two points $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$, we get an isometry: $\langle \boldsymbol{z}_1, \boldsymbol{z}_2 \rangle_{\mathcal{L}} = \langle \mathcal{K}(\boldsymbol{z}_1), \mathcal{K}(\boldsymbol{z}_2) \rangle$, where $\langle \cdot, \cdot \rangle$ is the standard dot product. We can construct such an isometry $\mathcal{K}$ by finding an orthonormal basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \ldots, \boldsymbol{v}_d\}$ (orthogonal with respect to the Lorentzian inner product) for any $\boldsymbol{v} \in \mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$, where $\boldsymbol{v}_1 = \frac{\boldsymbol{v}}{||\boldsymbol{v}||_{\mathcal{L}}}$. Then, for any point $\boldsymbol{z} \in \mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$, the isometry $\mathcal{K}$ is defined by

$$\mathcal{K}(\boldsymbol{z}) := \begin{bmatrix} \langle \boldsymbol{z}, \boldsymbol{v}_1 \rangle_{\mathcal{L}} \\ \langle \boldsymbol{z}, \boldsymbol{v}_2 \rangle_{\mathcal{L}} \\ \cdots \\ \langle \boldsymbol{z}, \boldsymbol{v}_d \rangle_{\mathcal{L}} \end{bmatrix} \in \mathbb{R}^d,$$

and its inverse map is given by

$$\mathcal{K}^{-1}(\boldsymbol{q}) := \sum_{i=1}^{d} \boldsymbol{q}(i) \boldsymbol{v}_i,$$

where $\boldsymbol{q} \in \mathbb{R}^d$. Proving that such a map $\mathcal{K}$ is an isometry is straight-forward: for any two points $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$, consider their expansions $\boldsymbol{z}_1 = \sum_{i=1}^{d} \alpha_i \boldsymbol{v}_i$ and $\boldsymbol{z}_2 = \sum_{i=1}^{d} \beta_i \boldsymbol{v}_i$, where $\alpha_i = \langle \boldsymbol{z}_1, \boldsymbol{v}_i \rangle_{\mathcal{L}}$ and $\beta_i = \langle \boldsymbol{z}_2, \boldsymbol{v}_i \rangle_{\mathcal{L}}$. Then, $\langle \boldsymbol{z}_1, \boldsymbol{z}_2 \rangle_{\mathcal{L}} = \sum_{i=1}^{d} \alpha_i \beta_i = \langle \mathcal{K}(\boldsymbol{z}_1), \mathcal{K}(\boldsymbol{z}_2) \rangle$.

However, the implementation of the rotation described above is involved, because the explicit construction of the isometric map $\mathcal{K}$ requires finding an orthonormal basis w.r.t. the Lorentzian inner product. Specifically, our implementation was based on the Gram–Schmidt process, which is numerically unstable.

As a remedy, we propose an alternative simpler implementation of the rotation based on the wrapped operations. We use the so-called mapping $\mathcal{P}$ that is a map between the tangent space and the Euclidean vector space $\mathbb{R}^d$. Then, we apply SVD in the Euclidean space, and use the inverse projection map $\mathcal{P}^{-1}$ to transfer back to the tangent space. Please note that no information is lost when we apply $\mathcal{P}$ (and $\mathcal{P}^{-1}$) due to the Lorentzian orthogonality.

We empirically compared the results of the two implementations of the rotation component: constructing and using the isometry, and using the wrapped operations, combining both with the same Riemannian translation and Riemannian scaling. The results on the bioinformatics datasets are presented in Table 5 and Table 6, where the rotation using isometric mapping is denoted by Iso($\mathcal{K}$). We see that the obtained results are comparable, with a slight advantage to the Riemannian wrapped rotation implementation (specifically, when the data have multiple labels and multiple batches). Since the Riemannian wrapped rotation (RWR) is simpler and computationally more efficient and stable, we choose RWR as the preferable implementation of rotation.

Table 6: The MMD value for rotation strategies.

Dataset	Baseline	RT+RS	HPA	Iso($\mathcal{K}$)
BC	0.2089 ± 0.0027	0.0066 ± 0.0008	0.0013 ± 0.0004	0.0019 ± 0.0004
ST&UM	0.1072 ± 0.0051	0.0231 ± 0.0023	0.0162 ± 0.0048	0.0173 ± 0.0032
ST&D-F	0.3213 ± 0.0152	0.0142 ± 0.0087	0.0122 ± 0.0042	0.0111 ± 0.0011
UM&D-M	0.0790 ± 0.0071	0.0172 ± 0.0083	0.0168 ± 0.0090	0.0183 ± 0.0043
P1 BT	0.0638 ± 0.0024	0.0018 ± 0.0007	0.0012 ± 0.0002	0.0012 ± 0.0003
P1 AT	0.0598 ± 0.0014	0.0014 ± 0.0001	0.0006 ± 0.0001	0.0008 ± 0.0001
P2 BT	0.0424 ± 0.0021	0.0014 ± 0.0004	0.0012 ± 0.0001	0.0014 ± 0.0002
P2 AT	0.0758 ± 0.0053	0.0013 ± 0.0003	0.0011 ± 0.0002	0.0012 ± 0.0002

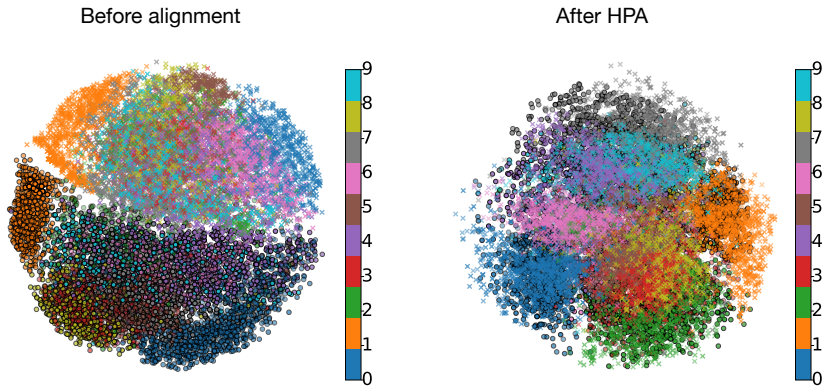

Figure D.3: The visualizations of HPA applied to the digit datasets, where $\times$ and $\bigcirc$ represent MNIST and USPS, respectively. The colors serve as the digit labels.

D.5 Out-of-sample extension

Table 7: Out-of-sample-extension (OOSE) in the CyTOF alignment task.

Seen set	Unseen set	HPA	OOSE performance
P1 BT	P2 BT	0.9329 ± 0.0186	0.8416 ± 0.0421
P2 BT	P1 BT	0.9401 ± 0.0068	0.8327 ± 0.0219
P1 AT	P2 AT	0.8453 ± 0.0798	0.7102 ± 0.0814
P2 AT	P1 AT	0.9329 ± 0.0011	0.6958 ± 0.0922

Here, we demonstrate the out-of-sample extension capabilities of HPA as follows. Recall that the CyTOF alignment task consists of $8 = 2 \times 2 \times 2$ batches of data collected from two patients under two different conditions (BT/AT) and at two different days (batches). First, using HPA, we build the batch correction map $\zeta : \mathbb{L}^d \to \mathbb{L}^d$, consisting of the Riemannian mean, scaling, and rotation, for removing the batch effects between the two days of one patient. Then, we apply the same map ζ, as is, for removing the batch effects between two days of the unseen data from the other patient.

Table 7 shows the obtained k-NN classification using the best choice of k (denoted OOSE performance). As a baseline, we also include the results from Table 1 without out-of-sample-extension (denoted HPA). We remark that the relatively poor results obtained for the AT condition (compared to the results obtained for the BT condition) could be due to the imbalance presence of the stimulated PBMCs in P2 AT.

D.6 Aligning digit datasets

Table 8: The k-NN AUC-ROC and MMD in the digits alignment task.

	S-Baseline	Baseline	HPA	RT	HOT-F	HOT-L	HOT-ME
k-NN	0.9705 ± 0.0076	0.3143 ± 0.0704	0.8655 ± 0.0020	0.4086 ± 0.0555	0.3716 ± 0.1277	0.3234 ± 0.0201	0.1143 ± 0.0167
MMD	-	0.4785 ± 0.0012	0.0007 ± 0.0001	0.0053 ± 0.0007	0.0012 ± 0.0004	0.0008 ± 0.0001	0.0013 ± 0.0003

One of the standard benchmarks for datasets alignment is the alignment of digit datasets, specifically, MNIST [31] and USPS [23]. Here, we wish to use this task in order to demonstrate the effectiveness of our HPA even in cases where the hierarchical structure is not inherent in the data (namely, the images of the digits in this case). We remark that we do not aim to achieve or beat the state-of-the-art performance in this task, because the representation in hyperbolic space might not be the most suitable for these particular data.

The resolution of the images in MNIST is 28×28, whereas the resolution in USPS is 16×16. Therefore, prior to applying HPA, we first resize each image in MNIST to the smaller size of the images in USPS. Then, the images are reshaped into column-stacks in $\mathbb{R}^{256}$, and in turn, are embedded in $\mathbb{L}^d$. Once we obtain the hyperbolic representations of the images, we apply HPA and the other competing alignment methods in the purely unsupervised (label-free) manner. Because the runtime of the competing methods (e.g., HOT-F) is increasing fast with the number of samples (images), we consider only a subset of 7000 images.

Fig. D.3 depicts a visualization of the embedding of the two digit datasets before and after the alignment using HPA. Same as before, for visualization purposes, we project the data from $\mathbb{L}^3$ to the 3D Poincaré ball. We see that before the alignment, the data are primarily divided according to dataset. In contrast, after HPA alignment, the data is organized according to digits. Table 8 presents the k-NN digit classification and the MMD. The MMD values are obtained by selecting ten random subsets of size 3500 from each of the datasets, which is the half size of the minimal dataset, same as in the other experiments. Same as in Section 4, the dimension of the hyperbolic space d is set to the dimension yielding the best digit classification applied to each dataset separately (without alignment). We observe that our HPA obtains an accurate unsupervised (label-free) alignment despite the probable use of sub-optimal representation.

E Classical PA and comparing HPA with other PA schemes

E.1 Background on Procrustes analysis

Classical Procrustes analysis consists of a sequence of three geometric transformations: translation, scaling and rotation, and it is usually applied in a Euclidean space [17]. In the context of data alignment, consider two (or more) sets of points, say $\mathcal{W} = \{\boldsymbol{w}_i \in \mathbb{R}^d\}_{i=1}^n$ and $\mathcal{W}' = \{\boldsymbol{w}'_i \in \mathbb{R}^d\}_{i=1}^{n'}$, where the goal is to find a global map $f : \mathbb{R}^d \to \mathbb{R}^d$ transforming the points from $\mathcal{W}'$ to $\mathcal{W}$. The application of Procrustes analysis to this problem translates to finding a map of the form $f(\boldsymbol{w}') = \frac{1}{b}\boldsymbol{V}^\top(\boldsymbol{w}'_i - \boldsymbol{w}'_m) + \boldsymbol{w}_m$, where $b \in \mathbb{R}$ is the scaling factor, $\boldsymbol{w}_m = \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i$ and $\boldsymbol{w}'_m = \frac{1}{n'}\sum_{i=1}^{n'} \boldsymbol{w}'_i$ are the mean vectors of the sets $\mathcal{W}$ and $\mathcal{W}'$, respectively, and $\boldsymbol{V} \in \mathbb{O}(d)$ is an orthogonal matrix [18]. If there is a known correspondence between the points, then the scaling b and the rotation matrix $\boldsymbol{V}$ could be the solution of an optimization problem minimizing some loss, e.g., the ℓ_2 distance between $\boldsymbol{w}_i$ and $f(\boldsymbol{w}')$. Otherwise, common practice is to set b and $\boldsymbol{V}$ so that the first moments of the sets are aligned, using, for example, singular value decomposition (SVD).

E.2 Comparison with PAH [51]

In Section 4.1, we present an empirical comparison to PAH. The comparison is done only in simulations, because PAH requires sample correspondence, and the tested datasets do not have sample correspondence. Both PAH [51] and our HPA are conceptually based on Procrustes analysis. However the problem setting and the approach are different. Specifically, our HPA makes use of the Riemannian geometry of hyperbolic spaces. We outline below a comparison on a technical level and emphasize the advantage of our method.

1. **Translation:** PAH proposed to translate the centroids of the two sets to the origin in $\mathbb{L}^d$, which is denoted by $\boldsymbol{\mu}_0$. In contrast, we proposed a Riemannian translation of the sets using parallel transport along the unique geodesic path connecting the Fréchet mean of one set to the Fréchet mean of the other set (or from the Fréchet mean of the set to the Fréchet mean of the Fréchet means when multiple sets are aligned). The translations of PAH and HPA are both isometries. However, the translation of our HPA also preserves the geodesic velocities. We also remark that translations cause rotations and distortions. Our translation using PT (derived from the Levi-Civita connection which is torsion-free) introduces the "minimal distortion" in the Riemannian sense. In addition, the result in Prop. 5 (which is an important property for alignment, allowing for a convenient multi-level alignments) is a direct consequence of the specific translation we propose and does not necessarily hold when the sets are translated to the origin as in PAH.
2. **Scaling:** PAH assumes that the hyperbolic sets are isometric and it does **not** address the scaling problem at all. In our HPA, the proposed Riemannian scaling is based on the geodesic path between each point and the Riemannian mean of the set. This way we can align the Riemannian second moment (defined as the dispersion) by taking into account the change of the Riemannian metric as we travel along the geodesic path on the manifold.
3. **Rotation:** PAH suggested a hyperbolic rotation map that requires one-to-one correspondence. We do not make this assumption, and our Riemannian wrapped rotation component (and the other components) can operate in a broader context. Such an assumption significantly limits the scope of PAH. Indeed, the batch correction applications we considered (nor the standard MNIST and USPS alignment we presented in the appendix) do not have one-to-one correspondence, and therefore, PAH cannot be tested on these applications. In addition, we note that the hyperbolic rotation map in PAH does not preserve the Riemannian mean if the mean does not coincide with the origin. In PAH, it does not raise a problem since the mean alignment is implemented by translating the sets to the origin. However, in general, and specifically when using the Riemannian translation of our HPA, such a rotation could cancel the mean alignment.

E.3 Euclidean Procrustes analysis on the tangent space

An alternative implementation of the alignment could be to first project the data on the (linear) tangent space, and then, apply standard Euclidean Procrustes analysis directly in the tangent space, that is, mean vector subtraction, vector scaling, and vector rotation using SVD in a Euclidean tangent vector space. Below, we compare this procedure to our HPA.

1. **Translation:** Unlike the Riemannian translation we propose, this tangent Euclidean translation (i) does not align the Riemannian means of the sets, (ii) does not preserve geodesic distances (i.e., it is not isometric w.r.t. the Riemannian distance), (iii) does not preserve the local geometry (in the sense we define in the paper by preserving geodesic velocities), and (iv) is not derived from the Levi-Civita connection, and therefore, it does not admit Prop. 5, which is an important property for alignment, allowing for a convenient multi-level alignment.
2. **Scaling:** The proposed Riemannian scaling on the manifold is equivalent to the modulation on the tangent space with the same scaling factor (See Lemma 4 in Appendix A).
3. **Rotation:** Applying SVD to the tangent space directly is not appropriate, because the resulting rotated points might violate the Lorentzian orthogonality constraint, and therefore, might not be in the same tangent space.

Another possible way to implement the alignment is to first project the points on the tangent space, then, to isometrically map them to a Euclidean space (the existence of such an isometric map between the tangent space $\mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d \subset \mathbb{R}^{d+1}$ and a Euclidean vector space $\mathbb{R}^d$ is guaranteed by definition), and finally, to apply Euclidean Procrustes analysis in the isometric Euclidean space. On the one hand, the use of the isometric mapping allows us to apply the Euclidean Procrustes analysis without conforming to the Lorentzian orthogonality constraint. On the other hand, as we detail below, it introduces its own challenges.

1. **Translation:** Each tangent space has its own Lorentzian orthonormal basis, and the Lorentzian orthonormal bases of different tangent spaces do not necessarily span the same

space. Therefore, similarly to the translation in the tangent space, this translation in the Euclidean isometric space does not admit items (i)-(iv) above.

2. **Scaling:** This scaling is equivalent to the proposed Riemannian scaling w.r.t. the Riemannian mean.
3. **Rotation:** Unlike applying SVD (for the purpose of rotation) directly in the tangent space, this rotation is a valid alternative. However, compared to the proposed Riemannian wrapped rotation, this alternative is computationally less efficient and stable, and it obtains slightly worse empirical results. See Appendix D.4 for more details and an empirical comparison.