Neural approximation of Procrustes-Wasserstein transport maps

Anonymous authors
Paper under double-blind review

Abstract

Optimal Transport (OT) provides a rigorous mathematical framework for comparing empirical distributions, such as point clouds. However, standard formulations based on the Wasserstein distances are highly sensitive to global rigid transformations in the feature space. While invariant OT formulations exist, they often disregard geometric discrepancies that are relevant for shape matching and can be computationally prohibitive (e.g. Gromov-Wasserstein). The Procrustes–Wasserstein (PW) distance addresses this limitation in a more scalable way by directly incorporating orthogonal alignments into the optimization problem. However, current approaches remain confined to discrete settings and lack generalization to unseen samples. In this work, we extend the PW framework by formally deriving its dual and semi-dual formulations. Leveraging results in both convex analysis and, more recently, neural optimal transport, we introduce NeuralPW, a framework that jointly learns rigid alignments and transport maps, being optimal whether the assumptions of the Brenier's theorem hold. We empirically study NeuralPW in a Gaussian setting, where both the optimal alignment and transport map admit closed-form solutions, and demonstrate that it accurately recovers geometric transformations, explicitly decoupling global pose alignment from local non-rigid deformation. Finally, we validate our approach on a real-world archaeozoological application involving the analysis of morphological deformations between biological specimens. The neural formulation enables high-fidelity matching of dense 3D point clouds at scales where discrete primal solvers become computationally intractable.

1 Introduction

By lifting the comparison of data from point-wise metrics to a geometric analysis of probability distributions, optimal transport (OT, Villani et al., 2009; Peyré et al., 2019) is attracting more and more interest in modern machine learning. Among the most impacted areas we mention generative modeling (Arjovsky et al., 2017; Gulrajani et al., 2017), imaging (Solomon et al., 2015; Feydy et al., 2017) and domain adaptation (Courty et al., 2016; Balaji et al., 2020). Central to this success is the Wasserstein distance (Kantorovich, 1942), which leverages the underlying geometry of a common metric space the distributions to be compared live in. While OT has broad utility, the Wasserstein distance has proven central in shape analysis and geometric matching (Eisenberger et al., 2020). In these settings, shapes are naturally represented as empirical measures, such as point clouds, and OT provides a theoretically grounded metric for comparisons, interpolations and deformation analysis. However, in some applications such as shape matching or clustering it is crucial to introduce some forms of invariance with respect to rigid transformations such as rotations or symmetries that, conversely impact the Wasserstein distance (e.g. the distance between a rotated shape and itself before rotation is not zero).

To address this "alignment gap", a branch of OT has focused on isometry invariant formulations that disentangle pose from shape. A first approach in this direction is represented by the Gromov-Wasserstein (GW) distance (Mémoli, 2011). GW relaxes the requirement of a shared ambient space by comparing distributions based on their intra-domain relational structures (e.g. pairwise similarity matrices) rather than absolute positions. This flexibility is powerful for matching objects across distinct metric spaces, such as in shape correspondence (Mémoli, 2011; Vayer et al., 2019), word embedding alignment (Alvarez-Melis & Jaakkola,

2018) and graph matching (Vayer et al., 2018; Vincent-Cuaz et al., 2021). Nevertheless, computing standard GW entails solving a non-convex quadratic optimization problem which is notoriously computationally expensive ($O(N^4)$ naively, with N denoting the number of points in the support of the empirical distributions to be compared). While approaches such as sliced (Vayer et al., 2019), quantized (Chowdhury et al., 2021) or regularized (Houry et al., 2026) versions of GW have improved scalability, GW methods often discard valuable ambient geometric information when distributions do reside in the same metric space.

For the specific but common scenario where distributions lie in the same space ($\mathbb{R}^d$) but differ by an unknown rigid transformation, the **Procrustes-Wasserstein** (PW, Zhang et al., 2017) distance offers a more specialized and efficient alternative. First formalized in the context of unsupervised word embedding alignment (Zhang et al., 2017; Grave et al., 2019; Alvarez-Melis et al., 2019), the PW distance couples the classical orthogonal Procrustes problem with OT matching. Zhang et al. (2017) formulated a joint optimization over a transport plan π and an orthogonal matrix $P \in \mathcal{O}(d)$ to simultaneously align the supports and match the masses of probability distributions. Subsequently, Grave et al. (2019) improved stability via convex relaxations over the permutation polytope, while Alvarez-Melis et al. (2019) introduced convexity annealing to eliminate the dependence on *ad hoc* initialization and generalized the PW framework to broader invariance classes. Recent contributions have further expanded both the application scope and the theoretical understanding of the PW framework. In the domain of cross-lingual learning, Ramírez et al. (2024) demonstrated that many prevailing unsupervised embedding alignment methods (e.g. MUSE and ICP) implicitly optimize a PW-like objective; they proposed a direct solver combining Sinkhorn iterations (Cuturi, 2013) with Procrustes analysis, yielding superior alignment over heuristic baselines. Beyond NLPs, Yang et al. (2022) applied the framework to unsupervised network alignment, introducing HackGAN, which integrates a PW loss within a generative (CycleGAN) architecture to map graph node embeddings across disparate networks. On the theoretical front, Even et al. (2024) established rigorous theoretical bounds for the PW problem, determining the precise sample complexity required to recover the ground-truth alignment in high-dimensional regimes ($d \gg \log N$) using alternating minimization strategies. Most recently, Adamo et al. (2025) formally established the metric properties of the PW distance over a quotient space of rotated/reflected discrete measures and provided an alternating minimization scheme, referred to as *ping-pong* in Even et al. (2024), for computing PW matching from samples. However, all these existing PW approaches remain confined to the discrete primal setting, meaning that they estimate a static matching matrix (transport plan) between samples and do not yield a transport map $T : \mathbb{R}^d \to \mathbb{R}^d$. Consequently, (i) they lack the ability to generalize to unseen samples or to provide a dense deformation field, limiting their insights and (ii) their computational cost is soon prohibitive.

Our contribution. We reformulate the Procrustes-Wasserstein optimization problem in the semi-dual form, possibly in the continuous domain, introducing a unified method that learns both rigid alignment and non-rigid deformation in a differentiable manner. Our main contributions are the following.

1. We provide a formal derivation of the dual and semidual formulations of the PW problem and an explicit characterization of the transport map (Theorem 3.2).

2. We introduce **NeuralPW** as a framework that leverage deep neural networks to estimate the PW transport map from the data.

3. We provide a detailed analysis of the PW problem between Gaussian distributions. The explicit closed-form solution for the PW distance and the associated map allows for the empirical demonstration of convergence of NeuralPW (Algorithm 1), which accurately disentangles global pose differences from local geometric variations, overcoming the limitations of Wasserstein neural OT solvers on unaligned data.

4. Finally, our approach is validated on a real-world scientific case study from archaeozoology, involving high-resolution 3D point clouds of animal skulls. We show that NeuralPW aligns and matches dense anatomical structures at scales where discrete primal solvers become computationally intractable.

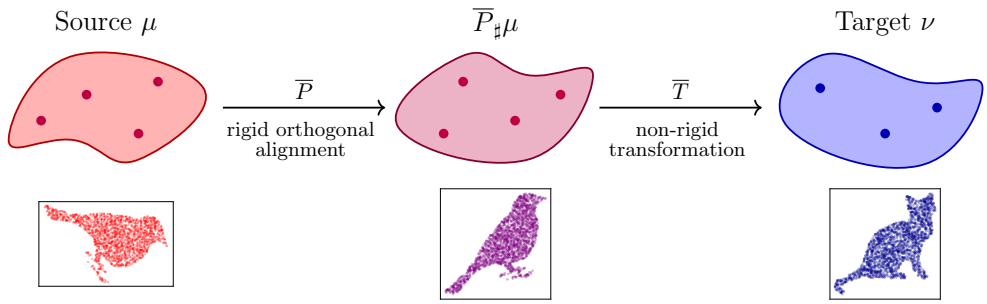

Figure 1: Overview of the Procrustes-Wasserstein landscape. A source and a target distribution, $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, differ by an orthogonal transformation (rotation and/or reflection) and a non-rigid deformation. The NeuralPW approach (Algorithm 1) jointly learns the optimal orthogonal alignment $\overline{P}$ and Monge transport map $\overline{T} : \mathbb{R}^d \to \mathbb{R}^d$ between $\overline{P}_\sharp\mu$ and ν.

2 OT background

Notation. In the following, we work with *centred* probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. In other words, when saying that $X \sim \mu$ we mean that X follows the discrete or continuous probability distribution μ, $\mathbb{E}(X) = 0$ and is intended to have finite second order moment, e.g. $\mathbb{E}(\|X\|_2^2) < \infty$. With a slight abuse of notation, $L^1(\mu)$ denotes the set of measurable functions f such that $\int f(x)d\mu(x) < \infty$.

2.1 Wasserstein distance

Given a pair of probability measures (μ, ν), the squared 2-Wasserstein problem is (Kantorovich, 1942)

$$W_2^2(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\|X - Y\|_2^2, \tag{OT}$$

where $\Pi(\mu, \nu)$ is the set of joint distributions (or *couplings*) $\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ with marginals μ and ν. In the above optimization problem, $W_2(\cdot, \cdot)$ defines a distance between μ and ν and for an in depth inspection of its properties the reader is referred to Villani et al. (2009); Santambrogio (2015); Peyré et al. (2019). Other Wasserstein distances (p-norms in place of the 2 norm in the above equation) as well as entropic optimal transport (Nutz, 2021) are not considered in this work. In case at least one between μ and ν has a density with respect to the Lebesgue measure (i.e. X or Y are continuous random variables) there exists a unique optimal transport or *Monge* map $T : \mathbb{R}^d \to \mathbb{R}^d$ linked with the (unique) optimal $\overline{\pi}$ from equation OT by

$$W_2^2(\mu, \nu) = \mathbb{E}_{(X,Y)\sim\overline{\pi}}\|X - Y\|_2^2 = \mathbb{E}_{X\sim\mu}\|X - T(X)\|_2^2,$$

where $T(X) \sim \nu$[1]. Moreover, equation OT admits the following dual formulation

$$W_2^2(\mu, \nu) = \sup_{(\varphi,\psi)\in\Phi} \big\{\mathbb{E}_\mu[\varphi(X)] + \mathbb{E}_\nu[\psi(Y)]\big\}, \tag{DOT}$$

where $\Phi := \big\{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu) \;\big|\; \forall(x, y) \quad \varphi(x) + \psi(y) \leq \|x - y\|_2^2\big\}$ and φ and ψ are usually referred to as the Kantorovich potentials. In case the Monge map T exists the above supremum is attained at $(\overline{\varphi}, \overline{\psi})$ and T is related to $\overline{\varphi}$ by the following relation

$$T(x) = \nabla(\|x\|_2^2 - 2\overline{\varphi}(x)). \tag{1}$$

The above facts relate to the Brenier's theorem and are proven in Peyré et al. (2019, Theorem 2.1) and Chewi et al. (2024, Theorem 1.14). Finally (for our purposes), the dual problem in equation DOT can be rephrased in a more convenient way, allowing us to make a parallel with convex analysis

$$W_2^2(\mu, \nu) = C_{\mu,\nu} - 2 \inf_{f\in L^1(\mu)} \big\{\mathbb{E}_\mu[f(X)] + \mathbb{E}_\nu[f^*(Y)]\big\}, \tag{2}$$

[1]This can be written in short-hand notation by means of the push-forward operator $\sharp$, i.e. $T_\sharp\mu = \nu$.

where $C_{\mu,\nu} := \mathbb{E}_{X\sim\mu}\left[\|X\|_2^2\right] + \mathbb{E}_{Y\sim\nu}\left[\|Y\|_2^2\right]$, $f(x) := \|x\|_2^2 - 2\varphi(x)$ and f^* is the convex conjugate of f (Appendix A). Note that, by definition of f and from equation 1 we have that $T(x) = \nabla\overline{f}(x)$. The *Monge* map and the semidual formulation of the Wasserstein problem are two key ingredients of this work.

Remark 2.1 (Transport maps and Monge maps). Brenier's theorem states that, under some assumptions, the unique *optimal* transport map or *Monge* map T exists and it is linked to the optimal Kantorovitch potential $\overline{\varphi}$ via equation 1. However, the dual and semidual formulations in equation DOT and equation 2, respectively, are independent from the assumptions of the Brenier's theorem. In case those assumptions do not hold (e.g. neither μ nor ν admit a density) one cas still compute a map $\nabla\overline{f}$ that (approximately) pushes forward μ to ν. This function is sometimes called barycentric map in the literature. We simply refer to it as *transport* map, keeping in mind that it might not be the optimal transport map.

2.1.1 Learning the Monge map from the data

Given observed realisations $x_1, \ldots, x_N$ and $y_1, \ldots, y_M$ from μ and ν, respectively, one might wish to estimate T from the data. This central topic is often treated "separately" in the statistical (e.g. Hütter & Rigollet, 2021; Manole et al., 2024) and deep learning (e.g. Makkuva et al., 2020; Korotin et al., 2022) rooms, although some works seek to make a bridge between the two communities (Divol et al., 2025). In the context of the 2-Wasserstein distance, the common point between most of the above cited works is the empirical counterpart of the semidual formulation of the Wasserstein problem 2, where μ and ν are replaced by their empirical cumulative distribution functions. A proper choice of the parametric functional class adopted to estimate the Kantorovich potential was proved to allow one to escape the curse of dimensionality and obtain a rate of convergence at least $\mathcal{O}(n^{-1/2})$ (Theorem 3.7, Chewi et al., 2024). In contrast, Manole et al. (2024) relies on the empirical counterpart of the primal formulation in equation OT and adopts a 1 nearest neighbor approach to map a new test data point x to the target domain, via the estimated transport plan $\hat{\pi}$. Still based on the semidual formulation in equation 2 the works of Makkuva et al. (2020) and Korotin et al. (2022) rely on neural networks to learn the *Monge* map. Makkuva et al. (2020) approximate a reparameterization of the dual potentials (φ, ψ) using input convex neural networks (ICNNs, Amos et al., 2017). This approach is discussed in more detail in next sections. More recently, Neural Optimal Transport (NOT, Korotin et al., 2022) rewrites the semidual objective as a saddle-point optimization problem (Corollary 1, Korotin et al., 2022). With the objective of finding a stochastic OT map $T(x, z)$, where z is a noise introduced to create stochasticity, NOT approaches the minimax problem directly parametrizing T and the dual potential g with general neural architectures such as multilayer perceptrons (MLPs). Indeed, sufficiently large NNs can arbitrarily well approximate stochastic transport maps between distributions (Theorem 1, Korotin et al., 2022).

3 The Procrustes-Wasserstein problem

The additional ingredients we now introduce are the orthogonal group $\mathcal{O}(d) = \{P \in \mathbb{R}^{d\times d} : P^T P = \mathbf{I}_d\}$ and the cost function $C_P(x, y) := \|Px - y\|_2$. The squared 2-Procrustes-Wasserstein cost is defined by

$$PW_2^2(\mu,\nu) := \inf_{\substack{P\in\mathcal{O}(d)\\ \pi\in\Pi(\mu,\nu)}} \mathbb{E}_{(X,Y)\sim\pi}\left[C_P(X,Y)^2\right]. \tag{Primal-PW}$$

This quantity was recently proven to be an OT distance with respect to a properly defined quotient space, which considers two measures as equivalent if they only differ by a rotation/symmetry of their support (Adamo et al., 2025). The following proposition links the PW_2 and W_2 distances.

Proposition 3.1. *For any $P \in \mathcal{O}(d)$, the Kantorovich transport cost between $P_\sharp\mu$ (the push-forward of μ by P) and ν is*

$$W_2^2(P_\sharp\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\,[C_P(X,Y)^2],$$

where $\Pi(\mu,\nu)$ is the set of couplings $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with marginals μ and ν.

Proof. See Appendix B. □

As an immediate consequence, one has

$$PW_2^2(\mu,\nu) = \inf_{P\in\mathcal{O}(d)} W_2^2(P_\sharp\mu,\nu), \tag{3}$$

meaning that solving the PW problem is equivalent to find the optimal orthogonal transformation P that minimizes the standard W_2 distance between the source $P_\sharp\mu$ and ν. The above formulation is particularly convenient since it immediately allows us to obtain the following **dual** formulation of equation Primal-PW

$$PW_2^2(\mu,\nu) = \inf_{P\in\mathcal{O}(d)} \sup_{(\varphi,\psi)\in\Phi_P} \mathbb{E}_{U\sim P_\sharp\mu}[\varphi(U)] + \mathbb{E}_{Y\sim\nu}[\psi(Y)], \tag{Dual-PW}$$

where $\Phi_P := \big\{(\varphi,\psi)\in L^1(P_\sharp\mu)\times L^1(\nu) \;\big|\; \forall(u,y)\ \varphi(u)+\psi(y)\le\|u-y\|_2^2\big\}$.

3.1 Optimal transport alignment problems

In the Procrustes-Wasserstein problem Primal-PW, the optimization is performed over the orthogonal group $\mathcal{O}(d)$. This setting is classical (Grave et al., 2019) and well motivated in the context of shape analysis (Adamo et al., 2025).

However, a natural question is whether the class of admissible transformations P can be enlarged. Following Alvarez-Melis et al. (2019) and Vayer (2020, Chapter 4), one may instead optimize over the Frobenius sphere

$$F_d := \{Q\in\mathbb{R}^{d\times d} \;:\; \|Q\|_{\mathcal{F}} = \sqrt{d}\},$$

which contains the orthogonal group as a strict subset, $\mathcal{O}(d)\subset F_d$. Under additional assumptions on the measures[2], it has been shown (Vayer, 2020, Theorem 4.2.1) (Alvarez-Melis et al., 2019, Lemma 4.3) that the corresponding OT alignment problem over F_d reduces to a particular inner-product Gromov–Wasserstein problem. The present work deliberately focuses on a different geometric setting. The two constraint sets are the spectral extremes of the Schatten-ℓ_p family of Alvarez-Melis et al. (2019). F_d is the $p=2$ (Frobenius) ball, while $O(d)$ is the $p=\infty$ corner. By restricting the alignment to orthogonal transformations, the PW framework preserves the Euclidean geometry of the ambient space and removes only rigid pose variability (rotations and reflections). In contrast, the larger class F_d contains transformations that may introduce anisotropic distortions while preserving only a global Frobenius-norm constraint. This distinction is particularly important in shape analysis applications. For example, in archaeological morphometrics (Section 7), rotations and reflections are generally regarded as nuisance transformations, whereas anisotropic deformations carry morphological information that should be quantified rather than absorbed into the alignment itself. The orthogonality constraint therefore provides a geometrically meaningful prior, allowing rigid alignment and shape deformation to be treated as separate phenomena (see Figure 2). A geometrical inspection of the PW-iGW correspondence is provided in Appendix C.

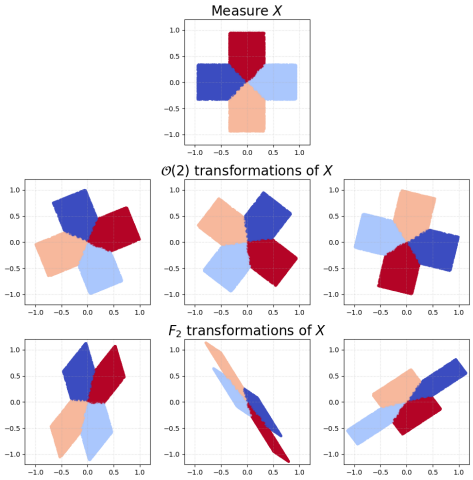

Figure 2: Examples of orthogonal and F_d transformations applied to a $d=2$ dimensional point cloud X. This visualization motivates the PW framework in specific shape analysis applications.

3.2 Semidual formulation of PW_2

We now focus on the dual formulation of $W_2(P_\sharp\mu,\nu)$. The inequality constraint in Φ_P can be read as

$$\begin{aligned}
&\varphi(u)+\psi(y)\le\|u-y\|_2^2\\
&\varphi(u)+\psi(y)\le\|u\|_2^2+\|y\|_2^2-2\langle u,y\rangle\\
&f(u)+g(y)\ge\langle u,y\rangle,
\end{aligned}$$

[2] Alvarez-Melis et al. (2019) consider a ν-whitening assumption on the target measure, or, equivalently, when every admissible P is angle-preserving.

where we adopted the same reparameterization seen in equation 2, namely $(f, g) \in L^1(P_\sharp\mu) \times L^1(\nu)$ and $f(u) = \|u\|_2^2 - 2\varphi$ and $g(y) = \|y\|_2^2 - 2\psi(y)$. Hence, for a fixed P, the dual problem now reads

$$W_2^2(P_\sharp\mu, \nu) = C_{\mu,\nu} - 2 \inf_{(f,g)\in\Phi'_P} \mathbb{E}_{U\sim P_\sharp\mu}[f(U)] + \mathbb{E}_{Y\sim\nu}[g(Y)], \tag{4}$$

with $\Phi'_P := \left\{(f, g) \in L^1(P_\sharp\mu)\times L^1(\nu) \;\middle|\; \forall(u, y) \quad f(u)+g(y) \geq \langle u, y\rangle\right\}$ and $C_{\mu,\nu}$ the same constant introduced previously. We can move one step forward by observing that, for a given, arbitrary $f \in L^1(P_\sharp\mu)$, the best we can do to minimize the above equation is to choose g as small as possible while satisfying the constraint $g(y) \geq \langle u, y\rangle - f(u)$, which is done by setting

$$g(y) = \sup_{u\in\mathrm{Dom}(f)} \{\langle u, y\rangle - f(u)\} =: f^*(y),$$

where f^* is the convex conjugate of f. Using the fact that f^* is convex by construction, and hence $f^{***} = f^*$ (Appendix A) as well as the symmetry of equation 4 in its arguments, it is not hard to show that the above problem simplifies to the following **semidual** formulation

$$W_2^2(P_\sharp\mu, \nu) = C_{\mu,\nu} - 2 \inf_{f\in L^1(P_\sharp\mu)} \mathbb{E}_{U\sim P_\sharp\mu}[f(U)] + \mathbb{E}_{Y\sim\nu}[f^*(Y)], \tag{5}$$

whose solution is given by a closed convex function $\bar{f} \in L^1(P_\sharp\mu)$. As a corollary, it holds that

$$\inf_{f\in L^1(P_\sharp\mu)} \mathcal{F}_{P_\sharp\mu,\nu}(f, f^*) = \inf_{\substack{f\in L^1(P_\sharp\mu)\\ f \text{ convex}}} \mathcal{F}_{P_\sharp\mu,\nu}(f, f^*). \tag{6}$$

Before discussing how to estimate the $\overline{f}$ solving the above minimization problem, we enounce the following

Theorem 3.2 (Characterization of the PW transport map)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with finite second order moments and assume μ is absolutely continuous w.r.t. the Lebesgue measure. Let $\overline{P} \in \mathcal{O}(d)$ be the optimal orthogonal transformation minimizing Primal-PW. Then the unique PW optimal transport map $T_{\mathrm{PW}} : \mathbb{R}^d \to \mathbb{R}^d$ is given by*

$$T_{\mathrm{PW}}(x) = \nabla\overline{f} \circ \overline{P}(x)$$

where there exists a unique convex function $\overline{f} : \mathbb{R}^d \to \mathbb{R}$, whose gradient is the Monge map between measures $\overline{P}_\sharp\mu$ and ν.

Proof. To prove the Theorem it suffices to explicit the link given by Proposition 3.1 and equation 3. Denoting $\overline{P}$ the minimizer of equation 3, then the problem to solve is $W_2^2(\overline{P}_\sharp\mu, \nu)$. Since μ is absolutely continuous with respect to the Lebesgue measure, the push-forward $\overline{P}_\sharp\mu$, under the diffeomorphism $x \mapsto \overline{P}x$, is also absolutely continuous. By Brenier's theorem (Peyré et al., 2019), there exists a unique Monge map $T_{\overline{P}_\sharp\mu\to\nu}$ (form $\overline{P}_\sharp\mu$ to ν), given by

$$T_{\overline{P}_\sharp\mu\to\nu}(u) = \nabla\overline{f}(u),$$

where $\overline{f}$ is a convex function. The PW transport map acting on x is therefore the gradient of f evaluated at $u = \overline{P}x$. □

4 Learning isometry invariant transport maps

Following the intuition of Makkuva et al. (2020), for any function $g \in L^1(\nu)$, differentiable ν-a.e., it holds that

$$\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y)) \leq \langle Y, \nabla f^*(Y)\rangle - f(\nabla f^*(Y)) = f^*(Y) \quad \nu\text{-a.e.}$$

due to the properties of the convex conjugate (Appendix A). In particular

$$\sup_{\substack{g\in L^1(\nu)\\ g \text{ diff.le } \nu\text{-a.e.}}} \mathbb{E}_{Y\sim\nu}\left[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))\right] = \mathbb{E}_{Y\sim\nu}[f^*(Y)].$$

Thus using that $\sup f = \inf(-f)$ we can re-rewrite equation 5 as

$$W_2^2(P_\sharp\mu,\nu) = C_{\mu,\nu} + 2 \sup_{f\in L^1(P_\sharp\mu)} \inf_{\substack{g\in L^1(\nu)\\ g \text{ diff.le } \nu\text{-a.e.}}} \mathcal{V}_{P_\sharp\mu,\nu}(f,g),$$

where $\mathcal{V}_{P_\sharp\mu,\nu}(f,g) := \mathbb{E}_{Y\sim\nu}\left[f(\nabla g(Y)) - \langle Y, \nabla g(Y)\rangle\right] - \mathbb{E}_{U\sim P_\sharp\mu}[f(U)]$. Finally a *computationally tractable* version of the dual PW objective of Dual-PW is

$$PW_2^2(\mu,\nu) = \inf_{P\in\mathcal{O}(d)} W_2^2(P_\sharp\mu,\nu). \tag{7}$$

4.1 Optimization over P, f and g.

The dependence of equation 7 on P can be made explicit by the following display

$$PW_2^2(\mu,\nu) = C_{\mu,\nu} + 2 \inf_{P\in\mathcal{O}(d)} \sup_{f\in L^1(\mu)} \inf_{g\in L^1(\nu)} \mathcal{V}_{\mu,\nu}(P,f,g), \tag{NeuralPW}$$

where g is differentiable ν-a.e. and $\mathcal{V}_{\mu,\nu}(P,f,g) := \mathbb{E}_{Y\sim\nu}\left[f(\nabla g(Y)) - \langle Y, \nabla g(Y)\rangle\right] - \mathbb{E}_{X\sim\mu}[f(PX)]$.

Implementation. We provide estimates for the optimal rigid alignment ($\overline{P}$) and transport map ($\nabla\overline{f}$) starting from the semidual PW formulation of equation NeuralPW. We treat the entries of the matrix $P \in \mathbb{R}^{d\times d}$ as learnable parameters. The semidual formulation naturally separates the roles of the two optimization variables f and g. The potential f appears through its convex conjugate and, under Brenier's theorem, the optimal solution is a convex function. For this reason, following Makkuva et al. (2020, Remark 3.4), we parameterize it using an ICNN, which enforces convexity by construction and is therefore consistent with the theoretical properties of the semidual problem. In contrast, no convexity constraint is required for g (Makkuva et al., 2020, Remark 3.4). The optimization only requires g to be differentiable almost everywhere. Moreover, equation NeuralPW depends on g exclusively through its gradient ∇g. We therefore directly parameterize the vector field ∇g using a standard multilayer perceptron (MLP), avoiding the unnecessary estimation of the scalar potential itself. This in not an entirely new idea, see for instance Dam et al. (2019) in a generative context.

Remark 4.1 (Convexity and neural nets). Although ICNNs provide a principled parameterization for f, an interesting question is whether explicit convexity constraints are always necessary in practice. During preliminary experiments with unconstrained MLPs, we frequently observed that the learned potential empirically behaved as a convex function despite the absence of architectural constraints. However, this observation remains purely empirical. Since the optimization problem is itself non-convex and no theoretical guarantees currently exist (at the best of our knowledge) for unconstrained neural parameterizations of the semidual objective, we adopt ICNNs throughout this work. Understanding under which conditions general neural architectures implicitly converge toward convex solutions represents an interesting direction that we leave for future research.

We denote by θ_f and θ_g the network parameters of f and ∇g, respectively, and we solve the optimization problem NeuralPW as a min-max-min game. At each iteration, we draw batches $\{x_i\}_{i=1}^R \sim \mu$ and $\{y_j\}_{j=1}^R \sim \nu$. The empirical objective function is defined as

$$\min_P \max_{\theta_f} \min_{\theta_g} \mathcal{J}(P,\theta_f,\theta_g) \tag{8}$$

where

$$\mathcal{J}(P,\theta_f,\theta_g) := \frac{1}{R}\sum_{i=1}^R f(\nabla g(y_i)) - \langle y_i, \nabla g(y_i)\rangle - f(Px_i).$$

We solved this using Adam (Kingma, 2014) for the potential parameters θ_f and θ_g and stochastic gradient descent (SGD) for the alignment matrix P. With respect to the orthogonality constraint of P, we perform a projection step via a SVD after the gradient updates. The full training procedure is summarized in Algorithm 1.

Algorithm 1 NeuralPW

Input: source μ, target ν, batch size R, inner iterations K_g, K_f, K_P, total iterations T
Initialize: $P \leftarrow P_0$ (*e.g. spectral alignment*)
for $t \in (1, \dots, T)$ **do**
  for $k \in (1, \dots, K_g)$ **do**
    Sample batches $\{x_i\}_{i=1}^R \sim \mu$, $\{y_j\}_{j=1}^R \sim \nu$
    Update θ_g to minimize $\mathcal{J}(P, \theta_f, \theta_g)$ via Adam
  end for
  for $k \in (1, \dots, K_f)$ **do**
    Sample batches $\{x_i\}_{i=1}^R \sim \mu$, $\{y_j\}_{j=1}^R \sim \nu$
    Update θ_f to maximize $\mathcal{J}(P, \theta_f, \theta_g)$ via Adam
  end for
  for $k \in (1, \dots, K_P)$ **do**
    Sample batch $\{x_i\}_{i=1}^R \sim \mu$
    Update P to minimize $\mathcal{J}(P, \theta_f, \theta_g)$ via SGD
    Project: $P \leftarrow UV^\top$ where $U, S, V^\top \leftarrow \mathrm{SVD}(P)$
  end for
end for

Remark 4.2. The min-max formulation in equation 8 resembles the objective considered by Carrasco et al. (2023). However, the two optimization problems originate from fundamentally different geometric settings and enforce different constraints on the alignment matrix. Carrasco et al. (2023) solves the inner-product Gromov–Wasserstein problem and optimizes P over the Frobenius sphere F_p. This choice is done with the explicit aim of handling *incomparable* spaces $d_x \neq d_y$. Our NeuralPW formulation instead solves the PW problem in a common space $d_x = d_y$. As shown in Appendix C, this is not a cosmetic difference. The two methods implicitly maximize different spectral functionals of the same cross-covariance matrix, and they coincide only when the optimal cross-covariance is isotropic.

4.2 Initialization

In discrete PW approaches (see e.g. Alvarez-Melis et al., 2019; Adamo et al., 2025), the optimization of problem Primal-PW typically starts with an initial guess of the coupling $\pi \in \Pi(\mu, \nu)$. In contrast, our continuous formulation, based on the optimization of the semi-dual objective NeuralPW, can be initialized just in therms of the orthogonal transformation P. The aim is to find a good enough rotation and/or reflection matrix $P_0 \in \mathcal{O}(d)$ that aligns the source and target measures. Since the joint optimization of (P, f, g) is non-convex, a random initialization of P may lead to suboptimal local minima. To mitigate this, we adopt a *spectral* initialization strategy. We effectively treat the (centred) measures μ and ν as if they were Gaussian $\mathcal{N}(0, \Sigma_\mu)$ and $\mathcal{N}(0, \Sigma_\nu)$ and, leveraging the closed-form solution derived in Theorem 5.1, we set $P_0 = UV^\top$, where U and V are the eigenvector matrices of Σ_μ and Σ_ν respectively (eigenvectors/eigenvalues are sorted). Geometrically, this operation corresponds to an initial Principal Component Analysis (PCA) alignment, ensuring that the principal axes of the source distribution are pre-aligned with those of the target.

5 PW between centered Gaussian measures

In this section, we derive a closed-form expression for the PW distance between two centered Gaussian distributions. As a generalization of the known closed Wasserstein formulation for Gaussian measures (Gelbrich, 1990; Bhatia et al., 2019; Masarotto et al., 2019), we demonstrate that, within the PW framework, the optimal alignment admits an analytical solution based on the spectral decomposition of the covariance matrices. A similar result was recently explored by Toukam (2025). However, we derived it independently from Toukam (2025) and via an alternative proof relying on the nuclear norm formulation of the transport cost and the closed form of the PW map.

Consider two anisotropic centered Gaussian measures $\mu = \mathcal{N}(0, \Sigma_\mu)$ and $\nu = \mathcal{N}(0, \Sigma_\nu)$ on $\mathbb{R}^d$. $\Sigma_\mu, \Sigma_\nu \in \mathbb{S}^d_{++}$ are symmetric positive definite matrices. Recall that the classical squared 2-Wasserstein distance between μ and ν has a closed form

$$W_2^2(\mu, \nu) = \mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(\Sigma_\nu) - 2\mathrm{tr}\big[(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}\big].$$

The squared 2-Procrustes-Wasserstein distance incorporates an optimization over the orthogonal group $\mathcal{O}(d)$ to align the distributions

$$PW_2^2(\mu, \nu) = \inf_{P \in \mathcal{O}(d)} W_2^2(\mu, P_\sharp \nu)$$

where $P_\sharp \nu = \mathcal{N}(0, P\Sigma_\nu P^\top)$, e.g. the covariance of the rotated target is given by $P\Sigma_\nu P^\top$. Substituting this into $W_2^2(\mu, P_\sharp \nu)$, we obtain

$$PW_2^2(\mu, \nu) = \inf_{P \in \mathcal{O}(d)} \big\{\mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(P\Sigma_\nu P^\top) - 2\mathcal{T}(P)\big\}, \tag{9}$$

where $\mathcal{T}(P) := \mathrm{tr}\big[(\Sigma_\mu^{1/2} P\Sigma_\nu P^\top \Sigma_\mu^{1/2})^{1/2}\big]$. Observing that the trace is invariant under cyclic permutations and thanks to the orthogonality of of P, we have $\mathrm{tr}(P\Sigma_\nu P^\top) = \mathrm{tr}(\Sigma_\nu)$. Consequently, minimizing the PW distance is equivalent to maximizing the coupling term $\mathcal{T}(P)$. We seek thus for

$$\overline{P} = \arg\max_{P \in \mathcal{O}(d)} \mathcal{T}(P).$$

Let the spectral decomposition of the covariances be

$$\Sigma_\mu = U\Lambda_\mu U^\top \qquad \text{and} \qquad \Sigma_\nu = V\Lambda_\nu V^\top,$$

where Λ_μ, Λ_ν are diagonal matrices containing the sorted eigenvalues (either both in ascending or descending order) of Σ_μ and Σ_ν, and $U, V \in \mathcal{O}(d)$ are the respective eigenvector matrices.

Theorem 5.1. *The squared PW_2 distance between centered Gaussian measures μ and ν admits the closed form*

$$PW_2^2(\mu, \nu) = tr(\Sigma_\mu) + tr(\Sigma_\nu) - 2\sum_{i=1}^{d} \sqrt{\lambda_\mu^i \lambda_\nu^i}, \tag{10}$$

where $\{\lambda_\mu^i\}_{i=1}^d$ and $\{\lambda_\nu^i\}_{i=1}^d$ are the sorted eigenvalues of Σ_μ and Σ_ν, respectively. Moreover, the optimal alignment is $\overline{P} = UV^\top$ and the optimal PW transport map is

$$T_{PW}(x) = V\Lambda_\nu^{1/2}\Lambda_\mu^{-1/2}U^\top x.$$

Proof. See Appendix D. □

Theorem 5.1 has the following geometric intuition: denoting $X \sim \mu$ and $Y \sim \nu$ the source and the target measures, respectively, the PW transport map T_{PW} first rotates (via $U^\top$) the source X to align its principal axis with the canonical ones; then $\Lambda_\mu^{-1/2}$ scales to the unit sphere and $\Lambda_\nu^{1/2}$ stretch and match to the target's eigenvalues. Finally V optimally align to the target's orientation (the eigenvector basis). A practical behaviour of such effects is reported in Figure 3 and particularly visible (in decoupled actions) in Figure 8 in the Appendix.

6 Experiments

In this section we validate NeuralPW on synthetic Gaussian measures and demonstrate its utility for point cloud matching based on the MNIST dataset. MNIST data are centered and scaled to the unit ball. Moreover, in the specific case of MNIST dataset, we initialized NeuralPW using the spectral strategy described in Section 4.2.

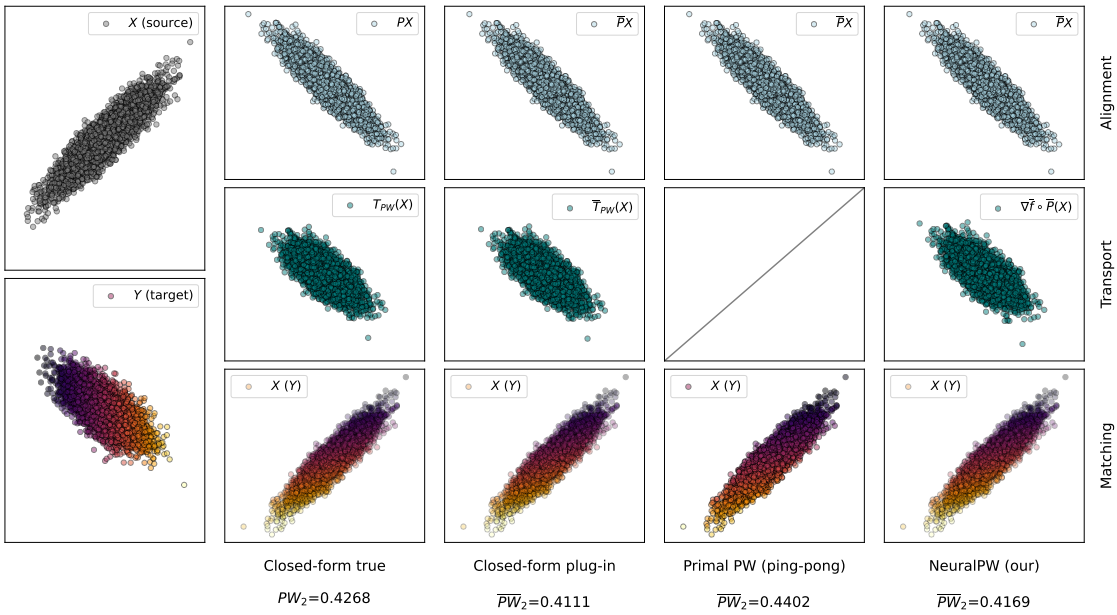

Figure 3: PW comparison between two Gaussian measures X and Y (left). Top row: the source is aligned to the target. Middle row: transported source. Bottom row: visualized source-target matching (color transfer). Estimated quantities are reported with an overline ($\bar{\cdot}$).

6.1 The Gaussian case

We first consider a controlled setting where the ground truth PW distance is known (Theorem 5.1). We first generate source $X \sim \mathcal{N}(0, \Sigma_\mu)$ and target $Y \sim \mathcal{N}(0, \Sigma_\nu)$ point clouds ($N = M = 5000$) as shown in Figure 3. Then our aim is to match and transport (whenever possible) the source data to the target with three different approaches: (i) plug-in estimation, i.e. the sample covariance matrices (and their spectral decomposition) are injected the closed-forms in Theorem 5.1; (ii) discrete OT, solving the primal PW optimization problem via the alternating minimization solver referred to as *ping-pong* (Adamo et al., 2025; Even et al., 2024); (iii) our NeuralPW approximation.

Figure 3 reports in rows the estimated alignment ($\overline{P}_\sharp\mu$), transport ($\overline{T}_\sharp(\overline{P}_\sharp\mu)$) and matching across methods (in columns). All methods correctly estimate the true rotation $\overline{P}$ (first row). In the second row one sees the transported X, not available for primal based methods. The bottom row shows the color gradient transfer from Y to X, based on the estimated matching between points. Whereas in primal based methods this can be done via the estimated transport plan, for all other methods (including ground truth), we obtain the discrete matching by assigning each transported source point $\overline{T}_{PW}(x_i)$ to its nearest neighbor in the target Y. NeuralPW achieves a squared distance of 0.4169, closely matching the analytical ground truth 0.4268 and recovers a smooth transport map that closely approximate the theoretical one.

6.2 Semantic clustering of MNIST digits

We now illustrate the advantage of PW over classical optimal transport distances on a simple point clouds clustering task. Standard neural Wasserstein solvers (Makkuva et al., 2020) are highly sensitive to the global pose of the input distributions. We compare this behaviour with two rotational invariant alternatives: the discrete PW solver (either ping-pong algorithm or our semidual approach), and the Gromov-Wasserstein (GW) distance.

We consider a subset of the MNIST dataset composed of six images of the digit "1" and six images of the digit "3", converted into two-dimensional point clouds. Two experimental settings are investigated: (i) *aligned*, where the digits retain their original orientation; and (ii) *rotated*, where each point cloud is independently

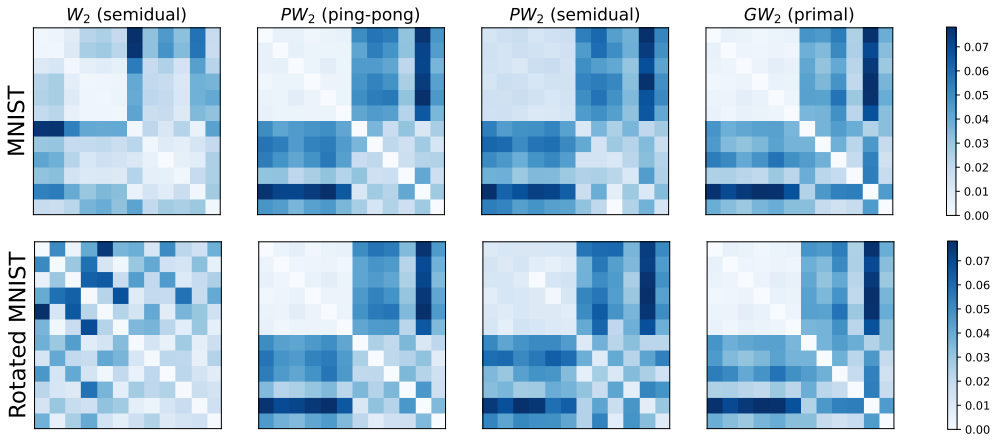

Figure 4: OT-based distance matrices between twelve MNIST digits. The first six rows/columns of each matrix are related to the handwritten digit "1" of MNIST, while the last six rows/columns are related to the handwritten digit "3". The first column of the grid displays Wasserstein distances (W_2) computed using the approach of Makkuva et al. (2020) (semidual). The second and third columns display Procrustes-Wasserstein distance matrices (PW_2) computed using the discrete primal solver (ping-pong) and our NeuralPW approach (semidual). Finally the last column displays Gromov-Wasserstein distances computed with the POT toolbox. In the first row, MNIST digits "1" and "3" (transformed into point clouds) are directly compared, with the pose and orientation that they have in the original dataset. In the second row, the same digits are compared after applying a random rotation.

Table 1: Agglomerative hierarchical clustering performance, with $K = 2$, obtained from the OT-based pairwise distance matrices of Figure 4. The experiments are performed on the original MNIST point clouds (top row of Figure 4) and the rotated version (bottom row of Figure 4). NeuralPW consistently recovers the correct partition in both settings, while the standard Wasserstein distance is sensitive to pose variations and the Gromov-Wasserstein distance struggles to capture the intra-class variability of handwritten digits.

Dataset	Metric	W_2 (semidual)	PW_2 (ping-pong)	PW_2 (semidual)	GW_2 (primal)
MNIST	ARI	0.6648	1.0000	1.0000	0.0638
	NMI	0.6615	1.0000	1.0000	0.2313
Rotated MNIST	ARI	-0.0879	1.0000	1.0000	0.0638
	NMI	0.0000	1.0000	1.0000	0.0638

transformed by a random rotation $R \in SO(2)$. For each setting, we compute the complete 12×12 pairwise distance matrix using four different OT-based metrics: the neural semidual approximation of W_2 (Makkuva et al., 2020); the discrete PW ping-pong solver; our NeuralPW semidual approach; and the primal GW solver from the POT library (Flamary et al., 2024).

The resulting distance matrices are displayed in Figure 4. To assess how well each OT metric captures the semantic class structure, we perform agglomerative (hierarchical) clustering with $K = 2$ clusters directly on the distance matrices and evaluate the obtained partition using the Adjusted Rand Index (ARI) and the Normalized Mutual Information (NMI, Vinh et al., 2010). The corresponding quantitative results are reported in Table 1.

Several observations emerge from this experiment. First, the standard Wasserstein geometry, being not invariant with respect to isometries, is already affected by the intrinsic pose variability present in the aligned dataset. It achieves ARI and NMI scores around 0.66, which correspond to a single clustering mistake among the twelve samples. As expected, in the case of rotated MNIST, the W_2 distance drastically deteriorates. In

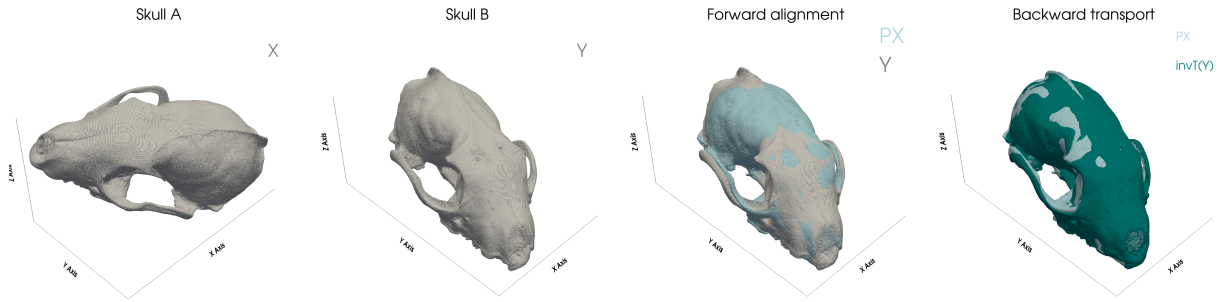

Figure 5: Procrustes–Wasserstein matching between two high-resolution Pine Marten skulls. The source specimen (CA025, first column) and target specimen (CA023, second column) contain approximately $N = 10^6$ vertices each. NeuralPW was trained on $20K$ uniformly sampled points from each scan and subsequently evaluated on the complete clouds, yielding a dense deformation field at full resolutions at no additional training cost. The learned orthogonal transformation P aligns the source skull with the target orientation (PX, third column), while the learned transport map ∇g (denoted "InvT") produces a dense non-rigid deformation field that morphs target anatomy toward the aligned source (fourth column).

contrast, both PW approaches correctly recover the underlying class structure. Both the discrete ping-pong solver, initialized with spectral strategy (Adamo et al., 2025, Fiedler-W, Section 4.1), and our NeuralPW semidual formulation achieve perfect clustering (ARI=NMI=1.0) in the aligned and rotated settings. These results confirm the expected invariance of the PW framework to rigid transformations and demonstrate that NeuralPW faithfully reproduces the geometric behaviour of the discrete PW formulation while relying on a fundamentally different optimization strategy. Finally, the GW distance performs considerably worse in the experiment. Although GW is intrinsically invariant to rigid motions, it compares shapes solely through their internal pairwise distance matrices. The variability in the handwritten realizations of the digit 3 produces substantial differences in these intrinsic geometries, making samples from the same semantic class less similar than expected. Consequently, GW fails to recover a coherent cluster corresponding to the digit 3, resulting in low ARI and NMI scores in both experimental settings. While these quantitative differences should not be over-interpreted given the small dataset regime, they nevertheless illustrate the different geometric inductive biases of the considered OT formulations. In this setting, where the dominant nuisance variability arises from rigid pose changes together with moderate handwriting deformations, the PW geometry is substantially the most discriminative.

Overall, this experiment demonstrates that the observed improvement originates from the PW geometry itself rather than from a particular numerical implementation. Both the discrete ping-pong algorithm and the proposed NeuralPW framework recover the correct clustering. This motivates the use of the PW geometry for point cloud analysis, while the proposed neural semi-dual formulation naturally extends its applicability to larger-scale problems, as illustrated in the following section on dense real-world objects.

7 NeuralPW for the study of morphological deformation in archaezoology

To illustrate both the scalability and the practical utility of NeuralPW when dealing with large 3D point clouds, we present a real-world case study from archaeozoology. Due to its established theoretical foundations, optimal transport has the potential to be a valuable tool in support of archaeozoologists seeking to document historical animal biodiversity and quantify the impact of anthropogenic pressures on faunal communities over time. In particular, we used NeuralPW to detect refined delineation of extinct taxa and lineages exhibiting specific adaptations to prolonged environmental fluctuations and human-induced selective pressures. To address similar tasks, bioarchaeologists have relied on established approaches such as Geometric Morphometrics (GMM, Bookstein, 1997) and diffeomorphometry (Durrleman et al., 2013). These methods allow one to quantify the morphological variability in both living and fossil organisms and are particularly well suited to the archaeological record. One of their main strengths lies in their adaptability to a wide range of biological structures (organs, bones and teeth). However, while powerful, they often face

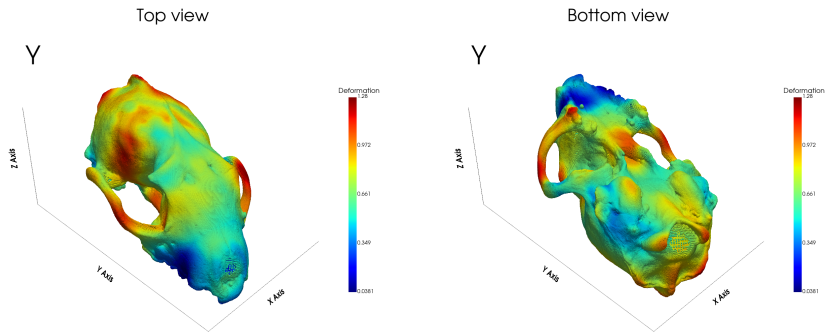

Figure 6: Deformation magnitude applied to the target Skull B (CA023) to become Skull A (CA025). Colors represent the norm of the displacement induced by the learned inverse transport field ($\overline{\nabla g}$), highlighting anatomical regions requiring larger geometric adaptations to match the aligned source specimen. Warmer colors indicate stronger local morphological differences, whereas cooler colors correspond to regions exhibiting high structural similarity between the two skulls.

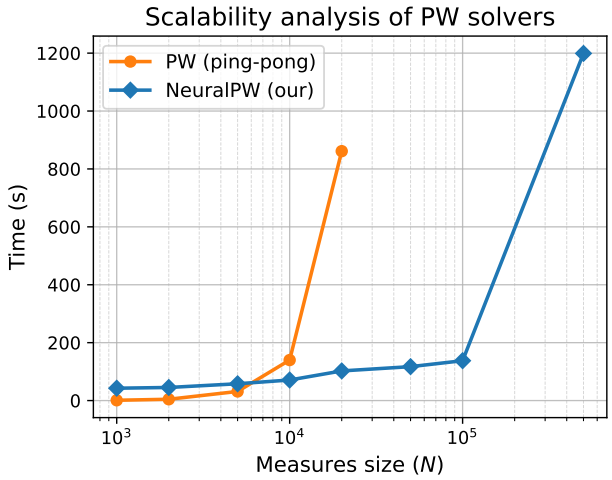

Figure 7: Runtime comparison (in seconds) between the discrete *ping-pong* solver (Adamo et al., 2025) and our NeuralPW framework. The x-axis represents the point cloud size N on logarithmic scale. While the discrete solver suffers from prohibitive computational scaling as N increases, NeuralPW maintains a manageable computational footprint, enabling a robust transport of high-density distributions ($N > 10^5$).

two critical limitations: (i) reliance on manual registration or landmark placement to align 3D shapes, introducing human bias; and (ii) the use of simplified morphological templates (averages) to infer evolutionary scenarios.

NeuralPW overcomes these limitations by performing automated pairwise comparisons of high-resolution point clouds. By jointly learning the alignment and transport , our framework eliminates the need for preprocessing steps like ICP (Chetverikov et al., 2002) registration. Furthermore, the learned transport map acts as a data-driven diffeomorphism, capturing the dense deformation between specimens without reducing complex anatomies to sparse landmarks. This facilitates a more robust interpretation of species evolution from global pose to local anatomical details.

7.1 Case study: Pine Marten skulls

We selected two complete skulls of modern Pine Martens (*Martes martes*) from the osteological collection of [Anonymous Laboratory] (Figure 10, in Appendix). The specimens, referred to as Skull A (ID: CA025)

and Skull B (ID: CA023), correspond to two adult males (Figure 11, in Appendix). Standard osteometric measurements (Von Den Driesch, 1976) reveal distinct size variations. Skull B measures 8.3cm in length (Akrokranion to Prosthion) and 5.2cm in Zygomatic breadth, while Skull A is "smaller", measuring 7.7cm and 4.8cm, respectively (see Table 2, in Appendix, for a full description). The specimens were digitized using an Artec Spider surface scanner following the protocol in Vuillien (2020), resulting in high-density meshes ($N \approx 1$ million vertices). The use of high-resolution 3D models is strictly necessary to provide a detailed and reliable assessment of the morphological features of the two martens, allowing the discussion of intraspecific characteristics.

We perform a PW "matching" between the two scans, considering a uniform downsample of 20K vertices. We use these downsampled clouds to train our NeuralPW framework with a fixed batch size of 1024 and for a maximum of 30K epochs. The potential f and gradient potential ∇g are optimized with learning rates of $3 \cdot 10^{-4}$, while for the orthogonal matrix P we fixed a learning rate of $\cdot 10^{-3}$.

Remark 7.1 (Direction of transport). In this archaeological experiment we estimate the transport from ν to $P_\sharp \mu$ rather than the forward map $T_{PW} : \mu \to \nu$. This is a deliberate consequence of the asymmetric roles played by f and g in equation NeuralPW. The objective is a max-min problem in which f is the outer (maximized) variable and g the inner (minimized) one. At each iteration, ∇g is trained to near-optimality against the current f, whereas f receive less ascent step per inner loop (i.e. $K_f < K_g$). Therefore, throughout training, ∇g is a better-optimized object that ∇f. Moreover, at the optimum $\nabla \overline{g} = \nabla \overline{f}^*$ is the inverse of the forward map, so it pushes ν onto $\overline{P}_\sharp \mu$ and is itself a valid transport map. The same choice is made by (Makkuva et al., 2020). Concerning the application, the deformation field of interest is the correspondence between the two skull anatomies, and reporting it as the morphing of Skull B onto the aligned Skull A is as informative as the converse.

Scalability of PW solvers. A key advantage of our neural approximation is its ability to scale to regimes where discrete solvers become intractable. Discrete OT solvers, including the alternating *ping-pong* algorithm (Even et al., 2024; Adamo et al., 2025), require instantiating an $N \times M$ cost matrix at each step, and therefore scale at least quadratically with the number of samples, $O(N^2)$ per iteration for entropic inner solvers and $O(N^3 \log N)$ for exact linear programming. In contrast, NeuralPW has a per-iteration cost governed by the batch size R and the network capacity. However, reporting a single runtime for NeuralPW is ambiguous, since we empirically observe two distinct convergence regimes. The estimated cost $\overline{PW}_2$ stabilizes after relatively few iterations, whereas the transport field ∇g requires substantially longer to converge. This gap is consistent with the structure of the objective in equation NeuralPW. The PW value is a functional of the potentials themselves, while the map is recovered through their gradients, so that a near-optimal value is compatible with a still-inaccurate map. To make the fair comparison with ping-pong, which returns a coupling and hence a cost, but no map, we report the time for NeuralPW to reach convergence of the PW cost alone. Figure 7 reports runtimes across varying downsampled resolutions of Skull A and Skull B. While *ping-pong* exhibits explosive computational cost, becoming prohibitive and memory-constrained beyond $N = 10^4$, NeuralPW remains tractable up to $N \approx 5 \times 10^5$. This allows bioarchaeologists to leverage the high resolution of modern 3D scanners without artificial downsampling.

Discussion. The benefit of the neural semidual approach (NeuralPW) is in its capability of learning alignments and transport maps that extends beyond the training samples. As the archaeological application demands, the focus on high-resolution point clouds is extremely important. Our NeuralPW framework estimates an optimal alignment $\overline{P}$ and inverse transport $\nabla \overline{g}$, yielding an estimated out-of-sample squared distance of $PW_2^2 = 3.1079$. More importantly for the application, as shown in Figure 5, the method accurately disentangles pose from shape. The aligned source $\overline{P}X$ matches the orientation of the target Y, while the transported target $\nabla \overline{g}(Y)$ successfully morphs the anatomical features of Skull B to match Skull A with high-fidelity. It is then possible to identify the specific anatomical regions affected by deformation, including the zygomatic process and the basioccipital region (Figures 12 and 13). Within the zygomatic region, deformation is most pronounced in the frontal process and along the orbital surface of the zygomatic bone (Barone, 1976). In the basioccipital region, significant differences are observed in the shape of the occipital condyles and the nuchal tubercle. Specifically, Skull A exhibits distinct curvature in the region

between the condyles and the tubercle compared to Skull B. This characteristic is specifically visible by the transported target $\nabla\overline{g}(Y)$ in green (see Figure 13, in appendix).

The primary contribution of this work is to enable the investigation of long-term morphological changes in species by visualizing and quantifying localized deformations of anatomical regions from 3D point clouds. By considering the complete external morphology of skeletal remains, our approach opens new avenues for investigating the long-term evolution of animal biodiversity. It enables the reconstruction of deformation trajectories in response to biological, environmental, and anthropogenic constraints, thereby providing a more precise characterization of morphological changes at both the inter- and intra-population levels. As such, this work extends the existing methodological framework developed to address these questions (Durrleman et al., 2009; Beaudet et al., 2018; Cucchi et al., 2023; 2024; Vuillien et al., 2026) by introducing a novel perspective based on NeuralPW. Future work will focus on extending NeuralPW to population-level analysis, and possibly highlighting the connections with PW barycenters. Beyond archaeology, this work contributes to the statistical modeling of 3D shapes, with potential applications in paleontology, anthropology, evolutionary biology, and ecology.

8 Conclusion

In this work, we have extended the Procrustes–Wasserstein framework beyond the discrete primal setting to which it had so far been confined. Starting from the observation that solving PW amounts to minimizing a Wasserstein problem over the orbit of the source measure under orthogonal transformations, we derived the dual and semidual formulations of the problem and characterized the PW transport map as $T_{PW} = \nabla f \circ P$, the composition of a rigid alignment with a Brenier-type map. We have then estimated the PW cost and maps, via a neural network-based approach named NeuralPW, an adversarial optimization scheme that parameterizes the potentials with neural networks and treats the alignment as a constrained learnable matrix. We validated the approach in a Gaussian setting, for which we derived a closed-form expression of both the PW distance and its map, and where NeuralPW recovers the analytical solution while explicitly decoupling pose from deformation. On MNIST point clouds, we have demonstrated that the advantage over other OT geometries stems from the PW geometry itself rather than from a particular solver, highlighting the value of such isometry invariant framework. Finally, an archaeozoological case study on high-resolution skulls demonstrated that the neural formulation operates at scales where discrete primal solvers become intractable, producing interpretable deformation fields over anatomical regions of morphological interest.

References

Davide Adamo, Marco Corneli, Manon Vuillien, and Emmanuelle Vila. An in depth look at the procrustes-wasserstein distance: properties and barycenters. *arXiv preprint arXiv:2507.00894*, 2025.

David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1881–1890, 2018.

David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1870–1879. PMLR, 2019.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International conference on machine learning*, pp. 146–155. PMLR, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

Robert Barone. Anatomie comparée des mammifères domestiques. 2ème édition revue et augmentée, tome 1, 2 fascicules. *Vigot Freres, Paris*, pp. 281, 1976.

Amélie Beaudet, Jean Dumoncel, Frikkie De Beer, Stanley Durrleman, Emmanuel Gilissen, Anna Oettlé, Gérard Subsol, John Francis Thackeray, and José Braga. The endocranial shape of australopithecus africanus: surface analysis of the endocasts of sts 5 and sts 60. *Journal of Anatomy*, 232(2):296–303, 2018.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones mathematicae*, 37(2):165–191, 2019.

Fred L Bookstein. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical image analysis*, 1(3):225–243, 1997.

Xavier Aramayo Carrasco, Maksim Nekrashevich, Petr Mokrov, Evgeny Burnaev, and Alexander Korotin. Uncovering challenges of solving the continuous gromov-wasserstein problem. *arXiv preprint arXiv:2303.05978*, 2023.

Dmitry Chetverikov, Dmitry Svirko, Dmitry Stepanov, and Pavel Krsek. The trimmed iterative closest point algorithm. In *2002 International Conference on Pattern Recognition*, volume 3, pp. 545–548. IEEE, 2002.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 3, 2024.

Samir Chowdhury, David Miller, and Tom Needham. Quantized gromov-wasserstein. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 811–827. Springer, 2021.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Thomas Cucchi, H Harbers, D Neaux, Marie Balasse, L Garbé, D Fiorillo, H Bocherens, D Drucker, Clément Zanolli, R Cornette, et al. 4500 years of morphological diversification in western europe wild boars (sus scrofa) and the consequences of the neolithic transition. *Quaternary Science Reviews*, 309:108100, 2023.

Thomas Cucchi, D Neaux, L Féral, F Goussard, H Adriensen, F Elleboudt, G Sansalone, and R Schafberg. How domestication, feralization and experience-dependent plasticity affect brain size variation in sus scrofa. *Royal Society Open Science*, 11(9):240951, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Nhan Dam, Quan Hoang, Trung Le, Tu Dinh Nguyen, Hung Bui, and Dinh Phung. Three-player wasserstein gan via amortised duality. In *International Joint Conference on Artificial Intelligence 2019*, pp. 2202–2208. Association for the Advancement of Artificial Intelligence (AAAI), 2019.

Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Optimal transport map estimation in general function spaces. *The Annals of Statistics*, 53(3):963–988, 2025.

Stanley Durrleman, Xavier Pennec, Alain Trouvé, and Nicholas Ayache. Statistical models of sets of curves and surfaces based on currents. *Medical image analysis*, 13(5):793–808, 2009.

Stanley Durrleman, Xavier Pennec, Alain Trouvé, José Braga, Guido Gerig, and Nicholas Ayache. Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *International journal of computer vision*, 103(1):22–59, 2013.

Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, and Daniel Cremers. Deep shells: Unsupervised shape correspondence with optimal transport. *Advances in Neural information processing systems*, 33:10491–10502, 2020.

Mathieu Even, Luca Ganassali, Jakob Maier, and Laurent Massoulié. Aligning embeddings and geometric random graphs: Informational results and computational approaches for the procrustes-wasserstein problem. *Advances in Neural Information Processing Systems*, 37:70730–70764, 2024.

Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *International conference on medical image computing and computer-assisted intervention*, pp. 291–299. Springer, 2017.

Rémi Flamary, Cédric Vincent-Cuaz, Nicolas Courty, Alexandre Gramfort, Oleksii Kachaiev, Huy Quang Tran, Laurène David, Clément Bonet, Nathan Cassereau, Théo Gnassounou, Eloi Tanguy, Julie Delon, Antoine Collas, Sonia Mazelet, Laetitia Chapel, Tanguy Kerdoncuff, Xizheng Yu, Matthew Feickert, Paul Krzakala, Tianlin Liu, and Eduardo Fernandes Montesuma. Pot python optimal transport (version 0.9.5), 2024. URL https://github.com/PythonOT/POT.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1880–1890. PMLR, 2019.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Roger A Horn and Charles R Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.

Guillaume Houry, Jean Feydy, and François-Xavier Vialard. Gromov-wasserstein at scale, beyond squared norms. *arXiv preprint arXiv:2602.06658*, 2026.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pp. 199–201, 1942.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, 2024.

Valentina Masarotto, Victor M Panaretos, and Yoav Zemel. Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhya A*, 81(1):172–213, 2019.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

Marcel Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Guillem Ramírez, Rumen Dangovski, Preslav Nakov, and Marin Soljačić. On a novel application of wasserstein-procrustes for unsupervised cross-lingual alignment of embeddings. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, pp. 1–11, 2024.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

Kevine Meugang Toukam. Procrustes wasserstein metric: A modified benamou-brenier approach with applications to latent gaussian distributions. *arXiv preprint arXiv:2503.16580*, 2025.

Titouan Vayer. A contribution to optimal transport on incomparable spaces. *arXiv preprint arXiv:2011.04447*, 2020.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. *arXiv preprint arXiv:1805.09114*, 2018.

Titouan Vayer, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced gromov-wasserstein. *Advances in Neural Information Processing Systems*, 32, 2019.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty. Online graph dictionary learning. In *International conference on machine learning*, pp. 10564–10574. PMLR, 2021.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

Angela Von Den Driesch. *A guide to the measurement of animal bones from archaeological sites*, volume 1. Peabody museum press, 1976.

Manon Vuillien. *Systèmes d'élevage et pastoralisme en Provence et dans les Alpes méridionales durant la Protohistoire: Nouvelles perspectives en archéozoologie*. PhD thesis, Université Côte d'Azur, 2020.

Manon Vuillien, Emmanuelle Vila, Jwana Chahoud, Bea De Cupere, Moussab Albesso, Agraw Amane, Sofiane Bouzid, Homa Fathi, Pauline Garberi, Haeedeh Laleh, et al. Exploring biological and ecological components of sheep astragalus size and shape variation using 3d geometric morphometrics: Towards a bioarchaeological proxy. *Journal of Archaeological Method and Theory*, 33(4):58, 2026.

Linyao Yang, Xiao Wang, Jun Zhang, Jun Yang, Yancai Xu, Jiachen Hou, Kejun Xin, and Fei-Yue Wang. Hackgan: Harmonious cross-network mapping using cyclegan with wasserstein–procrustes learning for unsupervised network alignment. *IEEE Transactions on Computational Social Systems*, 10(2):746–759, 2022.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1934–1945, 2017.

A Convexity

Consider $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, then f is said to be convex if and only if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \qquad \forall x, y, \in \mathbb{R}^d,\ \forall t \in [0,1].$$

If $f \in \mathcal{C}^1(\mathbb{R}^d)$ then f is convex iff $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$. Moreover, a function $f \in \mathcal{C}^2(\mathbb{R}^d)$ is convex iff $\mathrm{H}f(x) \geq 0$ (i.e. the Hessian is at least positive semi-definite).

Definition A.1. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. The *convex conjugate* of f is the function $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined by

$$f^*(x^*) := \sup_{x \in \mathbb{R}^d} \{\langle x, x^*\rangle - f(x)\}.$$

Note that for every function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, f^* is a convex function. Moreover f^* is closed (e.g. $(f^*)^{-1}_{(-\infty,\ a]}$ is closed in $\mathbb{R}^d$).
From the definition of the convex conjugate we have the *Fenchel-Young inequality* that reads

$$f(x) + f^*(y) \geq \langle x, y\rangle, \tag{11}$$

where equality holds if and only if $y \in \nabla f(x)$, e.g. $f(x) + f^*(\nabla f(x)) = \langle x, \nabla f(x)\rangle$. By symmetry

$$f(\nabla f^*(y)) + f^*(y) = \langle \nabla f^*(y), y\rangle \tag{12}$$

B The Procrustes-Wasserstein problem

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and let $X \sim \mu$, $Y \sim \nu$ be a source and target distributions, respectively. Fix the orthogonal group $\mathcal{O}(d) = \{P \in \mathbb{R}^{d\times d} : P^T P = I\}$. For $P \in \mathcal{O}(d)$ define the cost $c_P(x, y) := \|Px - y\|_2$.

B.1 Equivalent primal formulation

We start with the following

Proposition B.1. *For any $P \in \mathcal{O}(d)$, the Kantorovich transport cost between $P_{\#}\mu$ (the pushforward of μ by P) and ν is*

$$W_2^2(P_{\#}\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{X,Y\sim\pi}\ [c_P(X, Y)^2],$$

where $\Pi(\mu, \nu)$ is the set of couplings $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with marginals μ and ν.

Proof. First let us remark that if two random variables $X \sim \mu$ and $U \sim P_{\#}\mu$ are considered, with μ admitting a density denoted by $p_X(\cdot)$, than by change of variables in multiple integrals we have

$$\mathbb{E}(U) = \mathbb{E}(PX) = \int_{\mathbb{R}^D} Pxp_X(x)dx = \int_{\mathbb{R}^D} up_X(P^T u)|\det P^T|du = \int_{\mathbb{R}^D} up_X(P^T u)du$$

where the last equality comes from the orthogonality of P and shows that

$$p_U(u) = p_X(P^T u) \tag{13}$$

where $p_U(\cdot)$ is nothing but the density of $P_{\#}\mu$. Now, by definition of Wasserstein (squared) distance it holds that $W_2^2(P_{\#}\mu, \nu) := \min_{\gamma\in\Pi(P_{\#}\mu,\nu)} \mathbb{E}_{U,Y\sim\gamma}\left[||U - Y||_2^2\right]$. Moreover

$$\begin{aligned}\mathbb{E}_{U,Y\sim\gamma}\left[||U - Y||_2^2\right] &= \int_{\mathbb{R}^d\times\mathbb{R}^d} ||u - y||_2^2 \gamma(u, y)dudy \\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} ||Px - y||_2^2 \gamma(Px, y)dxdy = \mathbb{E}_{X,Y\sim\pi}||PX - Y||_2^2,\end{aligned}$$

where change of variables was employed again and we defined $\pi(x, y) := \gamma(Px, y)$. It is easy to check that π is a well defined density in $\Gamma(\mu, \nu)$. In particular

$$\int_{\mathbb{R}^d} \pi(x, y)dy = \int_{\mathbb{R}^d} \gamma(Px, y)dy = p_U(Px) = p_X(x),$$

where the penultimate and last equalities come from the marginal constraints on γ and Eq. equation 13, respectively. Finally, since $\pi = \gamma \odot (P, Id)$ and P is an isomorphism, the correspondence between π and γ is one to one and the proposition is proven. □

Thence, the squared 2-Procrustes-Wasserstein cost is defined by

$$PW_2^2(\mu, \nu) := \min_{P \in \mathcal{O}(p)} \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{X,Y\sim\pi} \left[c_P(X, Y)^2\right] = \min_{P \in \mathcal{O}(p)} W_2^2(P_\# \mu, \nu),$$

telling us that we "just" need to find the orthogonal transformation P that minimizes the usual W_2^2 distance between the rotated source $P_\# \mu$ and ν.

B.2 Dual formulation

In order to derive the dual formulation of the Procrustes-Wasserstein problem, we come back to the marginal constraint imposed to the joint distribution π. We can reformulate these constraint, by saying that $\pi \in \Pi(\mu, \nu)$ if for all bounded measurable functions $\varphi, \psi : \mathbb{R}^d \to \mathbb{R}$

$$\mathbb{E}_{x,y\sim\pi} \left[\varphi(x)\right] = \mathbb{E}_{x\sim\mu} \left[\varphi(x)\right], \qquad \mathbb{E}_{x,y\sim\pi} \left[\psi(y)\right] = \mathbb{E}_{y\sim\nu} \left[\psi(y)\right].$$

Let us now fix $P \in \mathcal{O}(d)$. We introduce potential functions (Lagrange multipliers) $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$. The Lagrangian that enforces marginal constraints if

$$\begin{aligned}\mathcal{L}(\pi, \varphi, \psi) &= \mathbb{E}_{X,Y\sim\pi} \left[c_P(X, Y)^2\right] + \mathbb{E}_{X\sim\mu} \left[\varphi(X)\right] - \mathbb{E}_{X,Y\sim\pi} \left[\varphi(X)\right] + \mathbb{E}_{Y\sim\nu} \left[\psi(Y)\right] - \mathbb{E}_{X,Y\sim\pi} \left[\psi(Y)\right] \\ &= \mathbb{E}_{X,Y\sim\pi} \left[c_P(X, Y)^2 - \varphi(X) - \psi(Y)\right] + \mathbb{E}_{X\sim\mu} \left[\varphi(X)\right] + \mathbb{E}_{Y\sim\nu} \left[\psi(Y)\right].\end{aligned}$$

By the classical Kantorovich duality, the inner infimum over π admits the dual formulation

$$W_2^2(P_\# \mu, \nu) = \sup_{(\varphi,\psi)\in\Phi_P} \left\{\mathbb{E}_{X\sim\mu}[\varphi(X)] + \mathbb{E}_{Y\sim\nu}[\psi(Y)]\right\},$$

where

$$\Phi_P := \left\{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu) \;\middle|\; \forall (x, y) \quad \varphi(x) + \psi(y) \leq \|Px - y\|_2^2, \quad (\mu, \nu) - a.e.\right\}.$$

Putting back the outer infimum over the orthogonal transformations, the Procrustes-Wasserstein dual formulation becomes

$$PW_2^2(\mu, \nu) = \inf_{P \in \mathcal{O}(p)} \sup_{(\varphi,\psi)\in\Phi_P} \left\{\mathbb{E}_{X\sim\mu}[\varphi(X)] + \mathbb{E}_{Y\sim\nu}[\psi(Y)]\right\}.$$

C Geometric comparison of different optimal transport alignment problems

In this section we precise the comparison introduced in Section 3.1 between the Procrustes–Wasserstein (PW) problem and the inner-product Gromov–Wasserstein (iGW) problem (Vayer, 2020, Chapter 4). We show that, when the two measures live in a common space $\mathbb{R}^d$, both problems can be written as the optimization of a spectral functional of the same cross-covariance matrix, and that they differ functional used: the nuclear norm for PW and a Frobenius norm for iGW. This isolates the geometric meaning of the orthogonality constraint and explains why PW is not, in general, an instance of inner-product GW alignment problem.

Consider $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. For a coupling $\pi \in \Pi(\mu, \nu)$ we define the cross-covariance matrix

$$V_\pi := \int yx^\top \mathrm{d}\pi(x, y) \in \mathbb{R}^{d\times d}. \tag{14}$$

The inner-product Gromov–Wasserstein problem. The inner-product GW objective between μ and ν is

$$\mathrm{iGW}(\mu,\nu) \;=\; \inf_{\pi\in\Pi(\mu,\nu)} \iint \big(\langle x,x'\rangle - \langle y,y'\rangle\big)^2 \,\mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y'). \tag{15}$$

Under finite fourth moment for μ and ν, (Vayer, 2020, Theorem 4.2.1) shows that equation 15 is equivalent to the bilinear problem

$$\sup_{\pi\in\Pi(\mu,\nu)} \sup_{P\in F_d} \int \langle Px, y\rangle \,\mathrm{d}\pi(x,y), \qquad F_d := \big\{P\in\mathbb{R}^{d\times d} : \|P\|_F = \sqrt{d}\big\}. \tag{MaxOT}$$

We first recall the elementary expansion that brings V_π into play. The MaxOT formulation can be rewritten as

$$\sup_{\pi\in\Pi(\mu,\nu)} \sup_{P\in F_d} \int \langle Px, y\rangle \mathrm{d}\pi(x,y) = \sup_{\pi\in\Pi(\mu,\nu)} \sup_{P\in F_d} \langle P, V_\pi\rangle.$$

Now, the inner supremum over P of MaxOT is a Cauchy–Schwarz maximization on the Frobenius sphere. Vayer (2020, Lemma 4.2.2) shows that, for a fixed coupling π,

$$\sup_{P\in F_d} \langle P, V_\pi\rangle \;=\; \sqrt{d}\,\|V_\pi\|_F, \quad \text{attained at } P^\star = \sqrt{d}\,V_\pi/\|V_\pi\|_F. \tag{16}$$

Substituting equation 16 into MaxOT probelm gives the surrogate objective

$$\sup_{\pi\in\Pi(\mu,\nu)} \sqrt{d}\|V_\pi\|_F. \tag{17}$$

The Procrustes–Wasserstein problem. Here we place PW in the same cross-covariance frame. Expanding the quadratic Wasserstein cost and using $W_2^2(P_\sharp\mu,\nu) = \min_\pi \mathbb{E}_\pi\|Px-y\|_2^2$ (Proposition 3.1), we can write

$$\begin{aligned}\mathrm{PW}_2^2(\mu,\nu) &= \inf_{P\in O(d)} \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi\big[\|Px\|_2^2 + \|y\|_2^2 - 2\langle Px, y\rangle\big] \\ &= C_{\mu,\nu} - 2 \sup_{\pi\in\Pi(\mu,\nu)} \sup_{P\in O(d)} \langle P, V_\pi\rangle. \end{aligned} \tag{18}$$

Now, we recall now that as a direct consequence of the Von Neumann's trace inequality, we have that for any matrix $A\in\mathbb{R}^{d\times d}$,

$$\max_{P\in\mathcal{O}(d)} \langle P, A\rangle = \|A\|_*,$$

with the maximum attained at $P^\star = UV^\top$, where $U\Sigma V^\top$ is a singular value decomposition of A.

We can therefore use this result in the PW problem 18, to get the following optimization problem

$$\sup_{\pi\in\Pi(\mu,\nu)} \|V_\pi\|_*, \tag{19}$$

which reads PW in terms of a nuclear-norm maximization of the coupling cross-covariance.

Comparison. Equations equation 17 and equation 19 place the two problems on a common footing. Both maximize a norm of V_π over the *same* feasible set of couplings $\Pi(\mu,\nu)$, and differ only in the norm,

$$\underbrace{\sup_\pi \|V_\pi\|_*}_{\mathrm{PW}} \qquad \text{vs.} \qquad \underbrace{\sup_\pi \sqrt{d}\,\|V_\pi\|_F}_{\mathrm{iGW}}.$$

The relation between the two functionals is governed by the spectral inequality below.

Proposition C.1 (PW and iGW relation)**.** *For every $\pi \in \Pi(\mu, \nu)$,*

$$\|V_\pi\|_* \;\leq\; \sqrt{d}\,\|V_\pi\|_F,$$

with equality if and only if all singular values of V_π are equal, i.e. $V_\pi = c\,Q$ for some $c \geq 0$ and $Q \in \mathcal{O}(d)$. Consequently if some maximizer $\pi^\star$ of the PW problem yields a $V_{\pi^\star}$ with all singular values equal, then PW and iGW select the same alignment value at $\pi^\star$ However, in general, the two problems may be maximized at different couplings, since they reward different spectral profiles of V_π.

Proof. Let $\sigma(V_\pi) = (\sigma_1, \ldots, \sigma_d)$ denote the vector of singular values of V_π. Then $\|V_\pi\|_* = \sum_{i=1}^d \sigma_i = \langle \sigma(V_\pi), \mathbf{1}\rangle$, with $\mathbf{1}$ the unitary vector. By Cauchy–Schwarz inequality, $\|V_\pi\|_* \leq \|\sigma(V_\pi)\|_2 \|\mathbf{1}\|_2 = \sqrt{d}\|V_\pi\|_F$, since the Frobenius norm is the Euclidean norm of the singular value vector. Equality in Cauchy–Schwarz holds if and only if $\sigma(V_\pi) = c\mathbf{1}$ for some $c \geq 0$, that is, if and only if all singular values of V_π are equal. Finally, the PW and iGW objectives differ only by replacing the nuclear norm with the scaled Frobenius norm of the same matrix V_π. Therefore, for any fixed coupling π, the two objectives agree exactly when V_π has equal singular values, and otherwise satisfy the strict inequality. This completes the proof. □

D Proof for Theorem 5.1

We report here the detailed proof for Theorem 5.1.

Proof. **(Closed PW formulation)** We want to maximize the term $\mathcal{T}(P) = \mathrm{tr}\big[(\Sigma_\mu^{1/2} P \Sigma_\nu P^\top \Sigma_\mu^{1/2})^{1/2}\big]$.

First, notice that for any matrix A, the term $\mathrm{tr}((AA^\top)^{1/2})$ corresponds to the nuclear norm $\|A\|_* = \sum_i \lambda^i$, where λ^i are the singular values of A. Let us define $A := \Sigma_\mu^{1/2} P \Sigma_\nu^{1/2}$. Then

$$\mathcal{T}(P) = \|\Sigma_\mu^{1/2} P \Sigma_\nu^{1/2}\|_*.$$

Now, considering the diagonalization of the covariance matrix Σ_μ, we have that $\Sigma_\mu^{1/2} = U\Lambda_\mu^{1/2}U^\top$ (respectively for ν). Substituting the spectral decompositions and exploiting the unitary invariance of the nuclear norm, e.g. $\|XYZ\|_* = \|Y\|_*$ for unitary X, Z (X, Z are real orthogonal matrices), we have

$$\begin{aligned}\mathcal{T}(P) &= \|U\Lambda_\mu^{1/2}U^\top P V \Lambda_\nu^{1/2} V^\top\|_* \\ &= \|\Lambda_\mu^{1/2}(U^\top P V)\Lambda_\nu^{1/2}\|_*.\end{aligned}$$

Let $Q := U^\top P V$. Since $U, P, V \in \mathcal{O}(d)$, it follows that $Q \in \mathcal{O}(d)$. The optimization problem transforms into

$$\max_{Q \in \mathcal{O}(d)} \|\Lambda_\mu^{1/2} Q \Lambda_\nu^{1/2}\|_*.$$

Using the Von Neumann trace inequality (or the singular value inequality for products of matrices), we have the upper bound

$$\|\Lambda_\mu^{1/2} Q \Lambda_\nu^{1/2}\|_* \leq \sum_{i=1}^d \sqrt{\lambda_\mu^i}\sqrt{\lambda_\nu^i},$$

where the singular values of $\Lambda_\mu^{1/2}$ and $Q\Lambda_\nu^{1/2}$ are $\sqrt{\lambda_\mu^i}$ and $\sqrt{\lambda_\nu^i}$, respectively.

This upper bound is achieved when Q is the identity matrix, e.g. $\overline{Q} = \mathbf{I}_d$. Recovering the optimal rotation P, we have $\mathbf{I}_d = U^\top \overline{P} V \implies \overline{P} = UV^\top$. Finally, substituting the matrix value $\mathcal{T}(\overline{P}) = \sum_{i=1}^d \sqrt{\lambda_\mu^i \lambda_\nu^i}$ into Eq. equation 9 yields the stated result of Eq. equation 10.

(Recovering the PW map) The *aligned* target distribution is drawn from $\overline{P}_\sharp \nu$, where his covariance matrix is

$$\Sigma_{\overline{P}_\sharp \nu} = \overline{P}\Sigma_\nu(\overline{P})^\top = (UV^\top)(V\Lambda_\nu V^\top)(VU^\top) = U\Lambda_\nu U^\top.$$

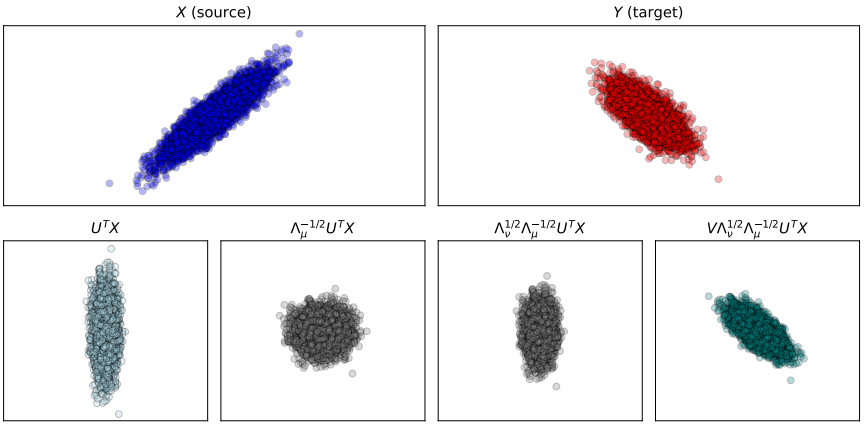

Figure 8: Geometric interpretation of the PW map between two Gaussian measures $X \sim \mathcal{N}(0, \Sigma_\mu)$ and $Y \sim \mathcal{N}(0, \Sigma_\nu)$.

Note that the source covariance Σ_μ and the *aligned* target covariance $\Sigma_{\overline{P}_\sharp\nu}$ share the same eigenvectors matrix U. Therefore, they commute.

We recall that the 2-Wasserstein *Monge* map between μ and $\overline{P}_\sharp\nu$ is $T(x) = \Sigma_\mu^{-1/2}(\Sigma_\mu^{1/2}\Sigma_{\overline{P}_\sharp\nu}\Sigma_\mu^{1/2})^{1/2}\Sigma_\mu^{-1/2}x$. Since, in our context, Σ_μ and $\Sigma_{\overline{P}_\sharp\nu}$ commute, the *Monge* map simplifies drastically to element-wise scaling of the eigenvalues (Masarotto et al., 2019), e.g.

$$T(x) = U\Lambda_\nu^{1/2}\Lambda_\mu^{-1/2}U^\top x.$$

Since T maps $X \sim \mu$ to $Y' \sim \overline{P}_\sharp\nu$, and we know that $Y' = \overline{P}Y$ (where $Y \sim \nu$), we can recover the target Y by applying the inverse rigid transform $(\overline{P})^\top$, e.g. $Y = (\overline{P})^\top Y' = (\overline{P})^\top T(X)$. Substituting we get

$$T_{PW}(x) = (VU^\top)(U\Lambda_\nu^{1/2}\Lambda_\mu^{-1/2}U^\top)x = V\Lambda_\nu^{1/2}\Lambda_\mu^{-1/2}U^\top x.$$

□

Remark D.1. While we prove the statement by exploiting the formulation of $\mathcal{T}(P)$ in terms of the nuclear norm, Toukam (2025) approaches the maximization via the singular value inequality for matrix product, i.e. $\sum_i \lambda_{AB}^i \leq \sum_i \lambda_A^i \lambda_B^i$ for couples of matrices A, B (Theorem 3.3.14, Horn & Johnson, 1994). Algebraically, these derivations are equivalent, as both rely on the Von Neumann trace inequality which dictates that the trace is maximized when the principal axes of the spectral decomposition are aligned (yielding to $Q = \mathbf{I}_d$). Nevertheless, we provide here the explicit formulation of the PW *Monge* map T_{PW}, while Toukam (2025) does not detail it. They only observe that, in this context, the *Monge* map is the one between the optimally rotated source and the target distributions.

E Experimental settings

We provide here the detailed experimental setup and hyperparameters to ensure the reproducibility of the results reported in Section 6.

NeuralPW implementation details. For the experiments of Section 6.2 and Section 7, we initialized the alignment matrix P using the spectral initialization strategy described in Section 4.2. The potential $f : \mathbb{R}^d \to \mathbb{R}$ was parameterized using an ICNN architecture with 4 hidden layers and Leaky ReLU activation as in (Makkuva et al., 2020), whereas $\nabla g : \mathbb{R}^d \to \mathbb{R}^d$ was parameterized using an MLP with 2 hidden layers and Softplus activation. We employed the alternating optimization scheme described in Algorithm 1 with the following inner iteration settings: $K_g = 10$, $K_f = 2$ and $K_P = 10$. A crucial detail of our implementation

concerns the orthogonality constraint on P. While standard projected Gradient Descent requires projecting P onto the manifold $\mathcal{O}(d)$ (via SVD) after every gradient step, we performed the orthogonal projection only once at the end of the outer loop (iteration t), rather than after each inner update $k = 1, \ldots, K_P$. We empirically observed that delaying the projection allows the optimization trajectory to traverse the ambient Euclidean space more freely before snapping back to the manifold, resulting in faster convergence and more stable minima.

Gaussian case. We generated source and target data by sampling $N = M = 5000$ points from two centered Gaussian measures $\mathcal{N}(0, \Sigma_\mu)$ and $\mathcal{N}(0, \Sigma_\nu)$. NeuralPW was trained for $T = 2000$ total iterations with a batch size of $R = 500$. The network parameters θ_f and θ_g were optimized using Adam with learning rates of $5 \cdot 10^{-4}$, while the matrix P was optimized using SGD with a learning rate of $2 \cdot 10^{-3}$. For the discrete baseline comparison, we utilized the alternating minimization (*ping-pong*) scheme from Adamo et al. (2025, Algorithm 1), setting the convergence tolerance to 10^{-4} and the maximum number of alternating steps (e.g. *ping-pong* steps) to 30, which was sufficient for convergence in all trials.

MNIST case. We selected the first six images from the MNIST test dataset for classes 1 and 3 (Figure 9, left). Each 28×28 image was converted into a 2D point cloud by extracting the coordinates of active pixels. To normalize the data, we centered each cloud at its Euclidean barycenter and scaled it to be enclosed within the unit sphere (Figure 9, center). This constitutes our *aligned* scenario. For the *rotated* scenario, we applied a random rotation to each point cloud individually (Figure 9, right). We compared NeuralPW against the standard neural OT solver of Makkuva et al. (2020). For the baseline W_2 (Makkuva et al., 2020), we parameterized the potentials f and g using two ICNNs with 4 hidden layers. We trained using the minimax game reported in Algorithm 1 of Makkuva et al. (2020) for $T = 1000$ iterations with a batch size of 50. The learning rates were set to $2 \cdot 10^{-3}$ for θ_f and 10^{-4} for the dual potential θ_g. For PW_2 (ping-pong) we followed the very same setting as for the Gaussian case, but using a Fiedler initialization for the initial transport plan. For NeuralPW, we maintained the same batch size ($R = 50$) and a maximum of training iterations of $T = 1000$, using the architecture and projection strategy described above to estimate the PW_2 (semidual) distance. Finally, we used the POT toolbox, with the default settings, for estimating the GW_2 matching.

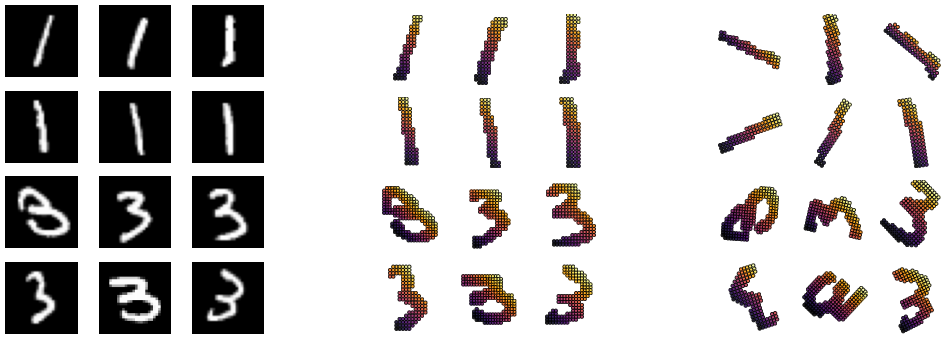

Figure 9: Left: the 12 MNIST images considered. Center: the associated 2D point clouds used in the *aligned* scenario. Right: the randomly *rotated* point clouds.

F Additional material for Section 7

Figure 10: Picture of a modern Pine Marten (*Martes martes*) © Martre des pins, un animal commun de nos forêts

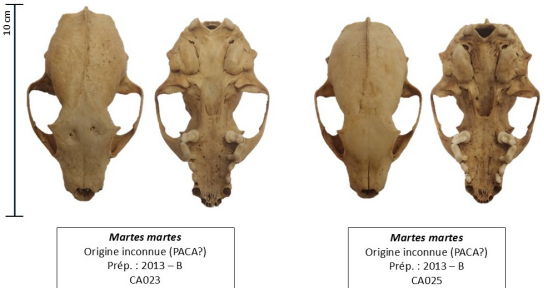

Figure 11: Pictures of Marten skulls used in the case study. Collection of © [Anonymous Laboratory].

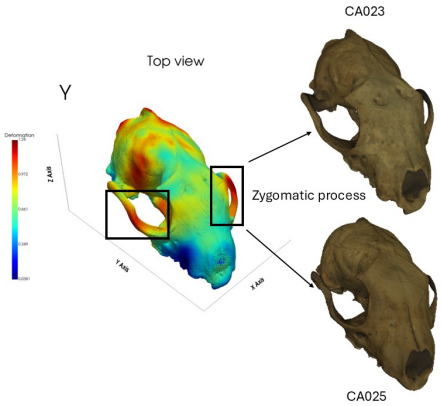

Figure 12: Morphological variation of the zygomatic region between the two Pine Marten specimens. On the right: textured 3D model view © [Anonymous Author]

Table 2: Inventory for the two modern specimens of *Martes martes* belonging to [Anonymous Laboratory]. All the measurements are in millimeters. Following Von Den Driesch (1976), columns refer to: (1) total length between Akrokranion - Prosthion; (23) Zygomatic breadth; (18) greatest mastoid breadth; (19) greatest breadth of the occipital condyles; (20) greatest breadth of the foramen magnum; (21) height of the foramen magnum.

Specimen	(1)	(23)	(18)	(19)	(20)	(21)
CA023	82.53	51.75	39.10	20.78	11.43	9.80
CA025	77.25	48.03	38.87	18.84	10.54	7.30

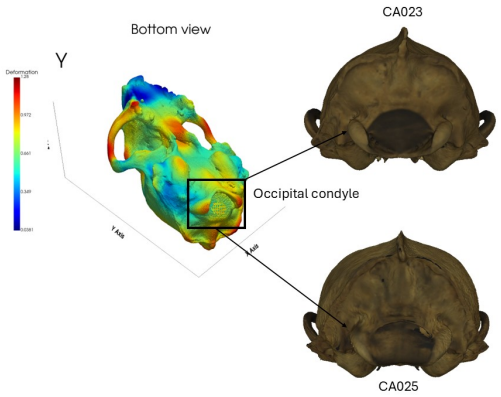

Figure 13: Morphological variation of the basioccipital region between the two Pine Marten specimens. On the right: textured 3D model view © [Anonymous Author]