Euclidean Embedding of Data Using Local Distances

Dimitris Arabadjis
Department of Statistics and Actuarial-Financial Mathematics
University of the Aegean
Karlovassi, Samos, PA 83200
darampatzis@aegean.gr

Abstract

We study the problem of recovering a globally consistent Euclidean embedding of data, given only a local distance graph and propose a method that optimally represents these distances. The method operates solely on a neighborhood graph weighted by pairwise distances, without requiring any prior vector representation of the data. The embedding is obtained by solving a variational problem that matches local, on-graph distances to the Euclidean metric, induced by the differentials of the embedding functions. The resulting Euler–Lagrange equations are derived in a coordinate-free form, enabling direct evaluation of all operators from the distance graph alone. Though non-linear and missing an explicit expression for their non-linearity, these equations are shown to be resolved as an iteratively updated sparse linear problem. The main contributions of the proposed approach are (a) the derivation of the functional equations governing the optimal Euclidean embedding in the continuum, (b) a representation-free formulation that requires only a neighborhood distance graph and no feature vectors and (c) an estimation procedure based exclusively on local graph operations. We experimentally evaluate the resulting non-parametric algorithm on synthetic manifolds and real datasets, demonstrating consistent preservation of local metric structure and neighboring relations, while approximating the global isometric embedding.

1 Introduction

The intention of this work is to introduce an approach to the problem of embedding data into Euclidean vector spaces, based on rigorous optimality guarantees (Sect. 3), while retaining generality and applicability. Specifically, the developed methodology avoids any reference to prior representations of the considered data, other than their distance graph. Moreover, the operations applied remain local, processing and producing sparse data.

Actually, the driving aim of the proposed approach is to approximate the globally optimal Euclidean embedding that IsoMap [1] achieves via Multidimensional Scaling (MDS) [2] on the full matrix of the datapoints' pair-wise geodesic distances, using only local distance data.

The methodological gap related with this aim is identified to the fact that methods which operate on local data, though capable to retain these data in the computed embedding, they fail to lift them globally and they distort the geodesic distances of the original data manifold. This is not a methodological deficiency but an implication of the optimization problems that are resolved in these methods. Specifically, according to the Minimum-Distortion prototype, introduced in [3], many of the existing local embedding methods resolve, point-wise, a(n) (un)constrained optimization problem formulated in terms of the embedding's local distances. Within this framework, even if one asks that these distances match the true data distances, the approximation holds only locally, due to the fact that the embedding is computed pointwise, without evaluating that there should be a (continuous/smooth) mapping underlying this embedding. Consequently, if pointwise optimization is not lifted, only the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

data distances involved in the optimization procedure could be approximated by the embedding's Euclidean distances, thus rendering the global embedding methods unavoidable.

The proposed approach drops the pointwise formulation of the embedding problem and uses functional optimization to determine the optimal mapping of the data manifold in the Euclidean space that locally approximates data distances. Then, integrability of the local approximations lifts this correspondence globally, without incorporating the actual shortest path distances between non-neighboring datapoints.

Local embedding methods that fall within the Minimum-Distortion prototype mainly split in two classes; the *spectral methods* that identify the embedding with the null space of the linear evaluator of the data manifold features that should be retained and the *stochastic methods* that iteratively construct embeddings, whose local Euclidean distances induce a stationary distribution equivalent to the non-stationary distribution induced by the original data distances.

Spectral methods: a) Locally Linear Embedding (LLE) [4] that locally matches the metric induced by the data distances and the metric of the embedding (see also [5]), b) Local Tangent Space Alignment (LTSA) [6] that locally matches the eigenvectors of these metrics, c) Laplacian Eigenmaps [7] that detects the embedding of the minimal total variation, d) Hessian Eigenmaps [8] that detects the embedding that is in the kernel of the Hessian operator

Stochastic methods: a) t-SNE [9] is the source point of the whole class and it is based on KL-divergence minimization between the embedding and data distributions, b) LargeVis [10] modifies the distributions' comparison, incorporating non-local points' relationships and c) Uniform Manifold Approximation and Projection (UMAP) [11] employs fuzzy operations to symmetrize datapoints' similarity graph and cross-entropy to symmetrize distributions' comparison.

Concerning the optimality of the determined embedding, spectral methods determine the global minimizers of the considered quadratic optimization problems, while the non-linearity of the stationarity equations of the stochastic methods' probabilistic objectives, leads to iterative (stochastic) gradient descent approximations. The stationarity equations determined under the proposed approach are highly non-linear (Sect. 3.1). However, the iterative scheme developed to resolve them (Sct. 3.2) provably converges to the set of critical points of the approximation error. Moreover, these stationarity equations are functional, not point-wise algebraic ones, naturally interpreting the point-wise embedding results as evaluations of the embedding functions. So, though these equations are properly discretized on graphs (Sct. 4) to obtain pointwise evaluations of the embedding, the integrability of these evaluations is intrinsic in the equations' formulas and their discrete realization on graphs, thus lifting the sought local isometric approximations to a global approximation.

2 Graph-based formulation of the Euclidean embedding problem

The core idea underlying the proposed methodology is that a smooth embedding, whose Euclidean distances optimally approximate the corresponding geodesic distances on the data manifold, is a vector of scalar fields whose differentials optimally align with infinitesimal, on-manifold transports that trivialize the manifold's metric (they turn it into the identity matrix). While the directions of such transports can always be found via spectral decomposition of the metric, only flat manifolds allow for isometrically deform them into integrable entities. The determination of the closest integrable entities to metric-trivializing infinitesimal transports is the problem considered in the proposed methodology.

Technically, this problem is formulated and resolved in terms of differential 1-forms, which are the integrands of line integrals, thus formally representing infinitesimal transports. However, these entities and their associated operators that resolve the problem have immediate analogues on graphs, which are analytically derived in appendix B and outlined in the following Scts. 2.1 and 2.2.

2.1 Local Euclidean frames from distances

Let $\mathcal{G}$ be a distance graph, namely an undirected graph, weighted by the connected nodes' mutual distances. While centered on an arbitrary vertex, say v_i, if the shortest path (geodesic) distances between its neighbors were available, then one could construct a Euclidean embedding for this neighborhood, by locally applying multidimensional scaling (Gram matrix computation from distances and spectral decomposition). Specifically, if M^{v_i} is the matrix of squared geodesic distances between the neighbors of v_i, then the Gram matrix, G^{v_i} of the underlying Euclidean representation of the

neighborhood, having v_i at the origin, reads

$$G^{v_i}|_{v_j,v_k} = 1/2(\, M^{v_i}|_{v_i,v_j} + M^{v_i}|_{v_k,v_i} - M^{v_i}|_{v_j,v_k})$$

where v_j, v_k are neighbors of v_i. Then, the eigen-decomposition $G^{v_i} = V^{v_i} S^{v_i} (V^{v_i})^T$, recovers a best approximating N-dimensional Euclidean embedding of the neighborhood, centered at v_i, as

$$E^{v_i} = V^{v_i} (S^{v_i})^{1/2} \mathbf{1}_N \tag{1}$$

where $\mathbf{1}_N$ denotes a (# of neighbors) - by - N identity matrix, selecting the largest N eigenvalues of the Gram matrix. Actually, E^{v_i} is a matrix whose rows correspond to the vertices of the neighbors of v_i and its rows to the embedding coordinates, representing the geodesics from v_i to its neighbors as straight lines in $\mathbb{R}^N$. However, geodesics between the neighbors of v_i are not necessarily straightened, since this straightening is controlled by these vertices' neighborhoods.

For this reason, E^{v_i} is not computed on the basis of the subgraph $\Gamma(v_i)$ of the immediate neighbors of v_i but on the subgraph $\Gamma^2(v_i)$ of the second order neighborhood of v_i. With such an E^{v_i}, except of its rows that straighten the paths between v_i and the vertices of $\Gamma^2(v_i)$, rows' differences $E^{v_i}|_{v_k} - E^{v_i}|_{v_j}$, if $v_j, v_k \in \Gamma(v_i)$, also straighten geodesic paths between these vertices. However, computation of E^{v_i} for the vertices of $\Gamma^2(v_i)$ requires valid geodesic distances between them, which in turn, require the subgraph of all neighborhoods of the vertices of $\Gamma^2(v_i)$ that is actually the subgraph $\Gamma^3(v_i)$ of the third order neighborhood of v_i.

Consequently, in short, for each vertex v_i and for a constant local embedding dimension:

1. The subgraphs $\Gamma(v_i)$, $\Gamma^2(v_i)$ and $\Gamma^3(v_i)$ of the first, second and third order neighborhoods of v_i are correspondingly considered,
2. Geodesic distances between the vertices of $\Gamma^2(v_i)$ are computed via shortest paths in $\Gamma^3(v_i)$
3. The Gram matrix G^{v_i} is computed on the basis of these distances and its rank - N spectral decomposition offers E^{v_i} via (1)
4. Only the rows of E^{v_i} that correspond to vertices of $\Gamma(v_i)$ are maintained.

2.2 Metric-trivializing transports and functions' differentials on graphs

The geodesics' straightening frames E^{v_i} allow for representing points within the convex hull of the neighborhood of each vertex v_i, via barycentric coordinates λ^{v_i}. Namely, for any such point p its v_i - centered Euclidean representation $x^{v_i}(p)$ reads

$$x^{v_i}(p) = \sum_{v_j \in \Gamma(v_i)} \lambda^{v_i}|_{v_j}(p)\; E^{v_i}|_{v_j} \tag{2}$$

Using these coordinates for p, if p lies along the shortest paths of $\Gamma^2(v_i)$ connecting the vertices of $\Gamma(v_i)$, the infinitesimal transports $dx(p)$ correspond to transports along geodesics, thus trivializing the metric. Now, if p lies anywhere else in the convex hull of $\Gamma(v_i)$, the identity matrix is a second order approximation of the metric of $dx^{v_i}(p)$ and specifically, with an $O(\min_{v_j \in \Gamma(v_i)} |x^{v_i}(p) - x^{v_i}(v_j)|^2)$ error. Consequently, up to second order approximation, the infinitesimal transport

$$\alpha^{v_i}(p) = \sum_{v_j \in \Gamma(v_i)} d\lambda^{v_i}|_{v_j}(p)\; E^{v_i}|_{v_j} \tag{3}$$

can be thought as trivializing the metric.

If the N components of this transport were integrable, then one could recover a global Euclidean isometry of the data manifold in $\mathbb{R}^N$. In order to do so approximately, we should determine N functions $\varphi = [\varphi_i]_{i=1..N}$, whose differential $d\varphi$ best approximates the trivializing transport, up to orthogonal N-dimensional transformations. The realization of the functions' differential on a distance graph, in terms of the local coordinates of (2) is derived in appendix B, thus allowing to approximate $d\varphi$, again up to the second order, by

$$d\varphi^{v_i}(p) = \sum_{v_j \in \Gamma(v_i)} d\lambda^{v_i}|_{v_j}(p)(\varphi(v_j) - \varphi(v_i)) \tag{4}$$

2.3 The quadratic approximation error

Having realized the metric trivializing transports α and the functions' differential $d\varphi$ on distance graphs, the next step is to formulate the error of approximating α by $d\varphi$ and evaluate it on distance graphs. Specifically, let

$$\varepsilon[\varphi, Q] = d\varphi - \alpha Q \tag{5}$$

the error of approximating the trivializing transports α by functions' differentials $d\varphi$, up to orthogonal transformations Q. Actually, ε is an infinitesimal transport between a Q - aligned version of the metric - trivializing α and its integrable approximate $d\varphi$. The quadratic measure of ε is formally offered by the inner product, as defined for differential forms, within the Hodge theory (e.g. refer to the ch. 0, par. 6 of [12]). In appendix B, it is shown that using on - graph realizations (3) and (4) for α and $d\varphi$, respectively, this quadratic error reads

$$\|\varepsilon[\varphi, Q]\|^2 = \sum_{v_i} \sum_{v_j, v_k \in \Gamma(v_i)} \Lambda^{v_i}|_{v_j, v_k} \left\langle \mathcal{E}^{v_i}|_{v_j}, \mathcal{E}^{v_i}|_{v_k} \right\rangle_{\mathbb{R}^{\mathbb{N}}} \mathrm{vol}(\Gamma(v_i)) \tag{6a}$$

$$\mathcal{E}^{v_i}|_{v_j} \equiv \varphi(v_j) - \varphi(v_i) - E^{v_i}|_{v_j} Q \tag{6b}$$

$$\Lambda^{v_i} : (E^{v_i})^T \Lambda^{v_i} E^{v_i} = I_{N \times N} \tag{6c}$$

Actually, $\mathcal{E}^{v_i}$ evaluates the failure of the local metric - trivializing frames E^{v_i} to be aligned with vertex functions' differences within graph's neighborhoods, while Λ^{v_i} is the rank - N pseudo-inverse of the Gram matrix G^{v_i}, properly weighting the on-graph inner product of $\mathcal{E}^{v_i}$ with itself.

3 The embedding method

3.1 The stationarity equations of the optimal Euclidean embedding

In order to formulate and resolve the optimal Euclidean embedding problem in a general setting, the error functional $\|\varepsilon[\varphi, Q]\|^2$ is considered independently of the on - graph realizations (3) and (4) of α and $d\varphi$, within the framework of differential forms. Then, allowing φ and Q to infinitesimally vary from their error minimizing counterparts $\tilde{\varphi}$ and $\tilde{Q}$, one gets the Euler - Lagrange equations that render the error functional stationary. This derivation is made in appendix C and within the manifold to be embedded, Ω, the stationary $\tilde{\varphi}$ and $\tilde{Q}$ are given by the equations

$$\mathrm{div}(d\tilde{\varphi}) = \mathrm{div}(\alpha\tilde{Q}) \ , \ \text{ with boundary conditions } \ d\tilde{\varphi}|_{\partial\Omega} = \alpha\tilde{Q}\Big|_{\partial\Omega} \tag{7a}$$

$$\tilde{Q} = R_{\tilde{\varphi}} {L_{\tilde{\varphi}}}^T \ , \ \text{ where } (L_{\tilde{\varphi}}, \Sigma_{\tilde{\varphi}}, R_{\tilde{\varphi}}) = \mathrm{svd}(B_{\tilde{\varphi}}) \ \text{ for } B_{\tilde{\varphi}} : \langle d\tilde{\varphi}_i, \alpha_j \rangle = [B_{\tilde{\varphi}}]_{ij} d\mathrm{vol} \tag{7b}$$

To clarify how these equations lead to the determination of $\tilde{\varphi}$ and $\tilde{Q}$, we should first recall that $d\varphi$ and α contain N infinitesimal transports on Ω, arranged in $1 \times N$ matrices and that Q is the $N \times N$ orthogonal matrix that is free to properly align α with $d\varphi$, without modifying the metric - trivializing identity of α. Then, (7b) indicates that $\tilde{Q}$, which best aligns α with a given $d\varphi$ results from the left and right singular vectors of the matrix of the integrands of the inner products between all components of $d\varphi$ and α. Actually, this matrix contains the directional derivatives of the components of φ along the directions of the infinitesimal transports of α. In turn, (7a) indicates that the elements of $\tilde{\varphi}$ result from corresponding Neumann PDE problems that enforce their differentials to differ from the properly aligned α by divergence free transports. Actually, on the left hand side of (7a) one can identify the on-manifold Laplacian of the elements of $\tilde{\varphi}$, while on the right hand side there is the on-manifold divergence of the directions of the infinitesimal transports of α, transformed by the $\tilde{\varphi}$ - dependent $\tilde{Q}$.

Consequently, the Euler-Lagrange equations that govern the optimal Euclidean embedding $\tilde{\varphi}$ become coupled, with non-linear and non-explicit terms with respect to the derivatives of $\tilde{\varphi}$, introduced by the singular value decomposition of (7b). Thus, $\tilde{\varphi}$ cannot be determined in one step, since its equation lacks an explicit formula. In order to overcome this deadlock, an alternating procedure is adopted, which, at its fixed point, determines $\tilde{\varphi}$ and $\tilde{Q}$, simultaneously.

3.2 Iterative determination of the embedding

In order to break the interdependency of the approximation error minimizers $\tilde{\varphi}$ and $\tilde{Q}$, a plausible approach is to consider $\tilde{Q}$ fixed while computing $\tilde{\varphi}$ by (7a), then use this $\tilde{\varphi}$ to re-compute $\tilde{Q}$ by (7b), return to re-estimate $\tilde{\varphi}$ and so on. This iterative scheme is summarized by the following recursion,

$$\tilde{\varphi}|_{n+1} : \begin{cases} \mathrm{div}(d\ \tilde{\varphi}|_{n+1}) = \mathrm{div}\left(\alpha \tilde{Q}|_n\right) \\ (d\ \tilde{\varphi}|_{n+1})\big|_{\partial\Omega} = \left(\alpha \tilde{Q}|_n\right)\Big|_{\partial\Omega} \end{cases} \tag{8a}$$

$$\tilde{Q}|_{n+1} = R_{\tilde{\varphi}|_{n+1}} {L_{\tilde{\varphi}|_{n+1}}}^T \tag{8b}$$

where $L_{\tilde{\varphi}|_{n++1}}$ an $R_{\tilde{\varphi}|_{n+1}}$ result by the SVD of (7b), applied on $d\tilde{\varphi}|_{n+1}$.

In appendix D, it is shown that instead of the non-linear and explicitly intractable equation (7a), one can resolve the linear, Poisson-like equation (8a), recursively aligning its right-hand side with the acquired solutions.

Theorem 1. *Within an N-dimensional manifold Ω, the error functional $\|\varepsilon[\varphi, Q]\|^2 = \int_\Omega \langle \varepsilon[\varphi, Q], \varepsilon[\varphi, Q] \rangle$, with $\varepsilon[\varphi, Q]$ defined by* (5) *and with $\langle , \rangle$ denoting the inner product of differential forms, for the iterative process defined in* (8)*, after fixing the $\int_\Omega \tilde{\varphi}|_n$, $\|\varepsilon[\varphi|_n, Q|_n]\|^2$ is decreasing and converges, successive changes of $\tilde{\varphi}|_n$ vanish, these maps asymptotically satisfy the Euler-Lagrange equations* (7) *and their limit set consists of maps $\tilde{\varphi}_\infty : \|d\tilde{\varphi}_\infty\|$ is constant. Additionally, if the bounded $d\tilde{\varphi}|_n$ is also compact and almost everywhere of full-rank, every accumulation point of $(\tilde{\varphi}_n, \tilde{Q}_n)$ belongs to the invariant set $[\tilde{\varphi}_\infty, \tilde{Q}_\infty]$ of the Euler-Lagrange critical points of $\|\varepsilon[\varphi, Q]\|^2$.*

4 Implementation of the method on distance graphs

While α and $d\varphi$ have explicit, on-graph realizations, given by (3) and (4), their inner product and their divergence, lack such realizations and they are necessary for evaluating (8b) and (8a), respectively. These realizations are given in the following Sct. 4.2, thus allowing for the computational implementation of the embedding method derived in Sct. 3.2, which is given in Sct. 4.3

4.1 Infinitesimal transports

Formula (3), derived for the metric-trivilizing transports actually sets the prototype for representing any infinitesimal transport, on distance graphs. So, let ζ be such a transport, then for any point p in the neighborhood $\Gamma(v_i)$ of a certain vertex v_i of the graph, ζ may be written as

$$\zeta^{v_i}(p) = \sum_{v_j \in \Gamma(v_i)} d\lambda^{v_i}|_{v_j}(p)\ Z^{v_i}|_{v_j} \tag{9}$$

with λ^{v_i} induced by the local Euclidean representation (2) of $\Gamma(v_i)$. Consequently, an arbitrary infinitesimal transport ζ is represented by its edge function Z, which in turn, is computationally realized by a sparse matrix in the same filling pattern as the underlying graph's adjacency matrix.

Moreover, definition (3) for the metric trivializing transports α reveals the necessity for a computational representation of vectors of infinitesimal transports, which, on-graph, correspond to vectors of edge functions that computationally, are represented by tuples of sparse matrices, following the filling pattern of the graph's adjacency matrix. So, the local frames $E^{v_i}|_{v_j}$ that determine the on-graph realization of α, as determined in Sct. 2.1, they correspond to N edge functions (the coordinates of the Euclidean embedding of $\Gamma(v_i)$ in $\mathbb{R}^N$), which are computationally represented as a tuple, `Frm`, of N sparse matrices, `F`, in the same filling pattern as the distance graph's adjacency matrix. This representation is identified by the correspondence

$$E^{v_i}|_{v_j} \equiv \left[(\texttt{Frm}(k).\,\texttt{F})^{v_i}|_{v_j}\right]_{k=1\ldots N} \tag{10}$$

4.2 Inner product, divergence and Laplacian

The key operation, whose on-graph realization enables the implementation of the iterative embedding scheme of (8) is the inner product of the infititesimal transports . In appendix B, using the representation of (9), the on-graph realization of the inner product between two arbitrary transports ζ and η is

derived. Moreover, there is given the evaluation of the corresponding formula as a vector product of N edge functions. The constituent elements of this evaluation are a) the sparse matrices Z, H that represent ζ, η, respectively and b) the N edge functions ${E_-}^{v_i}|_{v_j} \in \mathbb{R}^N$ that result from the local frames E^{v_i} of (1) as

$$ {E_-}^{v_i}|_{v_j} = \sqrt{\mathrm{vol}(\Gamma(v_i))} {V}^{v_i}|_{v_j} (S^{v_i})^{-1/2} \mathbf{1}_N \tag{11} $$

where, again, $\mathbf{1}_N$ denotes a (# of neighbors) - by - N identity matrix, selecting the largest N eigenvalues of S^{v_i} (please refer to (1)) and $\mathrm{vol}(\Gamma(v_i))$ denotes the volume of the N-dimensional convex hull of $\Gamma(v_i)$. In correspondence with the computational representation (10) of $E^{v_i}|_{v_j}$, ${E_-}^{v_i}|_{v_j}$ is represented as a tuple, `FrmInv`, of N edge functions `F`, via the correspondence

$$ {E_-}^{v_i}|_{v_j} \equiv \left[(\texttt{FrmInv}(k).\,\texttt{F})^{v_i}|_{v_j} \right]_{k=1\ldots N} \tag{12} $$

On the basis of these data, the inner product of the arbitrary infinitesimal transports ζ and η is implemented as

$$ \langle \eta, \zeta \rangle = \sum_{k=1\ldots N} (\texttt{sum}\,(\texttt{FrmInv}(k).\,\texttt{F} \odot H, \texttt{dim} = 2) \odot (\texttt{sum}\,(\texttt{FrmInv}(k).\,\texttt{F} \odot Z, \texttt{dim} = 2) \tag{13} $$

where $\odot$ denotes the element-wise product of matrices and `sum` denotes the summation of matrix elements, along the dimension `dim`.

Using the on-graph realization of the inner product and (4) for the functions' differential, in appendix B, the on-graph realization of the divergence operator is determined, as the inner product - dual of the differential. The computational evaluation of the operator follows from (13) as

$$ \begin{aligned} \mathrm{div}(\zeta) = & \sum_{k=1\ldots N} \texttt{FrmInv}(k).\texttt{F}^T \texttt{sum}\,(\texttt{FrmInv}(k).\texttt{F} \odot Z, \texttt{dim} = 2) \\ & - \sum_{k=1\ldots N} \texttt{sum}\,(\texttt{FrmInv}(k).\texttt{F}, \texttt{dim} = 2) \odot \texttt{sum}\,(\texttt{FrmInv}(k).\texttt{F} \odot Z, \texttt{dim} = 2) \end{aligned} \tag{14} $$

Using (14) and substituting in ζ a function's differential, *df*, as realized on distance graphs via (4), one has $Z^{v_i}|_{v_j} = f|_{v_j} - f|_{v_i}$, thus obtaining the implementation of the Laplacian operator

$$ \begin{aligned} \mathcal{L} = & \sum_{k=1\ldots N} \texttt{FrmInv}(k).\texttt{F}^T \texttt{FrmInv}(k).\texttt{F} + \texttt{diag}((\texttt{sum}\,(\texttt{FrmInv}(k).\texttt{F}, \texttt{dim} = 2))^2) \\ & - \texttt{sum}\,(\texttt{FrmInv}(k).\texttt{F}, \texttt{dim} = 2) \odot (\texttt{FrmInv}(k).\texttt{F} + \texttt{FrmInv}(k).\texttt{F}^T) \end{aligned} \tag{15} $$

where `diag` turns its input vector to a sparse diagonal matrix and the squaring evident in its argument is supposed to act element-wise. In both, (14) and (15) the element-wise matrix operations $+, -, \odot$ are considered to act as their vectorized implementation do, namely by operating column- or row-wise if one of their operands is a column or resp. row vector.

4.3 The embedding algorithm

Input: A - the distance graph's adjacency matrix, N - the dimension of the sought representation.

Initialization: Via the process of Sct. 2.1, compute the local Euclidean frames E^{v_i} and use them to estimate the volume $\mathrm{vol}(\Gamma(v_i))$ of the corresponding neighborhoods $\Gamma(v_i)$. Then, via (11), compute ${E_-}^{v_i}$. Collect E and E_- correspondingly to the tuples `Frm` and `FrmInv`, each containing N sparse matrices following the filling pattern of A. Using `FrmInv`, compute the Laplacian matrix $\mathcal{L}$, via (15). This is a sparse matrix following the filling pattern of A^2. Use $\mathcal{L}$ to compute a preconditioner (here a thresholded incomplete Cholesky factorization (ICT)) for the iterative sparse solver employed to invert it (here the Preconditioned Conjugate Gradient (PCG)). Initialize the #-vertices $\times\, N \times N$ matrices $\tilde{Q}$, with $\tilde{Q}_{i,*,*} = I_{N\times N}$, for all $i = 1$...#-vertices and `a`, with $\texttt{a}_{*,k,m} = \texttt{sum}\,(\texttt{FrmInv}(k).\texttt{F} \odot \texttt{Frm}(m).\texttt{F}, \texttt{dim} = 2)$.

Recursion: Compute the #-vertices $\times\, N$ matrix `DaQ`, evaluating $\mathrm{div}(\alpha\tilde{Q})$, by applying (14) in the form $\texttt{DaQ}_{*,m} = \sum_{k=1}^{N} \texttt{FrmInv}(k).\texttt{F}^T \texttt{a}_{*,k,m} - \texttt{sum}\,(\texttt{FrmInv}(k).\texttt{F}, \texttt{dim} = 2) \odot \texttt{a}_{*,k,m}$. Use a preconditioned iterative sparse solver to obtain the #-vertices $\times\, N$ matrix $\tilde{\varphi}$ that results from the solutions of $\mathcal{L}\tilde{\varphi}_{*,m} = \texttt{DaQ}_{*,m}$, which evaluate (8a) on distance graphs. Use equation (13) to compute the #-vertices $\times\, N \times N$ matrix $B_{\tilde{\varphi}}$ of (7b) in the form $[B_{\tilde{\varphi}}]_{*,m,n} =$

Table 1: Evaluation of synthetic datasets' representations' quality

	swiss roll					helix				
	Lcl-mtrc ↓	Lcl-F1 ↑	Glbl-Corr ↑	Glbl-mtrc ↓	Glbl-prm ↓	Lcl-mtrc ↓	Lcl-F1 ↑	Glbl-Corr ↑	Glbl-mtrc ↓	Glbl-prm ↓
LLE	0.59	0.49	0.76 / 0.76 / 0.57	1.00	0.06	0.59	0.53	0.45 / 0.49 / 0.35	0.99	0.49
HLLE	0.59	0.69	0.81 / 0.81 / 0.63	0.92	2e-3	0.57	0.94	0.96 / 0.99 / 0.92	0.79	0.01
LTSA	0.59	0.69	0.81 / 0.80 / 0.63	1.00	2e-3	0.59	0.94	0.96 / 0.99 / 0.92	0.99	0.01
LplcMps	0.59	0.43	0.85 / 0.82 / 0.65	1.00	0.89	0.59	0.92	0.98 / 1.00 / 0.98	1.00	5e-4
DffMps	0.58	0.46	0.27 / 0.29 / 0.22	0.93	0.09	1.50	0.78	0.86 / 0.86 / 0.68	0.80	0.11
IsoMap	1.19	0.92	1.00 / 1.00 / 0.99	0.73	4e-3	1.37	0.96	0.98 / 1.00 / 1.00	0.87	9e-5
proposed	0.31	0.94	1.00 / 1.00 / 0.98	0.01	3e-3	0.67	0.89	0.91 / 0.91 / 0.78	0.51	0.04

$\sum_{k=1}^{N}\left(\texttt{FrmInv}(k).\texttt{F}^T\tilde{\varphi}_{*,m} - \texttt{sum}\left(\texttt{FrmInv}(k).\texttt{F}, \texttt{dim} = 2\right) \odot \tilde{\varphi}_{*,m}\right) \odot \texttt{a}_{*,k,n}$. Apply SVD to each $N \times N$ matrix $[B_{\tilde{\varphi}}]_{i,*,*}$ and obtain via (7b) the corresponding elements $Q_{i,*,*}$ of the candidate update Q of the matrix $\tilde{Q}$. Compute the rigid-body rotation $Q_0 = R_0 {L_0}^T$, resulting from the $(L_0, S_0, R_0) = \text{svd}(\texttt{sum}\left(B_{\tilde{\varphi}}, \texttt{dim} = 1\right))$ and absorb it, setting $Q \leftarrow {Q_0}^T Q$. Evaluate the convergence error as $\texttt{err} = \underset{i}{\text{mean}}\left(\|\tilde{Q}^T_{i,*,*}Q_{i,*,*} - I\|_{2,2}{}^2\right)^{1/2}$ and break the recursion if $\texttt{err}$ falls below a predefined precision $\texttt{tol}$. Otherwise, update $\tilde{Q}_{i,*,*} \leftarrow \tilde{Q}_{i,*,*}Q_{i,*,*}$, $\texttt{a}_{i,*,*} \leftarrow \texttt{a}_{i,*,*}Q_{i,*,*}$ and iterate.

Output: $\tilde{\varphi}$ - the embedding of the graph's vertices, $\tilde{Q}$ - the orthogonal alignment matrices

Complexity: *Memory complexity* is set by the non-zero elements of the Laplacian (15), which is $O(\#\text{ - vertices})$. The computation of $\text{vol}(\Gamma(v_i))$ could dominate this, if convex hull computations were held for large N. So, for $N \geq 7$, Kubota's projection formula is applied to 7-dimensional orthogonal projections of $\Gamma(v_i)$. *Time complexity* of the recursion is set by the preconditioned iterative solution of the the sparse system that determines $\tilde{\varphi}$, which is $O((\#\text{ - vertices})^{1+\epsilon})$, with $\epsilon \in [0, 1/N)$, controlled by the preconditioner.

5 Experimental evaluation of the methodology

The experimental setup starts by determining a proper distance graph, common for the proposed and the baseline methods. This is driven by the determination of each dataset's intrinsic dimension, N, which is estimated by the method of [13], on the basis of the two nearest neighbors' distances' ratio. Then, the N - nearest neighbors' graph is determined and its adjacency matrix A is rendered symmetric by $A \leftarrow \max(A, A^T)$, so that it corresponds to a distance graph. Concerning the convergence tolerance, employed in the performed experiments, for the proposed embedding algorithm, we have left the algorithm fully converge to the single precision machine epsilon, 10^{-7} and to the double precision one, 10^{-16}, for the convergence error sequential variation.

The proposed embedding method has been tested against two requirements : a) local and global geometric fidelity and b) maintenance or enhancement of the data discrimination.

The metrics selected to evaluate a) are for the *local characteristics*: **Lcl-dist**: local distances' absolute relative error, **Lcl-ang**: local frames' angles' absolute error in degrees, **Lcl-mtrc**: strain of the local metric reconstruction, **Lcl-prec**, **Lcl-rec** and **Lcl-F1**: neighborhoods' precision, recall and F1 score, while for the *global characteristics*: **Glbl-mtrc**: stress of the reconstructed geodesic distances, **Glbl-corr**: Pearson, Spearman and Kendall correlation between the Euclidean and the geodesic distances, **Glbl-prm**: procrustes error of the affine alignment between the determined embedding and the underlying manifold's parameterization (only for synthetic datasets). On the other hand, to evaluate the capability of the embedding to discriminate data clusters, for each data point of a labeled dataset, we consider its N - nearest neighbors, with respect to the tested embedding. Then, the label of each neighbor is weighted by the logistic function of its Euclidean distance from the query point. The query point obtains the label of the maximum average weight. The data points' grouping is measured by **ACC**: its accuracy, **NMI**: its Normalized Mutual Information with the ground truth and **ARI**: its Adjusted Rand Index.

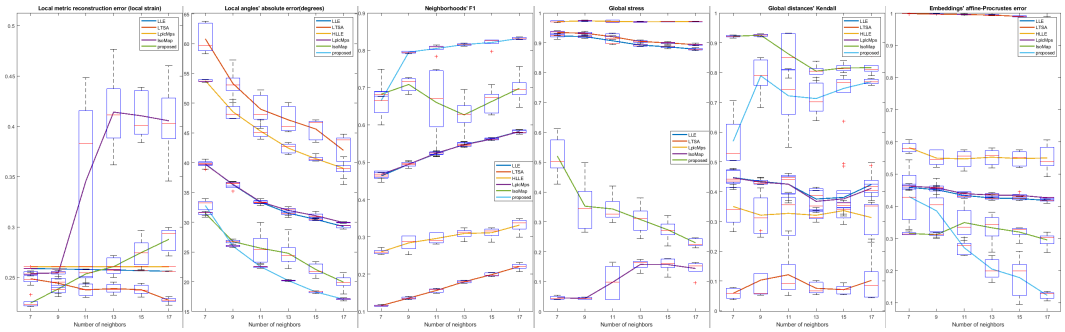

Figure 1: "difficult" manifold's unfolding experiment ($\mathbb{R}^{10} \rightarrow \mathbb{R}^5$) for varying neighborhood size. The measures' box plots depict the maximum/minimum boundaries at whiskers, the boxes of the populations' 25th-75th percantile and the median values in red. The plotted lines depict the variation of the metrics' mean values with respect to the neighborhood size.

The embedding method has been evaluated on both synthetic and real data and compared to representative embedding methods. Specifically, the dimensionality reduction toolbox of [14] is employed, using its implementations of local embedding methods, with the exception of UMAP implementation [15]. Namely, additionally to UMAP and t-SNE [9] from the class of stochastic local methods, the proposed method is compared to IsoMap [1], Diffusion Maps (DffMps) [16] and their hybrid PHATE [17] (implemented in [18]), which are global methods, expected to perform better than the proposed one in reproducing the global manifold structure and LLE [4], LTSA [6], Laplacian Eigenmaps (LplcMps) [7] and Hessian Eigenmaps (LplcMps) [8], which are local methods expected to distort global manifold structure. For fairness, the local methods' implementations have been properly modified so as to operate on graphs rather than features vectors, following the adaptation proposed in [5].

The considered *synthetic datasets* are: the swiss roll ($\mathbb{R}^3 \rightarrow \mathbb{R}^2$), the "difficult" dataset of [14] ($\mathbb{R}^{10} \rightarrow \mathbb{R}^5$) and the helix ($\mathbb{R}^3 \rightarrow \mathbb{R}^2$) , while the considered *real datasets* are the test sets of MNIST and FMNIST images' datasets ($\mathbb{R}^{784} \rightarrow \mathbb{R}^{13}$) and an RNA-seq data set, related to lung cancer and collected from [19] (1625 cases with 51 different diagnoses, $\mathbb{R}^{60660} \rightarrow \mathbb{R}^{30}$ embedding).

5.1 Synthetic datasets' results

A first comparison of the proposed method with the baselines, against the representation of low dimensional manifolds is summarized in Table 1. The results indicate that in the "swiss roll" 2D embedding the proposed is the only local method approximating the performance of IsoMap, in terms of global consistency, while retaining the local geometric data and the neighboring relations. The case of "helix" is somewhat different since it is locally an 1-dimensional manifold, however at least 2 dimensions are necessary to isometrically embed it. Consequently, the rank-2 local metric estimates are ill-conditioned and the local methods HLLE, LTSA and Laplacian Eigenmaps fail when they operate on the minimal graph that offers such metric estimates (it was the 5NN graph). This graph suffices for the proposed method and for LLE to estimate an embedding, however, for the other local embedding methods, we have increased the number of neighbors, up to the one that lifts failure.

Actually, the second class of experiments performed on synthetic data sets evaluates the embedding methods' sensitivity to the number of nearest neighbors. Specifically, starting from the minimal graph that returned local metric of full rank estimates, we have considered 5 consecutive increases of the neighborhood size by 2 neighbors. At each neighborhood size we have performed 15 independent random samplings, each of 1500 points. For the "difficult" data set, the variation of the embedding quality is summarized in the box-plots of figure 1 . The proposed method is the only local one, approximating IsoMap in the metrics of global consistency (Global stress that IsoMap minimizes and Kendall correlation between the Euclidean and the original geodesic distances), while retaining local consistency of neighborhoods' representation. A remark should be made on the preservation of the local metric; increase in the number of neighbors results to an increase of the discrepancy between the local distances and the corresponding actual geodesic distances and this adversarial relation is evident in the first plot of figure 1 . The proposed method and IsoMap, being methods that approximately

Table 2: Evaluation of the clusters' representation for the RNA-seq data set, related to lung cancer

	Diagnosis classification within embeddins' neighborhoods							
	LLE	LTSA	LplcMps	IsoMap	PHATE	t-SNE	UMAP	proposed
ACC ↑	0.6954	0.6640	0.6985	0.6825	0.6855	0.6917	0.7040	0.7126
NMI ↑	0.2934	0.2660	0.3006	0.2811	0.2855	0.3048	0.3213	0.3298
ARI ↑	0.3356	0.2525	0.3525	0.3020	0.3181	0.3461	0.3742	0.3821
	Local/global geometric consistency							
Lcl-F1 ↑	0.5258	0.3099	0.5410	0.5822	0.5838	0.6788	0.7160	0.7698
Lcl-mtrc ↓	0.0756	0.0752	0.0761	0.1087	0.0884	3.2433	0.1788	0.0566
Glbl-mtrc ↓	0.9428	0.9383	0.9903	0.1519	0.6291	34.28	3.1886	0.1589

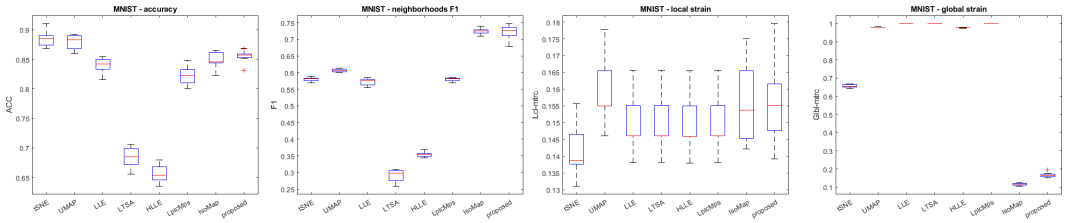

Figure 2: MNIST dataset 10-fold experiment. The measures' box plots depict the maximum/minimum boundaries at whiskers, the boxes of the populations' 25th-75th percantile and the median values in red. Possible outliers are marked by red crosses. The proposed method performs practically the same as Isomap, either in terms of clustering or local/global geometric consistency. The stochastic methods slightly outperform their clustering accuracy (probably outliers within neighborhoods)

preserve geodesic distances, both distort local metric as the neighborhood size increases, though this decay is suppressed in the proposed method, compared to the step-size distortion of IsoMap.

5.2 Real datasets' results

The most characteristic case, where embeddings' loca/global geometric consistency affects data clustering is the RNA-seq dataset (table 2). There, the 51 different diagnoses' classes are highly imbalanced and not fully separated by global distances. The consistency between the embeddings' neighborhoods and the original ones (F1 score) and the preservation of local distances are the key elements for reasoning the embeddings' clustering capability. One may observe in table 2 that for methods retaining local distance data, plausibly, neighborhoods' consistency is the key for representing data clustering. IsoMap distorts local consistency by collapsing neighboring geodesics, thus being outperformed. PHATE has a closely related performance, however not retaining local distances, due to the substitution of geodesics by diffusion paths. t-SNE and UMAP, by design, do not preserve distances, however they do it so as to amplify separation of dense local structures. Thus, these methods' capability to preserve neighborhoods dominates the distortion of local distances. However, UMAP partially retains these distances and outperforms t-SNE. The proposed method's embedding succeeds in best representing, local structures thus achieving best clusters' representation, while it's global geometric consistency is practically the one of IsoMap.

The MINST and FMNIST datasets' test sets were used to test the stability of the embedding methods under sparse random sampling within high dimensional spaces. Specifically, the 10k data points of these datasets were randomly split in 10 disjoint groups and the embedding methods have been evaluated on these groups. The MNIST results are summarized in Fig. 2 and the ones of FMNIST are included in appendix E.6. The FMNIST results practically follow the ones of RNA-seq dataset, indicating that the preservation of the neighborhoods' structure (F1 score) sets the baseline for the embeddings' capability to represent data clusters. This is not that clear-cut in the MNIST results. While, the proposed method and IsoMap are top-scored both in terms of neighborhood preservation (F1 score) and in terms of global distances' reconstruction (global strain), they are outperformed by the density based methods, t-SNE and UMAP, in the accuracy of data clusters' representation. In this case, the clustering capability of the original graph's neighborhoods sets the barrier for the performance of the proposed method and IsoMap that prserve and arange them according to the

original graph structure. On the other hand, t-SNE and UMAP modify the neighborhoods and arrange them according to the homogeneity of the data points' density, which represents clustering better than the actual datapoints' distances.

6 Limitations and extensions

The fundamental restriction that the proposed approach imposes to the data embedding problem is that any sought data - embedding correspondence should be encoded in terms of a symmetric relation between datapoints that are considered neighboring according to his relation. This setting fundamentally excludes requirements in connection with relations between non-neighboring datapoints (e.g. negative sampling, Barnes–Hut approximation) and even worse, requirements about the global shape of data. Such requirements are crucial for algorithms like UMAP [11] and t-SNE [9] that are highly efficient in representing data clustering. In terms of the proposed approach, the restriction to local-only relations is a cost paid in favor of the variational formulation of the embedding problem and the requirement that the method does not rely on dense pairwise data. Namely, incorporation of distant relations in a variational formulation would lead to covariance-like operators that are densely represented on data, thus violating the sought sparsity of the processed data.

The consistency between the operators included in the variational formulation of the embedding problem and the locality of the distance graph's edges also imposes the main technical challenge to the implementation of the optimal embedding method of Sct. 3. Though, the iterative scheme of (7) is theoretically guaranteed to approximate locally optimal embeddings, the on-graph realization of the involved differential operators is up to a second order approximation. Thus the presence of outliers in the neighborhoods of the distance graph, though it does not affect the convergence, it may result to poor operators' estimates, thus affecting the quality of the embedding.

Concerning the proposed method's computational profile, its memory demands remain linear in the number of datapoints, K, in accordance with all local embedding methods. However, its time complexity, per iteration, is between $O(K)$ and $O(K)^{1+1/\text{dimensions}}$, while UMAP [11] and LargeVis [10] iterations' time demands are $O(K)$. Shifting to an algebraic multigrid (AMG) preconditioner for the Laplacian inversion (Sct. 4.3) guarantees $O(K)$, however imposing challenges in the convergence of the iterative scheme, due to the non-stationarity of the AMG preconditioning, across iterations.

Finally, the proposed on-graph implementation of the optimal embedding scheme of (7), though general in terms of data diversity and sparse in terms of data volume, it necessarily remains recursive and non-parametric. As a consequence, the embedding of new data points, requires their incorporation in the distance graph and usage of Nyström approximation to extend the embedding. To restore optimality of these extensions, we should use them to extend the alignment matrices $Q|_\infty$ and then repeat the iterative embedding scheme updating only the new points. As it becomes evident, such an inference is computationally expensive, fetching the whole training data for each out of sample extension.

On the other hand, parametric embedding schemes that provide such a functional form, require vector representations for the processed data, thus lacking the representation-free feature that allows for operating heterogeneous data. Autoencoders [20], [21], is the source point of the parametric, non-linear data embedding. The objectives optimized by the learned models focus on reconstructing the data from a mapping (encoder) that embeds them in a low-dimensional vector space. The coupling between the embedding and the reconstruction (decoder) expands the class of the optimal embeddings to the whole class of diffeomorphisms, having only regularization of the embedding to impose restrictions within the class. Thus, local (differential) or global (integral) data - embedding correspondences cannot be imposed in the Autoencoders' objective function. Shifting to deep architectures (convolutional [22] or attention-based [23]) partially resolves this, if the datapoints' graph is fixed and gridded. To cover the fully intrinsic case of irregular data graphs, one should shift to the graph deep learning framework (e.g. Graph Attention [24]), however not avoiding the $O(\log(K))$ graph update step per inference.

The proposed approach is not inherently bound to non-parametric estimation, since it fundamentally relies on the functional equation of the optimal embedding. The operators involved in these equations can be analytically evaluated on parametric models, thus enabling a possible reduction of the original operator-learning problem to a mapping-learning one. This is a technically involved extension, both in terms of theory and implementation and it is the major direction of our on-going research effort.

Acknowledgments and Disclosure of Funding

For the experimental evaluation, the shared computational resources of the IDEAL lab of the University of the Aegean have been used.

The work has not received any third-party funding and no financial relationships outside the submitted work have influenced it.

References

[1] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[2] W. S. Torgerson. *Theory and Methods of Scaling*. John Wiley and Sons, New York, 1958.

[3] Akshay Agrawal, Alnur Ali, and Stephen Boyd. Minimum-distortion embedding. *Foundations and Trends in Machine Learning*, 14(3):211–378, 2021.

[4] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[5] L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[6] Zhang Zhenyue and Zha Hongyuan. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004.

[7] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[8] David L. Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[9] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9: 2576–2605, 2008.

[10] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *WWW '16: 25th International Conference on World Wide Web*, pages 287–297, 2016.

[11] L. McInnes, J. Healy, N. Saul, and J. Melville. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 29(3):861, 2018.

[12] P. Griffiths and J. Harris. *Principles of Algebraic Geometry*. John Wiley and Sons, New York, 1978.

[13] Elena Facc, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7, 2017.

[14] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005, Tilburg University, 2009.

[15] Connor Meehan, Jonathan Ebrahimianand Stephen Meehanand, and Wayne Moore. Uniform manifold approximation and projection (umap), 2025. URL `https://www.mathworks.com/matlabcentral/fileexchange/71902-uniform-manifold-approximation-and-projection-umap`.

[16] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[17] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37:1482–1492, 2019.

[18] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Phate - visualizing transitions and structure for biological data exploration, 2025. URL `https://github.com/KrishnaswamyLab/PHATE/tree/main`.

[19] National Cancer Institute. Gdc data portal, 2025. URL https://portal.gdc.cancer.gov/. Accessed: 2025-12-17.

[20] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[21] Diederik P Kingma and MaxWelling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[22] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(1), 2023.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, 2017.

[24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[25] Francesco Bullo, Jorge Cortés, and Sonia Martínez. *Distributed Control of Robotic Networks: A Mathematical Approach to Motion Coordination Algorithms*. Princeton Series in Applied Mathematics. Princeton University Press, 2009.

[26] The MathWorks, Inc. Matlab version: 25.1.0 (r2025b), 2025. URL https://www.mathworks.com.

[27] The MathWorks, Inc. Parallel computing toolbox version 23.2 (r2023b), 2025. URL https://www.mathworks.com.

[28] F. Denti, D. Doimo, A. Laio, and A. Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005, 2022.

A Inner product of differential 1-forms and their divergence

Concerning the notation adopted for representing vectors, matrices and tensors and their contractions, Einstein summation convention is used and in the matrix form version of the expressions, the vectors are treated as columns.

An 1-form η on an N-dimensional manifold Ω can always be written as a linear combination of the manifold's local coordinates' differentials, e.g. for an N-dimensional manifold with $dx = [dx_i]_{i=1...N}$, η reads $\eta = dx_i h^i = dx^T h$, where $\eta = [h_i]_{i=1...N}$ is the tangent vector of the corresponding infinitesimal transport, with respect to the selected local coordinates.

Hodge $\star$ operator, allows for taking the space of transports orthogonal to the ones included in the operated differential form. Acting on the 1-form η, the operator returns an $(N-1)$-form, $\star\eta = \star dx_i h^i$, having η as normal vector. This is achieved if for each dx_i, $\star dx_i$ corresponds to differential transports normal to dx_i, which is equivalent to asking that for each dx_j, the exterior (wedge) product $dx_j \wedge \star dx_i$ equals the interior product of dx_j and dx_i. If such an identification is achieved, then, using the Hodge operator, one can determine the inner product of differential forms, point-wise as $\langle \eta_1, \eta_2 \rangle = \eta_1 \wedge \star\eta_2 = \eta_2 \wedge \star\eta_1$, which, in turn allows for identifying the squared L^2 norm of a differential form as $\langle \eta, \eta \rangle$

As a $(N-1)$-form, $\star dx_i$ can be expressed as a linear combination of the differential $(N-1)$-forms $[dx_{k^*}]_{k=1...N}$, where $dx_{k^*} : dx_k \wedge dx_{k^*} = dx_1 \wedge ... \wedge dx_N$. So, let $\star dx_i = dx_{k^*}\gamma_i^k$ for a proper matrix of coefficients γ. Then, $dx_j \wedge \star dx_i = \gamma_i^j (dx_1 \wedge ... \wedge dx_N)$ should equal the interior product of dx_j and dx_i. By the definition of the Riemann metric, let it be g, the interior product of the manifold's local coordinates' differentials reads $\langle dx_i, dx_j \rangle = (g^{-1})_i^j d\text{vol}$, while their exterior (wedge) product reads $dx_1 \wedge ... \wedge dx_N = d\text{vol}\sqrt{\det g^{-1}}$, with $d$vol denoting the volume form of the manifold. Consequently, the coefficients' matrix γ reads $\gamma = g^{-1}\sqrt{\det g}$, thus allowing us to write

$$\star\eta = \sqrt{\det g}\; h_i (g^{-1})_j^i dx_{j^*} \tag{16}$$

Having defined the $\star$ operator that allows to evaluate the inner product of differential 1-forms point-wise, integration of this evaluation over the whole manifold Ω determines the inner product as a functional, in the form

$$(\eta_1, \eta_2) \equiv \int_\Omega \langle \eta_1, \eta_2 \rangle \equiv \int_\Omega \eta_1 \wedge \star\eta_2 = \int_\Omega {h_1}^T g^{-1} h_2 d\text{vol} \tag{17}$$

In turn, this functional form of the inner product and the generalized Stokes' theorem allow for determining the dual action on the 1-forms of the differential of scalar functions (0-forms). Namely, if we consider the inner product of such a differential df with an 1-form η, Stokes theorem allows us to write $(df, \eta) \equiv \int_\Omega df \wedge \star\eta = \int_{\partial\Omega} f \star \eta - \int_\Omega f d \star \eta = \int_\Omega f(d\chi_\Omega \wedge \star\eta - d \star \eta) = (f, d\chi_\Omega \wedge \star\eta - d \star \eta)$, where χ_Ω stands for the characteristic function of Ω. Then the sought duality, asks for determining an operator div, acting on differential 1-forms, such that $(df, \eta) = (f, \mathrm{div}(\eta))$. Consequently, the divergence operator div that satisfies this duality reads

$$\mathrm{div}(\eta) = d\chi_\Omega \wedge \star\eta - d \star \eta \tag{18}$$

Since $d\chi_\Omega$ results to the Dirac evaluator of $\star\eta$, (18) determines the divergence of η as the co-differential operator $d \star \eta$ in Ω and on $\partial\Omega$, as the inner product $\eta|_{\partial\Omega}$ of η with the normal of $\partial\Omega$.

Then, equations of the form $\mathrm{div}(\eta) = \rho d\mathrm{vol}$ correspond to continuity equations with Neumann boundary conditions, thus allowing to simultaneously represent the equation and its boundary conditions with a single operator.

B Representation of differential 1-forms, their inner product and their divergence on distance graphs

Since differential 1-forms correspond to the integrands of line integrals and the graph's fundamenal line elements are its edges, the intrinsic basis of differential 1-forms should be associated with the lines underlying graph edges. Correspondingly, the vector fields to be integrated should be realized on-graph as edge functions. As a consequence, a generic differential 1-form ζ can be expressed in the neighborhood of a vertex v_i of a graph as

$$\zeta^{v_i} = \sum_{v_j \in \Gamma(v_i)} Z^{v_i}|_{v_j} e^{v_i}|_{v_j} \tag{19}$$

where $e^{v_i}|_{v_j}$ is the 1-form element along the edge $v_i \to v_j$, $Z^{v_i}|_{v_j}$ the corresponding component of the Z^{v_i} vector field and $\Gamma(v_i)$ the neighborhood of v_i.

In order to compare differential 1-forms, it is necessary to evaluate their inner product, which also allows for defining a quadratic error functional that measures this comparison. For any two differential 1-forms, ζ, η, their point-wise inner product in $\Gamma(v_i)$ is determined by the inner product of the 1-form elements

$$\langle e^{v_i}|_{v_j}, e^{v_i}|_{v_k} \rangle = g^{v_i}|_{v_j, v_k} d\mathrm{vol} \tag{20}$$

where $d\mathrm{vol}$ stands for the infinitesimal intrinsic volume and $g^{v_i}|_{v_j,v_k}$ is the Gram matrix of the tangents of the transports along the edges that source from v_i.

Considering the v_i - centered $\mathbb{R}^N$ coordinates of (2), the 1-forms of their differential, $\alpha^{v_i}(p)$, have inner products, whose corresponding Gram matrix is trivial along the geodesic paths that connect the vertices of $\Gamma(v_i)$. For any other p, within the convex hull of $\Gamma(v_i)$, if $v_*(p)$ is the geodesic projection of p to its closest edge, the Gram matrix of $\alpha^{v_i}(p)$ differs from the (trivial) one of $\alpha^{v_i}(v_*(p))$ by

$$\langle \alpha^{v_i}(p), (\alpha^{v_i}(p))^T \rangle = I_{N\times N} d\mathrm{vol}(v_*(p)) + \langle \alpha^{v_i}(v_*(p)), \epsilon(p)^T \rangle + \langle \epsilon(p), (\alpha^{v_i}(v_*(p)))^T \rangle + \langle \epsilon(p), \epsilon(p)^T \rangle \tag{21}$$

where $\epsilon(p) = \alpha^{v_i}(p) - \alpha^{v_i}(v_*(p))$. Since, $v_*(p)$ is the geodesic projection of p, the transports of $\epsilon(p)$ are orthogonal to the ones of $\alpha^{v_i}(v_*(p))$ at $v_*(p)$ and thus, $\langle \alpha^{v_i}(v_*(p)), \epsilon(p)^T \rangle$ is a second order term, varying from zero only due to extrinsic rotation of the transports of $\alpha^{v_i}(v_*(p))$, which is intrinsically measured by the curvature of the underlying manifold. Consequently,

$$\langle \alpha^{v_i}(p), (\alpha^{v_i}(p))^T \rangle = (I_{N\times N} + O(|x^{v_i}(p)|^2)) d\mathrm{vol}(v_*(p)) = (I_{N\times N} + O(\|G^{v_i}\|_2)) d\mathrm{vol}(v_*(p)) \tag{22}$$

where G^{v_i} is the Gram matrix of $\Gamma(v_i)$, computed as described in Sct. 2.1.

By substituting (3) to (22), one obtains that, up to a 2nd order approximation, the Gram matrix Λ^{v_i} : $\langle d\lambda^{v_i}|_{v_j}, d\lambda^{v_i}|_{v_k} \rangle = \Lambda^{v_i}|_{v_j,v_k} d\mathrm{vol}$ should satisfy

$$(E^{v_i})^T \Lambda^{v_i} E^{v_i} = I_{N\times N} \tag{23}$$

namely being the rank - N pseudo-inverse of G^{v_i}.

With this on - graph realization of $\langle d\lambda^{v_i}|_{v_j}, d\lambda^{v_i}|_{v_k} \rangle$ in hand and letting the on-edge 1-form elements $e^{v_i}|_{v_j}$ of (19) be identified with $d\lambda^{v_i}|_{v_j}$, the inner product of two arbitrary differential 1-forms ζ, η is the integral of their point-wise product over the whole manifold, Ω, thus reading

$$(\eta, \zeta) \equiv \int_\Omega \langle \eta, \zeta \rangle = \sum_{v_i} \sum_{v_j, v_k \in \Gamma(v_i)} \Lambda^{v_i}|_{v_j,v_k} \int_{|\Gamma(v_i)|} H^{v_i}|_{v_j} \, Z^{v_i}|_{v_k} \, d\mathrm{vol} \tag{24}$$

where $|\Gamma(v_i)|$ denotes the geometric realization of the N-dimensional convex hull of $\Gamma(v_i)$. As for the edge function H, it is for the representation (19) of η the analogue of the edge function Z of ζ.

This realization of the inner product allows for defining a quadratic error measuring the N discrepancies ε of (5) on the whole manifold Ω as

$$\|\varepsilon\|^2 \equiv \sum_{i=1}^{N} (\varepsilon_i, \varepsilon_i) \tag{25}$$

Then, letting the N edge functions $\mathcal{E}^{v_i}|_{v_j} \in \mathbb{R}^N$, induced by the representation $\varepsilon^{v_i} = \sum_{v_j \in \Gamma(v_i)} \mathcal{E}^{v_i}|_{v_j} d\lambda^{v_i}|_{v_j}$, enter in (24) , in place of H and Z, one gets (6) as the on-graph evaluation of (25). We should clarify here that constancy of $\mathcal{E}^{v_i}|_{v_j}$ within $\Gamma(v_i)$ holds identically for the coefficients $E^{v_i}|_{v_j}$ of α^{v_i} and up to a second order approximation for the coefficients $(\varphi(v_j) - \varphi(v_i))$ of $d\varphi^{v_i}$, a fact that will become evident later on by determining the on-graph realization (27) of the scalar functions' differential.

Namely, following (19), the differential df of a scalar function f may be written as

$$df = \sum_{v_j \in \Gamma(v_i)} \partial_{v_j} f^{v_i} d\lambda^{v_i}|_{v_j} \tag{26}$$

Then, at each point $p \in |\Gamma(v_i)|$, we can consider its v_i - centered $\mathbb{R}^N$ representation of (2) and evaluate $df(p)$ as $df(p) = \sum_{v_j \in \Gamma(v_i)} \nabla_x f|_p \cdot E^{v_i}|_{v_j} d\lambda^{v_i}|_{v_j}$. Consequently, $\partial_{v_j} f(p) = \nabla_x f|_p \cdot E^{v_i}|_{v_j} = f(v_j) - f(v_i) + O(\|G^{v_i}\|_2)$. By substituting $\partial_{v_j} f(p)$ in (26), we get

$$df(p) = \sum_{v_j \in \Gamma(v_i)} (f(v_j) - f(v_i)) d\lambda^{v_i}|_{v_j}(p) + o(\|G^{v_i}\|_2) \tag{27}$$

It should be clarified here that the remainder $\sum_{v_j \in \Gamma(v_i)} O(\|G^{v_i}\|_2) d\lambda^{v_i}|_{v_j}(p)$ is $o(\|G^{v_i}\|_2)$ because it vanishes if $O(\|G^{v_i}\|_2)$ is constant in $\Gamma(v_i)$, due to the barycentric constraint $\sum_{v_j \in \Gamma(v_i)} \lambda^{v_i}|_{v_j}(p) = 1$. Consequently (27) verifies the second order approximation of scalar functions' differential given in (4) in the main text.

Having in avail on-graph realizations for both the inner product of 1-forms and the scalar functions' differential, the divergence of 1-forms can be evaluated as the inner product - dual of the differential, namely letting one of the 1-forms η or ζ of (24)to be exact, so that inner product offers the divergence of the other. So, let $\eta = df$, where f is an arbitrary scalar function. Then, the divergence $\mathrm{div}(\zeta)$ of ζ results from the duality $(df, \zeta) = (f, \mathrm{div}(\zeta))$, which indicates that one should use (24) in order to transfer the realization of the vertex functions' differential operator to an operator acting on edge functions. Using this on-graph realization of scalar functions' differential, its inner product with an 1-form ζ, by (24), reads up to a second order approximation

$$(df, \zeta) = \sum_{v_i} \sum_{v_j, v_k \in \Gamma(v_i)} (f(v_j) - f(v_i))\ \Lambda^{v_i}|_{v_j, v_k}\ \hat{Z}^{v_i}|_{v_k} \mathrm{vol}(\Gamma(v_i)) \tag{28}$$

where $\hat{Z}^{v_i}|_{v_k}$ stands for the mean value of $Z^{v_i}|_{v_k}$ in $|\Gamma(v_i)|$. By rearranging the summands so that one vertex of f appears in each summation term, (28) is rewritten as

$$\begin{aligned}(df, \zeta) = -\sum_{v_i} f(v_i) \Bigg(&\sum_{v_j, v_k \in \Gamma(v_i)} \Lambda^{v_i}|_{v_j, v_k}\ \hat{Z}^{v_i}|_{v_k} \mathrm{vol}(\Gamma(v_i)) \\ &- \sum_{v_j : v_i \in \Gamma(v_j)} \sum_{v_k \in \Gamma(v_j)} \Lambda^{v_j}|_{v_i, v_k}\ \hat{Z}^{v_j}|_{v_k} \mathrm{vol}(\Gamma(v_j)) \Bigg)\end{aligned} \tag{29}$$

Since, on the right-hand side of this identity, no operation is applied to f, other than the inner product, the vertex function that multiplies f is the divergence of ζ, thus reading

$$\mathrm{div}(\zeta)|_{v_i} = \sum_{v_j : v_i \in \Gamma(v_j)} \sum_{v_k \in \Gamma(v_j)} \mathrm{vol}(\Gamma(v_j))\ \Lambda^{v_j}|_{v_i, v_k}\ \hat{Z}^{v_j}|_{v_k} - \sum_{v_j, v_k \in \Gamma(v_i)} \mathrm{vol}(\Gamma(v_i))\ \Lambda^{v_i}|_{v_j, v_k}\ \hat{Z}^{v_i}|_{v_k} \tag{30}$$

Having this second order realization of the divergence operator in hand, we can determine the corresponding realization of the Laplacian operator, by letting $\zeta = df$, for a certain scalar function f. As indicated by (27), up to a second order approximation this is equivalent to substituting $\hat{Z}^{v_j}|_{v_k} = f(v_k) - f(v_k)$ thus evaluating

$$\begin{aligned}\mathrm{div}(df)|_{v_i} = &\sum_{v_j : v_i \in \Gamma(v_j)} \sum_{v_k \in \Gamma(v_j)} \mathrm{vol}(\Gamma(v_j))\ \Lambda^{v_j}|_{v_i, v_k}\ (f(v_k) - f(v_j)) \\ &- \sum_{v_j, v_k \in \Gamma(v_i)} \mathrm{vol}(\Gamma(v_i))\ \Lambda^{v_i}|_{v_j, v_k}\ (f(v_k) - f(v_i))\end{aligned} \tag{31}$$

By writing this linear combination of the values of f in matrix form, we can evaluate the Laplacian operator as a sparse matrix $\mathcal{L}$ with entries $\mathcal{L}^{v_i}|_{v_k}$, for $v_k \in \Gamma^2(v_i)$ given by

$$\mathcal{L}^{v_i}|_{v_k} = \sum_{v_j : v_i \in \Gamma(v_j)} \mathrm{vol}(\Gamma(v_j))\ \Lambda^{v_j}|_{v_i,v_k} + \delta^{v_i}|_{v_k} \sum_{v_j, v_m \in \Gamma(v_i)} \mathrm{vol}(\Gamma(v_i))\ \Lambda^{v_i}|_{v_j,v_m} - \sum_{v_j \in \Gamma(v_k)} \mathrm{vol}(\Gamma(v_k))\ \Lambda^{v_k}|_{v_i,v_j} - \sum_{v_j \in \Gamma(v_i)} \mathrm{vol}(\Gamma(v_i))\ \Lambda^{v_i}|_{v_j,v_k} \quad (32)$$

where $\delta^{v_i}|_{v_k}$ stands for the sparse identity matrix, having 1 if $v_i = v_k$ and 0 otherwise.

In order to derive vectorized versions of the defining realizations (24) for the inner product, (30) for the divergence and (32) for the Laplacian, we should overcome the problem that Λ, although sparse, it has entries ordered with respect to three dimensions. However, Λ^{v_i}, as a pseudo-inverse of $E^{v_i}(E^{v_i})^T$ (refer to (23)), may be expressed as a vector product of N edge functions, namely the columns of the pseudo-inverses of the local frames E^{v_i} of (1). Thus, we can express inner product's weighting factors as

$$\Lambda^{v_i}|_{v_j,v_k}\ \mathrm{vol}(\Gamma(v_i)) = \left\langle {E_-}^{v_i}|_{v_j}\ ,\ {E_-}^{v_i}|_{v_k} \right\rangle_{\mathbb{R}^N} \quad (33)$$

$${E_-}^{v_i}|_{v_j} \equiv \sqrt{\mathrm{vol}(\Gamma(v_i))} V^{v_i}|_{v_j} (S^{v_i})^{-1/2} \mathbf{1}_N \quad (34)$$

where, again, $\mathbf{1}_N$ denotes a (# of neighbors) - by - N identity matrix, selecting the largest N eigenvalues of S^{v_i} (please refer to (1)), thus rendering $V^{v_i}(S^{v_i})^{-1/2}\mathbf{1}_N$ the rank-N pseudo-inverse of E^{v_i}.

By substituting (33) in (24), the inner product may itself be implemented as a vector product of N edge functions

$$\langle \eta, \zeta \rangle|_{v_i} = \left\langle \sum_{v_j \in \Gamma(v_i)} \hat{H}^{v_i}|_{v_j} {E_-}^{v_i}|_{v_j}\ ,\ \sum_{v_k \in \Gamma(v_i)} \hat{Z}^{v_i}|_{v_k} {E_-}^{v_i}|_{v_k} \right\rangle_{\mathbb{R}^N} \quad (35)$$

where the edge functions $\hat{H}, \hat{Z}$ are the H, Z coefficients of ζ, η evaluated at the mean value theorem - verifying points $p_*^{v_i} \in \Gamma(v_i) : \int_{|\Gamma(v_i)|} H^{v_i}|_{v_j} Z^{v_i}|_{v_k}\, d\mathrm{vol} = H^{v_i}|_{v_j}(p_*^{v_i}) Z^{v_i}|_{v_k}(p_*^{v_i}) \mathrm{vol}(\Gamma(v_i))$.

In turn, substitution of (33) in the realization (30) of the divergence, the operator takes the form of a vector product of N edge functions, thus reading

$$\mathrm{div}(\zeta)|_{v_i} = \sum_{v_j : v_i \in \Gamma(v_j)} \left\langle {E_-}^{v_j}|_{v_i}\ ,\ \sum_{v_k \in \Gamma(v_j)} {E_-}^{v_j}|_{v_k}\ \hat{Z}^{v_j}\Big|_{v_k} \right\rangle_{\mathbb{R}^N} - \left\langle \sum_{v_j \in \Gamma(v_i)} {E_-}^{v_i}|_{v_j}\ ,\ \sum_{v_k \in \Gamma(v_i)} {E_-}^{v_i}|_{v_k}\ \hat{Z}^{v_i}\Big|_{v_k} \right\rangle_{\mathbb{R}^N} \quad (36)$$

Finally, the same substitution in the realization (32) of the Laplacian operator, offer its vector product evaluation

$$\mathcal{L}^{v_i}|_{v_k} = \sum_{v_j : v_i \in \Gamma(v_j)} \langle {E_-}^{v_j}|_{v_i}\ ,\ {E_-}^{v_j}|_{v_k} \rangle_{\mathbb{R}^N} + \delta^{v_i}|_{v_k} \left\langle \sum_{v_j \in \Gamma(v_i)} {E_-}^{v_i}|_{v_j}\ ,\ \sum_{v_m \in \Gamma(v_i)} {E_-}^{v_i}|_{v_m} \right\rangle_{\mathbb{R}^N} - \left\langle {E_-}^{v_k}|_{v_i}\ ,\ \sum_{v_j \in \Gamma(v_k)} {E_-}^{v_k}|_{v_j} \right\rangle_{\mathbb{R}^N} - \left\langle {E_-}^{v_i}|_{v_k}\ ,\ \sum_{v_j \in \Gamma(v_i)} {E_-}^{v_i}|_{v_j} \right\rangle_{\mathbb{R}^N} \quad (37)$$

C The Euler - Lagrange equations of the optimal Euclidean embedding

Concerning the notation adopted for representing vectors, matrices and tensors and their contractions, Einstein summation convention is used and in the matrix form version of the expressions, the vectors are treated as columns.

Since the sought embedding in $\mathbb{R}^N$, $\varphi = [\varphi_j]_{j=1..N}$ should minimize the discrepancies of (5), $\varepsilon_j = d\varphi_j - \alpha_i Q^i_j$, the squared L^2 norm of these 1-forms is considered to measure their magnitude. Using the evaluation (17) of the inner product of differential 1-forms, the corresponding error functional reads

$$\mathcal{J}[\varphi, Q] = \frac{1}{2}\|(\varepsilon, \varepsilon)\|^2 = \frac{1}{2}\int_\Omega \langle \varepsilon_j, \varepsilon^j \rangle = \frac{1}{2}\int_\Omega \varepsilon_j \wedge \star \varepsilon^j \quad (38)$$

where Ω is the space to be embedded, $\wedge$ stands for the wedge (exterior) product and $\star$ stands for the Hodge duality operator.

By varying φ, Q around the optimal pair $\tilde{\varphi}, \tilde{Q}$ as $\varphi = \tilde{\varphi} + \epsilon_\varphi \delta\varphi$ and $Q = \tilde{Q} + \epsilon_Q \delta Q$ and due to the symmetry of $\langle , \rangle$, the error functional reads $\mathcal{J}[\varphi, Q] = \int_\Omega (\epsilon_\varphi \delta\varphi_j - \epsilon_Q \alpha_i \delta Q^i_j) \wedge \star\varepsilon^j|_{\tilde{\varphi},\tilde{Q}} + O({\epsilon_\varphi}^2) + O({\epsilon_Q}^2)$. As a consequence, the Q - variation of $\mathcal{J}[\varphi, Q]$ read

$$\frac{d}{d\epsilon_Q}\mathcal{J}[\varphi, Q]\Big|_{\epsilon_Q=0} = -\int_\Omega \alpha_i \delta Q^i_j \wedge \star\varepsilon^j \tag{39}$$

while the φ - variations read $\frac{d}{d\epsilon_\varphi}\mathcal{J}[\varphi, Q]\Big|_{\epsilon_\varphi=0} = \int_\Omega d\delta\varphi_j \wedge \star\varepsilon^j$. Then by the generalized Stokes theorem, applied to $d(\delta\varphi_j \wedge \star\varepsilon^j) = d\delta\varphi_j \wedge \star\varepsilon^j + \delta\varphi_j d \wedge \star\varepsilon^j$, the φ - variations of $\mathcal{J}[\varphi, Q]$ read

$$\frac{d}{d\epsilon_\varphi}\mathcal{J}[\varphi, Q]\Big|_{\epsilon_\varphi=0} = \int_{\partial\Omega} \star\varepsilon^j \delta\varphi_j - \int_\Omega \delta\varphi_j d \star \varepsilon^j = \int_\Omega \delta\varphi_j \mathrm{div}(\varepsilon^j) \tag{40}$$

Here the divergence, div, of 1-forms stands for the inner product - dual of the funcions' differential, as determined in (18) to combine both the co-differential operator $d\star$ in Ω and the inner product with the normals of $\partial\Omega$, on $\partial\Omega$.

Considering the Q - stationarity of $\mathcal{J}[\tilde{\varphi}, \tilde{Q}]$ first, we should point out that δQ is not free, since it should hold $Q^T Q = \tilde{Q}^T \tilde{Q} = I_{N\times N} \Leftrightarrow \delta Q^T \tilde{Q} + \tilde{Q}^T \delta Q = \epsilon_Q \delta Q^T \delta Q$. Consequently, at $\epsilon_Q = 0$, $\delta Q^T \tilde{Q} + \tilde{Q}^T \delta Q = 0$. Since $\tilde{Q}^T \delta Q$ should be skew-symmetric, the freely varying components are the lower (or upper) triangular components of $\tilde{Q}^T \delta Q$. Letting $\delta\theta^i_j$ denote these components, for $i > j$, and since $\tilde{Q}\tilde{Q}^T = I_{N\times N}$ the integrand of (39) reads $\alpha_i \delta Q^i_j \wedge \star\varepsilon^j = \delta\theta^k_j(\alpha_i \tilde{Q}^i_k \wedge \star\varepsilon^j - \alpha_i \tilde{Q}^i_j \wedge \star\varepsilon^k)$, with the summation convention applied for $k > j$. Due to the arbitrariness of $\delta\theta^k_j$, the Q - stationarity of $\mathcal{J}[\tilde{\varphi}, \tilde{Q}]$ is equivalent to the (k, j)-symmetry requirement $\alpha_i \tilde{Q}^i_k \wedge \star\varepsilon^j = \alpha_i \tilde{Q}^i_j \wedge \star\varepsilon^k$. Given that the term $\alpha_i \tilde{Q}^i_k \wedge \star\alpha_m \tilde{Q}^m_j = \langle \alpha_i \tilde{Q}^i_k, \alpha_m \tilde{Q}^m_j \rangle$ of $\alpha_i \tilde{Q}^i_k \wedge \star\varepsilon^j$ is (k, j)-symmetric, due to the symmetry of $\langle , \rangle$, the Q - stationarity of $\mathcal{J}[\tilde{\varphi}, \tilde{Q}$ reduces to

$$\langle \alpha_i \tilde{Q}^i_k, d\tilde{\varphi}^j \rangle = \langle \alpha_i \tilde{Q}^i_j, d\tilde{\varphi}^k \rangle \tag{41}$$

This (k, j)-symmetry requirement may be expressed as a matrix symmetry reqirement in the form

$$B_{\tilde{\varphi}} \tilde{Q} = \tilde{Q}^T {B_{\tilde{\varphi}}}^T \ , \ \text{where} \ B_{\tilde{\varphi}} : \langle \alpha_i, d\tilde{\varphi}^j \rangle = (B_{\tilde{\varphi}})^j_i d\mathrm{vol} \tag{42}$$

If the singular matrix decomposition of $B_{\tilde{\varphi}}$ reads $L_{\tilde{\varphi}} \Sigma_{\tilde{\varphi}} {R_{\tilde{\varphi}}}^T$, then $\tilde{Q} = R_{\tilde{\varphi}} \sigma {L_{\tilde{\varphi}}}^T$, where σ is a diagonal matrix with either 1 or -1. By substituting this $\tilde{Q}$ back in $\mathcal{J}[\tilde{\varphi}, \tilde{Q}]$, one gets

$$\mathcal{J}[\tilde{\varphi}, \tilde{Q}] = \frac{1}{2}\int_\Omega \langle d\tilde{\varphi}_j, d\tilde{\varphi}^j \rangle - \int_\Omega \mathrm{tr}(\Sigma_{\tilde{\varphi}} \sigma) d\mathrm{vol} + \frac{1}{2}\int_\Omega \langle \alpha_j, \alpha^j \rangle \tag{43}$$

and since singular values are non-negative, σ should be $I_{N\times N}$, in order to minimize $\mathcal{J}[\tilde{\varphi}, \tilde{Q}]$, thus offering (7b) for determining $\tilde{Q}$.

The next step is to consider the $\tilde{\varphi}$ - stationarity of $\mathcal{J}[\tilde{\varphi}, \tilde{Q}]$, which zeros the right-hand side of (40). Then, arbitrariness of $\delta\varphi$, leads to the requirement $\mathrm{div}(\varepsilon^j) = 0$, thus resulting to the Euler-Lagrange equations (7a) for $\tilde{\varphi}$

$$\mathrm{div}(d\tilde{\varphi}_j) = \mathrm{div}(\alpha_i \tilde{Q}^i_j) \Leftrightarrow \begin{cases} d(\star d\tilde{\varphi}_j) = d\left(\star\alpha_i \tilde{Q}^i_j\right) \\ (d\tilde{\varphi}_j)|_{\partial\Omega} = \left(\alpha_i \tilde{Q}^i_j\right)\Big|_{\partial\Omega} \end{cases} \tag{44}$$

where by $(\cdot)|_{\partial\Omega}$ we denote the inner product of an 1-form on $\partial\Omega$ with the normals of $\partial\Omega$.

D Proof of theorem 1

We consider the scheme (8) of alternating resolutions of (44) and (42) to test if it converges to the set of the solutions of the Euler-Lagrange equations

$$\tilde{\varphi}|_{n+1} : \mathrm{div}(d\ \tilde{\varphi}|_{n+1}) = \mathrm{div}\left(\alpha\tilde{Q}|_n\right) \tag{45a}$$

$$\tilde{Q}|_{n+1} : B_{\tilde{\varphi}|_{n+1}} \tilde{Q}|_{n+1} = \tilde{Q}|_{n+1}{}^T {B_{\tilde{\varphi}|_{n+1}}}^T \tag{45b}$$

Since $\mathcal{J}[\varphi, Q]$ is the squared L^2 norm of the approximation discrepancies $\varepsilon[\varphi, Q]$ of (5), based on inner product, $\mathcal{J}[\varphi, Q]$ is strictly convex with respect to ε and consequently, with respect to $d\varphi$ and convex with respect to Q. However, Q is restricted in the orthogonal group, which is non-convex, thus rendering $\mathcal{J}$, non-convex with respect to (φ, Q).

In appendix C , it is shown that given $\tilde{\varphi}|_n$, $\tilde{Q}|_n$, as results from (45b), it is a global minimizer of $\mathcal{J}[\tilde{\varphi}|_n, Q]$. Moreover, (45b) yields $\tilde{Q}|_{n+1}$ of (8b) as a unique solution, if the singular values of $B_{\tilde{\varphi}|_{n+1}}$ are unique. In any case,

$$\tilde{Q}|_{n+1} \in \left\{ Q : \arg\min_{Q} \mathcal{J}[\tilde{\varphi}|_n, Q] \right\} \tag{46}$$

Consequently $\mathcal{J}[\tilde{\varphi}|_n, \tilde{Q}|_n] \leq \mathcal{J}[\tilde{\varphi}|_n, \tilde{Q}|_{n-1}]$ and in order to verify that the iterative scheme (8) converges to the set of minimizer of $\mathcal{J}$, first one should test if $\mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] \leq \mathcal{J}[\tilde{\varphi}|_n, \tilde{Q}|_n]$, thus rendering $\mathcal{J}$ decreasing in a full iteration of (8).

Considering the Hodge decomposition of the 1-forms $\alpha\tilde{Q}|_n$, the defining equation of $\tilde{\varphi}|_{n+1}$ in (45a) is the one that determines the exact component of the decomposition. As a result, the discrepancies $\varepsilon[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] = d\tilde{\varphi}|_{n+1} - \alpha\tilde{Q}|_n$ consist of the remaining terms of the Hodge decomposition (co-exact and harmonic), thus satisfying

$$d(\star\varepsilon[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n]) = 0 \tag{47}$$

Now, if, instead of $\tilde{\varphi}|_{n+1}$, one employs functions $\varphi = \tilde{\varphi}|_{n+1} + \phi$, then the discrepancies read $\varepsilon[\varphi, \tilde{Q}|_n] = \varepsilon[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] + d\phi$ and the integrand of $\mathcal{J}[\varphi, \tilde{Q}|_n]$ becomes

$$\left\langle \varepsilon_j[\varphi, \tilde{Q}|_n], \varepsilon^j[\varphi, \tilde{Q}|_n] \right\rangle = \left\langle \tilde{\varepsilon}_j[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n], \tilde{\varepsilon}^j[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] \right\rangle + \langle d\phi_j, d\phi^j \rangle + 2 \left\langle d\phi_j, \tilde{\varepsilon}^j[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] \right\rangle \tag{48}$$

Using the identity (47), the third term of this expansion reads $\left\langle d\phi_j, \tilde{\varepsilon}^j[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] \right\rangle = d\left(\phi_j \star \tilde{\varepsilon}^j[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] \right)$. As a consequence, integration of $\left\langle \varepsilon_j[\varphi, \tilde{Q}|_n], \varepsilon^j[\varphi, \tilde{Q}|_n] \right\rangle$ over Ω offers

$$\mathcal{J}[\varphi, \tilde{Q}|_n] = \mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] + \frac{1}{2}\int_\Omega \langle d\phi_j, d\phi^j \rangle + \int_{\partial\Omega} \phi_j \star \tilde{\varepsilon}^j[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] \tag{49}$$

Since the boundary conditions that div imposes in (45a) are the ones of (44), $(d\tilde{\varphi}|_{n+1})|_{\partial\Omega} = \left(\alpha\tilde{Q}|_n\right)\Big|_{\partial\Omega}$ or equivalently, $\tilde{\varepsilon}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n]\Big|_{\partial\Omega} = 0$, thus offering

$$\mathcal{J}[\varphi, \tilde{Q}|_n] = \mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] + \frac{1}{2}\int_\Omega \langle d\phi_j, d\phi^j \rangle \Leftrightarrow \mathcal{J}[\varphi, \tilde{Q}|_n] > \mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] \tag{50}$$

So $\tilde{\varphi}|_{n+1}$, obtained by resolving the Neumann boundary problem of the Poisson - like equation (45a), is a global minimizer of $\mathcal{J}[\varphi, \tilde{Q}|_n]$ and (50) implies that $d\tilde{\varphi}|_{n+1}$ is unique and consequently, $\tilde{\varphi}|_{n+1}$ is unique modulo constants. Moreover, according to (50), the error functional decrease achieved by the update $\tilde{\varphi}|_n \to \tilde{\varphi}|_{n+1}$ reads

$$\mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] - \mathcal{J}[\tilde{\varphi}|_n, \tilde{Q}|_n] = -\frac{1}{2}\left\| d(\tilde{\varphi}|_{n+1} - \tilde{\varphi}|_n) \right\|^2 \tag{51}$$

Consequently, the considered iterative scheme of (8) renders the error functional $\mathcal{J}[\varphi, Q]$ decreasing, satisfying in a full iteration

$$\mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n] \leq \mathcal{J}[\tilde{\varphi}|_n, \tilde{Q}|_n] \leq \mathcal{J}[\tilde{\varphi}|_n, \tilde{Q}|_{n-1}] \tag{52}$$

Since also $\mathcal{J} \geq 0$, the error functional values $\mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n]$ converge and (51) & (52) imply that this convergence leads to $\lim_{n\to\infty} d(\tilde{\varphi}|_{n+1} - \tilde{\varphi}|_n) = 0$.

Also, (50) & (52) imply that each iteration of (8) splits in two alternating global optimization steps

$$d\tilde{\varphi}|_{n+1} = \arg\min_{\varphi} \mathcal{J}[\varphi, \tilde{Q}|_n] \tag{53a}$$

$$\tilde{Q}|_{n+1} \in \arg\min_{Q} \mathcal{J}[\tilde{\varphi}|_{n+1}, Q] \tag{53b}$$

and the discrepancy of $\tilde{\varphi}|_n$ from an Euler-Lagrange critical point is controlled by $\|d\tilde{\varphi}|_n - d\tilde{\varphi}|_{n+1}\|$. This is because, for this n, $\tilde{Q}_n$ already satisfies (8b) and $d\tilde{\varphi}|_{n+1}$ is the exact 1-form that satisfies satisfies (8a). Consequently, the iterative scheme of (8) asymptotically satisfies the Euler-Lagrange equations of (7).

Concerning the limit set of $(\tilde{\varphi}|_{n+1}, \tilde{Q}|_n)$ (i.e. the inverse image of the limit of $\mathcal{J}[\tilde{\varphi}_{n+1}, \tilde{Q}_n]$), one can use the following identity

$$\|d\tilde{\varphi}_{n+1}\|^2 = \int_\Omega \left\langle d\tilde{\varphi}_j|_{n+1}, d\tilde{\varphi}^j|_{n+1} \right\rangle \overset{(45a)}{=} \int_\Omega \left\langle d\tilde{\varphi}_j|_{n+1}, \alpha^i \tilde{Q}_i^j|_n \right\rangle \tag{54}$$

to obtain for $\mathcal{J}[\tilde{\varphi}_{n+1}, \tilde{Q}_n]$

$$\mathcal{J}[\tilde{\varphi}_{n+1}, \tilde{Q}_n] \equiv \frac{1}{2}(\|\alpha\|^2 + \|d\tilde{\varphi}_{n+1}\|^2) - \int_\Omega \left\langle d\tilde{\varphi}_j|_{n+1}, \alpha^i \tilde{Q}_i^j|_n \right\rangle = \frac{1}{2}(\|\alpha\|^2 - \|d\tilde{\varphi}_{n+1}\|^2) \tag{55}$$

Consequently, the limit set of $(\tilde{\varphi}|_{n+1}, \tilde{Q}|_n)$ consists of maps $\tilde{\varphi}_\infty$ with fixed $\|d\tilde{\varphi}_\infty\|^2$, which is the limit of the increasing sequence of $\|d\tilde{\varphi}|_n\|^2$.

If the map $d\tilde{\varphi}|_n \mapsto d\tilde{\varphi}|_{n+1}$ is continuous and $\{d\tilde{\varphi}|_n\}_{n\geq 1}$ is compact, convergence of the decreasing $\mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n]$ allows for applying the discrete LaSalle invariance principle (Theorem 1.19, Section 1.3.3, pp. 15–16, of [25]), which implies that $(\tilde{\varphi}|_{n+1}, \tilde{Q}|_n)$ converges to a limit set $[\tilde{\varphi}|_\infty, \tilde{Q}|_\infty]$, contained in the zero-decrease set of $\mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n]$. Then, (52) together with (51) imply that the zero-decrease of $\mathcal{J}[\tilde{\varphi}|_{n+1}, \tilde{Q}|_n]$ results to $d\tilde{\varphi}|_{n+1} = d\tilde{\varphi}|_n$. By fixing the value of $\int_\Omega \tilde{\varphi}|_n$, the zero-decrease set consists of $\tilde{\varphi}|_n$ – fixed points $\tilde{\varphi}_*$ and then all corresponding $\tilde{Q}_*$ are obtained by (45b) as

$$\tilde{Q}_* = R_{\tilde{\varphi}_*} {L_{\tilde{\varphi}_*}}^T \tag{56}$$

Also, (45a) implies that $\tilde{\varphi}_*$ is a solution of

$$\mathrm{div}(d\tilde{\varphi}_*) = \mathrm{div}\left(\alpha \tilde{Q}_*\right) \tag{57}$$

Equations, (57) and (56) are exactly the stationarity equations (7) of $\mathcal{J}[\varphi, Q]$, thus implying that $[\tilde{\varphi}|_\infty, \tilde{Q}|_\infty]$ consists of the Euler-Lagrange critical points of $\mathcal{J}[\varphi, Q]$.

Continuity of $d\tilde{\varphi}|_n \mapsto d\tilde{\varphi}|_{n+1}$ is controlled by the continuity of the composition of the singular vectors (needed to obtain $\tilde{Q}|_n$ from $d\tilde{\varphi}|_n$) with the projection of an 1-form to its exact component (needed to obtain $d\tilde{\varphi}|_{n+1}$ from $\tilde{Q}|_n$). The latter operator is continuous; two Q_1, Q_2 result to corresponding φ_1, φ_2, such that $\mathrm{div}(\varepsilon[\varphi_1, Q_1]) = \mathrm{div}(\varepsilon[\varphi_2, Q_2]) = 0$, thus forcing $\|d(\varphi_1 - \varphi_2)\|^2 = \int_\Omega \left\langle d((\varphi_1)_j - (\varphi_2)_j), \alpha^i (Q_1 - Q_2)_i^j \right\rangle \leq \|d(\varphi_1 - \varphi_2)\| \|\alpha(Q_1 - Q_2)\|$, implying continuity of the exact component $d\varphi$, with respect to its source 1-form αQ. So, the continuity of $\tilde{\varphi}|_n \mapsto \tilde{\varphi}|_{n+1}$ is controlled by the continuity of (7b), with respect to B_φ. If $d\varphi$ is of the same rank with α, B_φ is of full rank and Q of (7b) may be written as $Q = {B_\varphi}^T (B_\varphi {B_\varphi}^T)^{-1/2}$, thus being continuous.

E Details on the empirical evaluation and additional experimental results

E.1 Execution time requirements of the proposed algorithm

To evaluate the variation of the actual wall-clock time of each iteration of the embedding estimation, with respect to the number of data points and the embedding dimension, the following hardware set up is fixed: CPU – AMD EPYC™ 9655p (96 CPU cores at 2.60GHz), RAM – 128 GB, DDR5-6400 RDIMM, cache – L3 384 MB. Next, the N-dimensional embedding of V data points uniformly sampled on the $(N-1)$-sphere is computed, for all $V = 2^\nu$, $\nu = 10..14$ and $N = 3..11$. All performed iterations' wall clock times were registered and the results are summarized in figure 3. It should be noted that in this timing experiment, parallelization is adopted with respect to the embedding components only if the first iteration time exceeds 1sec. The algorithm is implemented in MATLAB R2025b [26], using the corresponding Parallel Computing Toolbox [27], with a threads-based profile, for the aforementioned parallelization.

Consistently with the theoretical time complexity of $O(V^{1+\epsilon})$, with $\epsilon \in [0, 1/N)$, the iteration timings slightly exceed the linear complexity pattern in the lower dimensions ($N = 3$ & $N = 4$). For higher embedding dimensions the iteration timings accurately follow the corresponding linear complexity pattern. When the first iteration's time exceeds 1sec, dimension-wise parallelization is adopted to restore a stable linear complexity pattern across different dimension (these are the cases $V \geq$ 16k and $N \geq 8$ or $V \geq$ 8k and $N \geq 11$).

E.2 Details on the intrinsic dimension estimation and its relation to the graph construction

The performed experiments, primarily aim at evaluating the proposed method's capability to retain local metric information, while recovering approximately isometric Euclidean embedding of the data. In this process, the embedding's underlying distance graph is constructed by the minimal neighborhoods that guarantee local metric estimates of rank equal to the data intrinsic dimension. For the intrinsic dimension, its minimal (2NN) neighborhood estimator is adopted, as the one that considers the most detailed approximation of the data density. The intention was to make the "most local" choices for the coupled number of neighbors and intrinsic dimension hyper-parameters, thus evaluating the proposed embedding method under minimal preliminary noise suppression.

The determined intrinsic dimension is considered as a lower bound for the dimensionality of the embedding space. The distortion imposed by asking an embedding of dimension lower than the intrinsic one is technically caused by the projection of the metric trivializing transports α to their dominant components. Consequently, the distortion depends on the considered manifold's metric, the worst case being metrics of low-dimensional manifolds with equal eigenvalues (conformally flat). If, moreover, we retain the rule to select the minimal neighbourhood that recovers the target dimensionality, decrease of the embedding dimension also affects the graph quality, practically causing abrupt decline in the quality of the computed embedding. This is a consequence of the

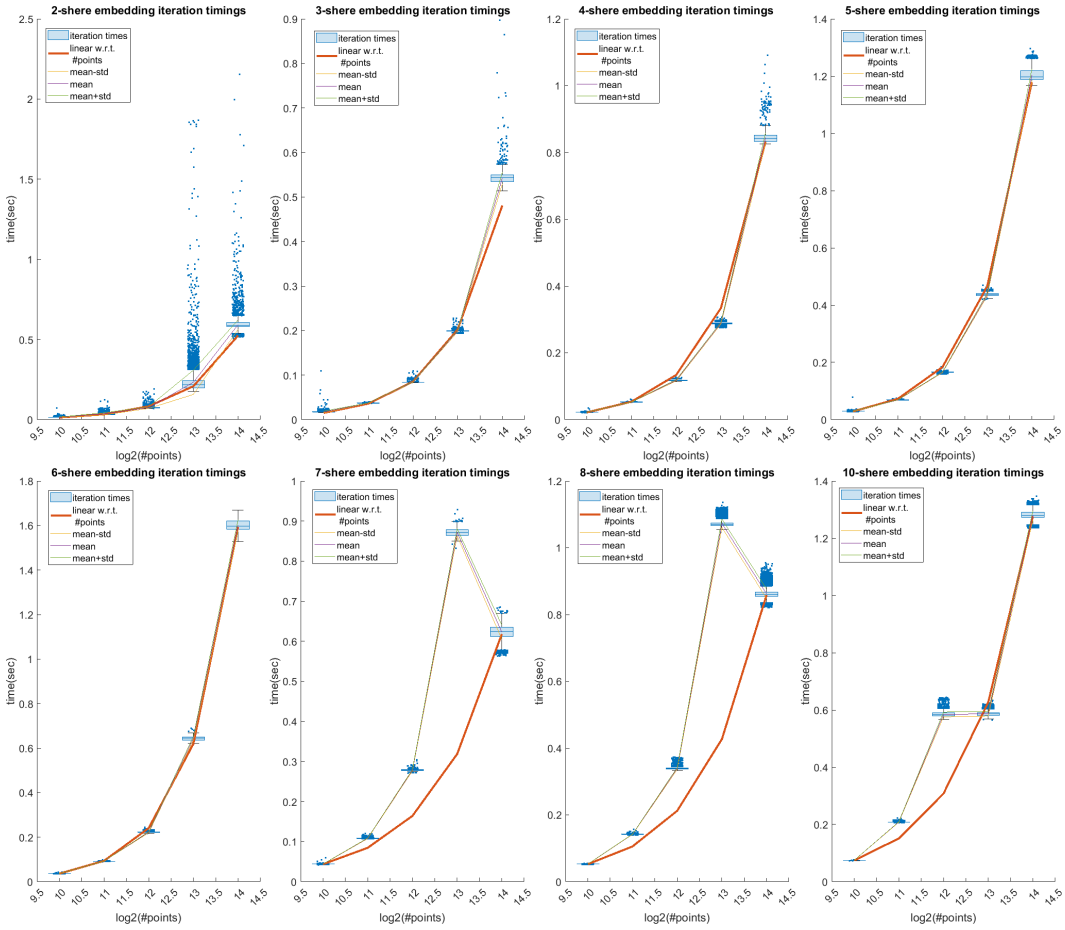

Figure 3: Iteration timings with respect to the number of data points and the embedding dimension. The timings practically follow a linear increase pattern (orange bold line) with respect to the numbers of points. In lower embedding dimensions $N = 3, 4$ the timings' increase slightly exceed the linear increase pattern, due to possible $O(\#\text{ points}^{1/N})$ convergence time of the preconditioned Laplace inversion. When the embedding dimension $N \geq 8$, dimension-wise parallelization is enabled, when the 1st iteration time exceeds 1sec to restore a stable linear pattern across different dimensions. For presentation issues, the 10-dimensional embedding of the 9-sphere is omitted, since its timing increase pattern practically follows the one of the 8-sphere.

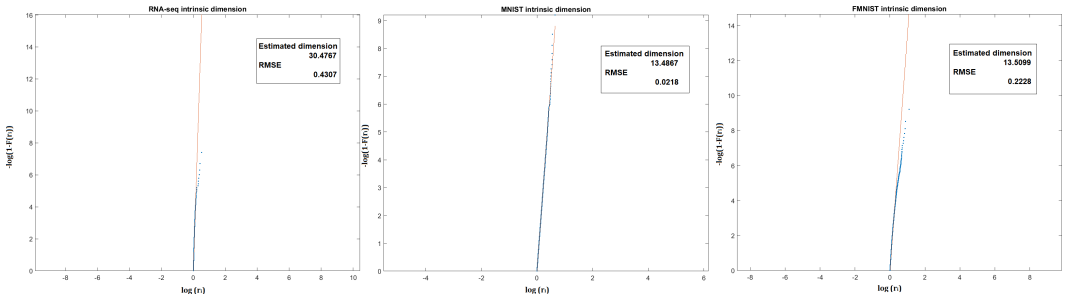

Figure 4: Intrinsic dimension minimal (2NN) neighborhood estimates, for the real data sets.

fact that setting the number of neighbors to be lower than the intrinsic dimension randomizes the dimensions represented in each local frame. On the other hand, projections of local frames of a full rank to lower dimensions represent principal directions consistently across different frames. Moreover, the neighborhood size – sensitivity experiments, whose results are summarized in figures 1 and 7, indicate that the decrease of the neighborhood size negatively affects the quality of the reconstructed embedding, both locally and globally.

Concerning the fidelity of the 2NN-estimate of the data intrinsic dimension, it may be evaluated by the accuracy of the line that best fits the pairs $(\log(r_i), -\log(1 - F(r_i)))$, where r_i is the ratio of the distances of the 2NNs from the neighborhood's origin and F stands for its empirical cumulative distribution (see [13]). In figure 4, visual evidence on this fit is provided, together with the root of its mean squared error and the estimated intrinsic dimensions of the real-world data sets. Namely, the estimated dimension for RNA-seq is 30.476, with RMSE 0.4307, for the MNIST test set it is 13.4867, with RMSE 0.0218 and for the FMNIST test set it is 13.5099, with RMSE 0.2228.

The minimal neighborhood intrinsic dimension estimate, though based on the most detailed approximation of the data density, may result in over-estimation of the data dimension, when the noise increases. To test how the intrinsic dimension estimates change as the neighborhood size increases, thus suppressing the effect of noise, the Generalized Ratios Intrinsic Dimension Estimators (GRIDE) [28] have been computed on the real data sets. Specifically, the graph constructed on the basis of the 2NN-estimate is considered, since it is the one that corresponds to the most noisy estimate, thus requiring the maximum number of neighbors. On this basis, the GRIDEs have been computed for all pairs of nearest neighbors and the corresponding distributions with respect to the possible neighbors' pairs are provide in figure 5. Comparing the estimate of [13] with the mean of the GRIDEs one, for the FMNIST and MNIST test sets, these estimates are very close, thus confirming the fidelity of the adopted intrinsic dimensions. Namely, for the FMNIST, the 2NN estimate is 13.5099 and the GRIDEs' mean is 13.1591 and for MNIST, the 2NN estimate is 13.4867 and the GRIDEs' mean is 12.1533. The case of RNA-seq is somewhat different; the 2NNs' estimate corresponds to narrow neighborhoods' intrinsic dimension, while there exist neighborhoods containing longer edges (probably isolated or clusters' boundary data points) whose intrinsic dimension is lower (see the first plot of fig. 4 and refer to the RMSE of 0.4307, which is significantly larger than the one of MNIST & FMNIST). This is possibly a consequence of the fact that data points form clusters of high dimensional, though limited variation, whose distant connections form low dimensional frames. Such a hypothesis is consistent with the approximate isometries of IsoMap and of the proposed method, visualized in 10, where the data points are concentrated in clusters, placed in a 2D arrangement. In numbers, the GRIDEs' mean, for the RNA-seq data set, is 22.8196, while the 2NN estimate is 30.476.

In the following appendix section E.6.1, embeddings are constructed and evaluated, replacing the 2NN by the GRIDEs' mean intrinsic dimension estimate. The results show a borderline modification in the embeddings' evaluation and in most cases, the diagnoses' discrimination is improved, thus signaling that the extra dimensions of the fine, 2NN estimate are not essential in the diagnoses' separation. However, the geometric consistency of the embeddings indicates that these extra dimensions carry geometric information. Local methods' evaluation slightly declines and global methods improve their global consistency, however distorting neighborhoods. This picture cannot be attributed to the methods' performance as the neighborhood sizes decrease. Figures 1 and 7 that summarize neighborhood sensitivity results, indicate that, as long as the neighborhoods' size does not affect dimensionality, the local metric reconstruction (local stain) is not affected by the change of neighborhoods' size in the embeddings of the local methods that do not consider global metric consistency (LLE, HLLE, LTSA, Laplacian Eigenmaps). Contrary, these methods' local metric reconstruction declines in the RNA-seq data set, when the neighborhood sizes decreased, along with the embedding's dimension. On the other hand, the global, geodesic distances' reconstruction (global stress) in the RNA-seq data behaves consistently with the neighborhood sensitivity pattern; global methods' performance declines as the NNs increases, while the proposed method's performance increases.

Consequently, the initial hypothesis that that the RNA-seq data points form high dimensional clusters, of limited variation, globally arranged in a lower dimensional frame is also supported by the lower dimensional embeddings' evaluation.

E.3 The adopted evaluation metrics' formulas

For the metrics **Lcl-*** that evaluate the preservation of the local metric structure, the distance graph's adjacency matrix M^{v_i} and the corresponding Gram matrix G^{v_i} (Sct. 2.1) of each v_i data point's neighborhood are respectively considered as the ground truth local distances' and metric's estimates. The corresponding estimates that result from the embedding in hand, $\tilde{\varphi}$, read

$$\tilde{M}^{v_i}|_{v_j} = \|\tilde{\varphi}(v_j) - \tilde{\varphi}(v_i)\| \tag{58}$$

$$\tilde{G}^{v_i}|_{v_j,v_k} = (\tilde{\varphi}(v_j) - \tilde{\varphi}(v_i)) \cdot (\tilde{\varphi}(v_k) - \tilde{\varphi}(v_i)) \tag{59}$$

On the basis of these estimates, the **Lcl-*** metrics are computed as

- **Lcl-dist**: $\underset{v_i}{\text{mean}} \left(\underset{v_j \in \Gamma(v_i)}{\text{mean}} \left| \frac{\tilde{M}^{v_i}|_{v_j}}{M^{v_i}|_{v_j}} - 1 \right| \right)$.

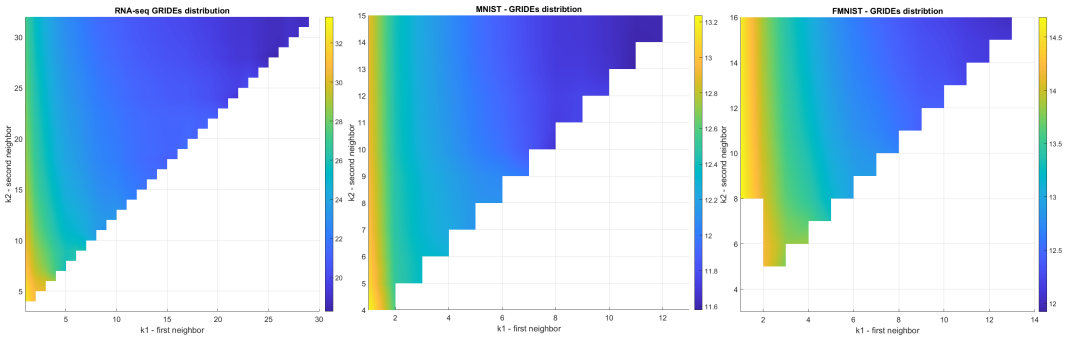

Figure 5: Generalized Ratios Intrinsic Dimension Estimators (GRIDE), for the real data sets. FMNIST and MNIST GRIDEs are close to the employed 2NNs estimates, while the GRIDEs of the RNA-seq data set vary from this elementary estimate significantly. This is possibly a consequence of the fact that data points form high dimensional clusters, though of limited variation, whose distant connections form low dimensional frames.

- **Lcl-ang**: $\underset{v_i}{\text{mean}}\left(\underset{(v_j,v_k)\in\Gamma(v_i)}{\text{mean}}\left|\tilde{\theta}^{v_i}|_{v_j,v_k}-\theta^{v_i}|_{v_j,v_k}\right|\right)$, where, in each neighborhood $\Gamma(v_i)$, only the unique pairs (v_j,v_k) of neighbors of v_i are considered and $\theta^{v_i}|_{v_j,v_k}=\arccos\left(\frac{G^{v_i}|_{v_j,v_k}}{M^{v_i}|_{v_j}M^{v_i}|_{v_k}}\right)$, $\tilde{\theta}^{v_i}|_{v_j,v_k}=\arccos\left(\frac{\tilde{G}^{v_i}|_{v_j,v_k}}{\tilde{M}^{v_i}|_{v_j}\tilde{M}^{v_i}|_{v_k}}\right)$.
- **Lcl-mtrc**: $\underset{v_i}{\text{mean}}\left(\underset{(v_j,v_k)\in\Gamma(v_i)}{\text{mean}}\left\|\left(\tilde{G}^{v_i}G_{-}{}^{v_i}\right)\Big|_{v_j,v_k}-1\right\|\right)$, where $G_{-}{}^{v_i}$ denotes the $\text{rank}\,\text{dim}\,\tilde{\varphi}$ – pseudoinverse of G^{v_i} and $\|\cdot\|$ the Frobenius norm.
- **Lcl-prec**, **Lcl-rec**, **Lcl-F1**: the precision, recall and F1 score of $\tilde{\Gamma}(v_i)$, with respect to the ground truth $\Gamma(v_i)$, where $\tilde{\Gamma}(v_i)$ is the neighborhood of v_i that has the same number of neighbors with $\Gamma(v_i)$, however resulting from $\tilde{M}^{v_i}$

Concerning the metrics **Glbl-*** that evaluate the reconstruction of the manifold's global structure, the shortest path distances are approximated via Dijkstra algorithm, forming a dense symmetric matrix $D^{v_i}|_{v_j}$, which is used as ground truth for the geodesics' straightening evaluation. Also, for the synthetic data sets, the underlying parameterization, $t\in\mathbb{R}^{\text{intrinsic dim}}$, is considered as the ground truth of the manifold's unfolding. On the basis of these two global entities, the metrics **Glbl-*** read

- **Glbl-mtrc**: $\sqrt{\underset{v_i}{\text{mean}}\left(\frac{\underset{v_j}{\text{mean}}\left(\tilde{M}^{v_i}|_{v_j}-D^{v_i}|_{v_j}\right)^2}{\underset{v_j}{\text{mean}}D^{v_i}|_{v_j}{}^2}\right)}$
- **Glbl-corr**: $\exp\left(\underset{v_i}{\text{mean}}\log\left|\underset{v_j}{\text{corr}}\left(\tilde{M}^{v_i}|_{v_j},D^{v_i}|_{v_j}\right)\right|\right)$, where corr is the Pearson / Spearman / Kendall correlation.
- **Glbl-prm**: $\underset{v_i}{\text{mean}}\left\|\tilde{t}(v_i)-t(v_i)\right\|^2$, where $\tilde{t}$ results from the affine procrustes alignment of the embedding $\tilde{\varphi}$ with t. Namely, $\tilde{t}=\text{cov}[t,\tilde{\varphi}]\text{cov}[\tilde{\varphi},\tilde{\varphi}]^{-1}\left(\tilde{\varphi}-\underset{v_i}{\text{mean}}[\tilde{\varphi}(v_i)]\right)+\underset{v_i}{\text{mean}}[t(v_i)]$, where cov stands for the covariance over the graph vertices v_i.

E.4 Unfolding of the Swiss roll

The widely used case of "Swiss roll" surface unfolding is considered for a visual comparison of the proposed embedding algorithm against baseline methods' embedding results. Specifically, the Swiss roll embeddings' visualizations in figure 6 depict that in small neighborhood sizes (there 7NNs), the proposed is the only local method that faithfully approximates the 2D isometry of the surface, optimally recovered by IsoMap. LLE and Laplacian Eigenmaps practically contract the points along the non-curved principal direction, while HLLE and LTSA failed to represent any of the principal directions and their embeddings have been omitted. On the other hand, tSNE clusters data points according to their density, however successfully representing both principal directions within clusters.

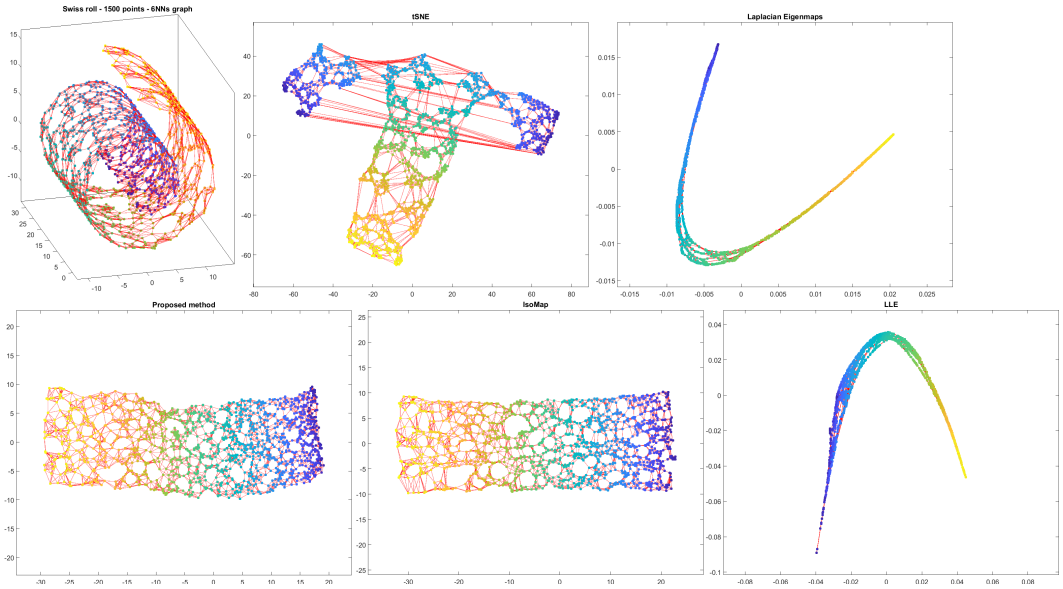

Figure 6: Reconstruction of the Swiss roll embedding in $\mathbb{R}^2$ from the 7 nearest neighbors' distance graph of 1500 points, randomly sampled on the surface. The coloring depicts the data points' ordering along the surface's dominant principal direction.

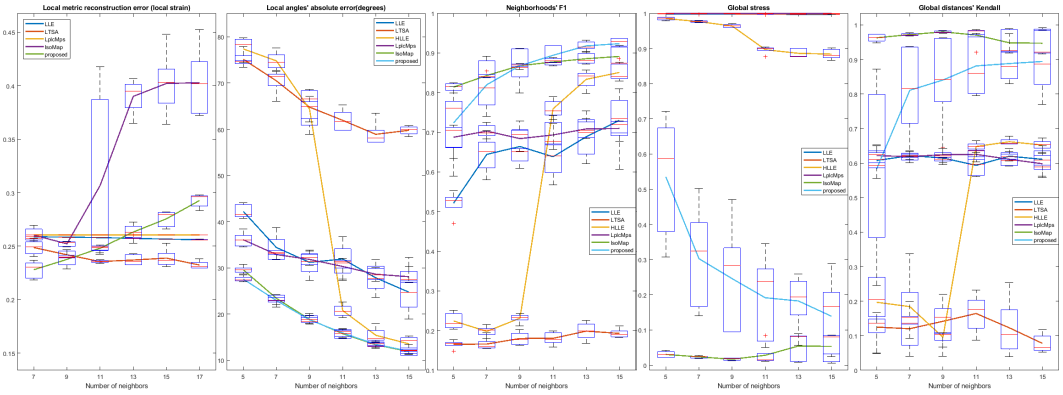

Figure 7: Swiss roll unfolding experiment for varying neighborhood size. The measures' box plots depict the maximum/minimum boundaries at whiskers, the boxes of the populations' 25th-75th percantile and the median values in red. The plotted lines depict the variation of the metrics' mean values with respect to the neighborhood size. At each neighborhood size, 15 random samplings were performed, each of 1500 datapoints.

Moreover, the Swiss roll has undergone the neighborhood size – sensitivity experiment described in Sct. 5.1. The evaluation of the embeddings' quality is summarized in Fig. 7. The proposed is the only local method approximating the performance of IsoMap, in terms of global consistency, while retaining the local geometric data and the neighboring relations. Local metric conservation is adversarial to the integrability of the computed embedding, due to the curvature of the underlying manifold. After all, the deformation of the local metric for the sake of integrability is the error functional minimized by the proposed method. Consistently, the proposed method results to a more conservative local metric distortion than IsoMap, however not following the other local methods' invariance with respect to the neighborhood size. Actually, while the number of neighbors increases, the intra-neighborhood distances deviate more intensely from the true geodesic ones and if a globally accurate isometric embedding is recovered, these distances should be stretched to approximate their true geodesic counterparts.

Table 3: Evaluation of embeddings' quality in synthetic data sets without smooth Euclidean isometries

	Klein's bottle					Torus				
	Lcl-mtrc ↓	Lcl-F1 ↑	Glbl-Pearson ↑	Glbl-mtrc ↓	Glbl-prm ↓	Lcl-mtrc ↓	Lcl-F1 ↑	Glbl-Pearson ↑	Glbl-mtrc ↓	Glbl-prm ↓
LLE	0.30	0.59	0.66	1.00	0.09	0.45	0.50	0.69	0.98	0.61
HLLE	3.88	0.23	0.07	0.90	0.80	13.5	0.16	0.06	1.25	0.97
LTSA	0.30	0.16	0.07	1.00	1.00	0.52	0.19	0.05	0.98	1.00
LplcMps	0.30	0.61	0.67	1.00	0.09	0.44	0.47	0.77	0.99	0.67
DffMps	0.27	0.86	0.76	0.75	0.03	0.83	0.56	0.87	0.25	0.45
IsoMap	0.77	0.81	0.98	0.09	0.00	0.60	0.55	0.94	0.14	0.67
proposed	0.44	0.91	0.95	0.23	0.07	1.14	0.76	0.80	0.41	0.48

E.5 Synthetic data sets without smooth Euclidean isometries

In order to test the proposed method on manifold embedding cases with non-existent Euclidean isometries, we have considered the cases of flat Torus and Klein's bottle and their smooth embedding in $\mathbb{R}^3$. While both surfaces are flat when embedded in $\mathbb{R}^4$ and $\mathbb{R}^5$, correspondingly, their smooth embedding (or immersion in the case of Klein's bottle) in $\mathbb{R}^3$ necessarily distorts the Euclidean metric. Consequently, in the performed experiment, the surfaces are sampled in $\mathbb{R}^4$ & $\mathbb{R}^5$ and embedded in $\mathbb{R}^3$. The evaluation of the obtained embeddings is summarized in Table 3, which is analogous to the Table 1 provided in the paper for the swiss roll and the helix, while the computed embeddings are presented in figures 8 and 9.

Flat torus: Parameters $(u, v) \in [0, 2\pi)^2$, equations of the flat torus in $\mathbb{R}^4$:

$$r(u, v) = (2\cos(u), 2\sin(u), \cos(v), \sin(v))$$

equations of the short embedding in $\mathbb{R}^3$

$$\begin{aligned} x(u, v) &= \cos(u)(2 + \sin(v)) \\ y(u, v) &= \sin(u)(2 + \sin(v)) \\ z(u, v) &= \cos(v) \end{aligned}$$

Klein's bottle: Parameters $(u, v) \in [0, 2\pi)^2$, isometric embedding of the Klein's bottle in $\mathbb{R}^5$:

$$r(u, v) = \left(\frac{1}{2}\cos(v)\cos(2u), \frac{1}{2}\cos(v)\sin(2u), \sin(v)\cos(2u), \sin(v)\sin(2u), \frac{\sqrt{3}}{2}\cos(v)\right)$$

equations of the short embedding in $\mathbb{R}^3$

$$\begin{aligned} x(u, v) &= 6\cos(u)(1 + \sin(u)) + 4(1 - 0.5\sin(u))(\cos(u)\cos(v)\chi(u \leq \pi) + \cos(v + \pi)\chi(u > \pi)) \\ y(u, v) &= 16\sin(u) + 4(1 - 0.5\cos(u))\sin(u)\cos(v)\chi(u \leq \pi) \\ z(u, v) &= 4(1 - 0.5\cos(u))\sin(v) \end{aligned}$$

As a ground truth for the Klein bottle's isometric immersion in $\mathbb{R}^3$, we have adopted the embedding of IsoMap.

E.6 Real data sets

E.6.1 Diagnoses' discrimination in the RNA-seq dataset

The connection between embeddins' local/global geometric consistency and their data discrimination capability, hypothesized by the results of table 2, becomes more evident via embeddings' visualization. In Fig. 10, the embeddings' three top-scaled dimensions are plotted and it becomes evident how the different methods distribute the datapoints, within their embeddings. The stochastic distribution – based tSNE and UMAP practically cluster data according to the cross-homogeneity of the neighborhoods' internal distribution. IsoMap and the proposed method offer distances preserving embeddings that respectively enforce continuity and smoothness, thus not amplifying data clustering, however retaining the original neighboring relations. The deviation in the picture that IsoMap and the proposed method offer for the dataset mainly reflects the fact that data points' alignment in IsoMap is determined by the original points' shortest paths, while the alignment in the proposed method's picture is determined by the Laplacian inversion – induced integration. As a result, any shortest paths' on-graph collapse is also evident in the IsoMap embedding, while Laplacian inversion imposes C^2 continuity, thus not allowing such collapses.

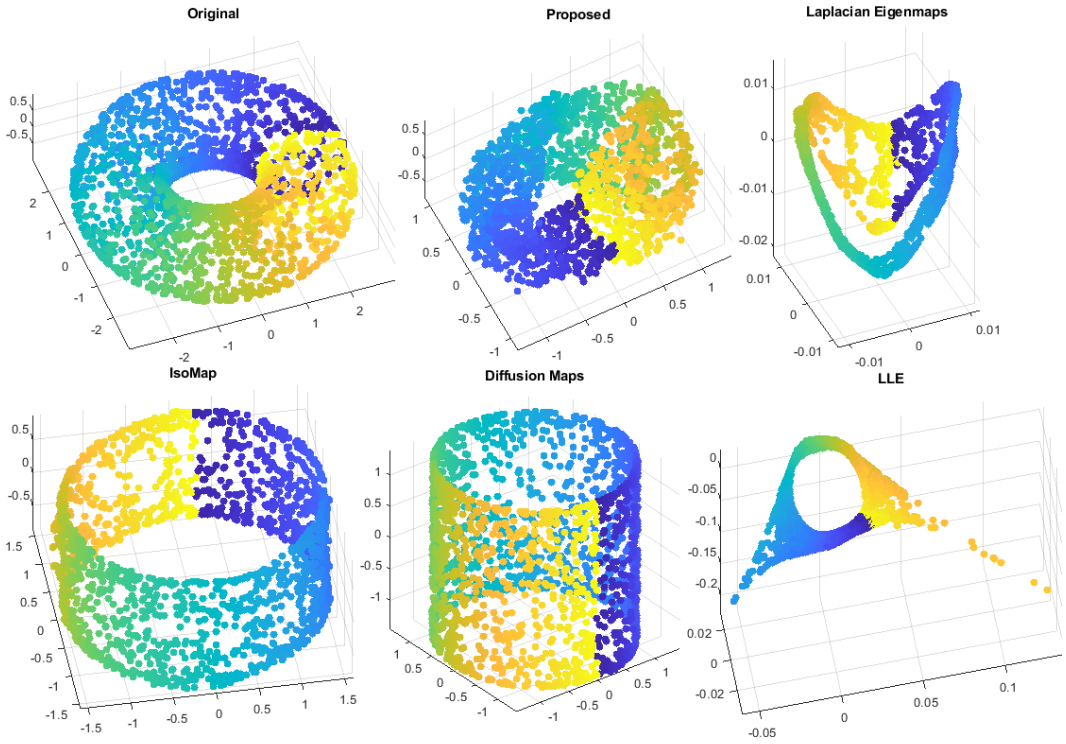

Figure 8: Embedding of the flat torus in $\mathbb{R}^3$. Reconstruction of this short embedding from the neighborhoods' graph of constant intrinsic dimension 2, across all neighborhoods. The coloring depicts the data points' polar angle in the original figure. The (u, v) - parameterization of the manifold is used as ground truth for evaluating the embeddings' capability to recover the manifold's product structure. The global embedding of Diffusion Maps best approximates it up to affine transformations, while the proposed local method follows, however up to orthogonal transformations (see also table 1). Visually, the proposed embedding method best approximates the support of the $\mathbb{R}^3$ isometry of the flat torus, while the baseline ones, either local or global, they flatten the v-coordinate.

Table 4: Evaluation of the diagnoses' representation for the RNA-seq data set, on the basis of 23-dimensional embeddings

	Diagnosis classification within embeddings' neighborhoods							
	LLE	LTSA	LplcMps	IsoMap	PHATE	t-SNE	UMAP	proposed
ACC ↑	0.6991	0.6695	0.6997	0.6788	0.6911	0.7065	0.7009	0.7138
NMI ↑	0.3012	0.2601	0.3033	0.2736	0.2979	0.3323	0.3226	0.3382
ARI ↑	0.3397	0.2630	0.3591	0.3012	0.3354	0.3857	0.3754	0.3996
	Local/global geometric consistency							
Lcl-F1 ↑	0.5261	0.2888	0.5507	0.5592	0.5512	0.6607	0.6976	0.7534
Lcl-mtrc ↓	0.0918	0.0912	0.0922	0.1321	0.1039	3.7875	0.1908	0.0690
Glbl-mtrc ↓	0.9516	0.9486	0.9908	0.1273	0.5397	32.50	2.7475	0.2238

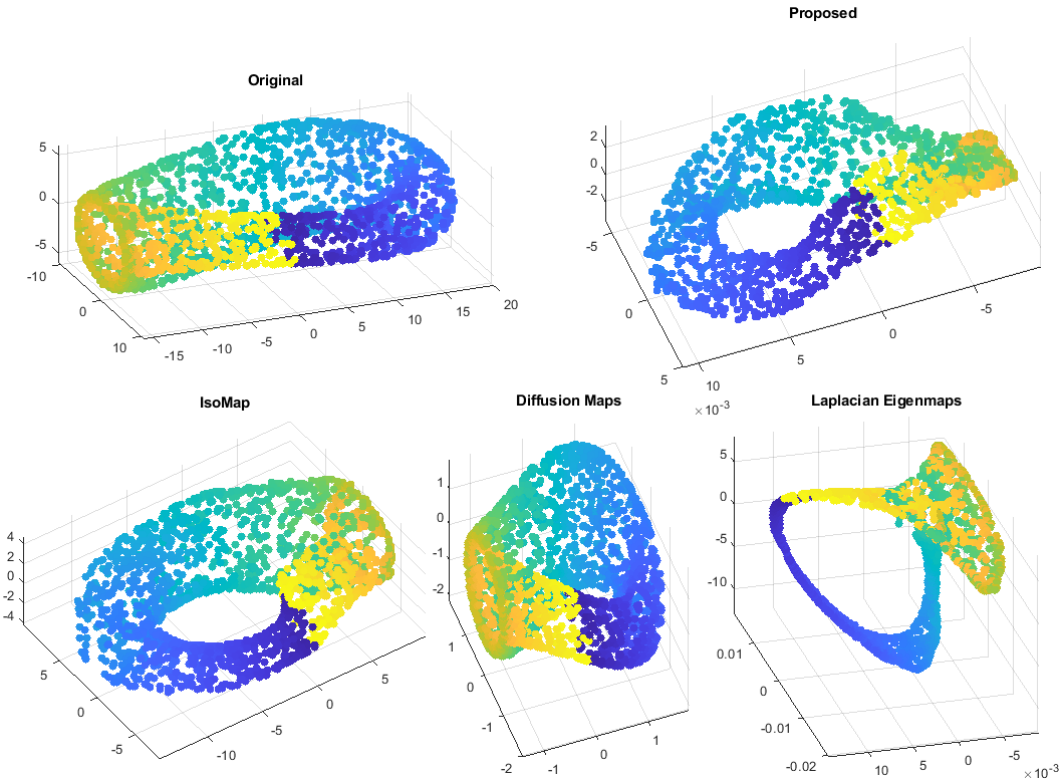

Figure 9: Reconstruction results on the Klein's bottle. Reconstruction of the original bottle figure from the neighborhoods' graph of constant intrinsic dimension 2, across all neighborhoods. The coloring depicts the data points' polar angle in the original figure. The embedding of IsoMap is used as ground truth for the approximate isometry of the original bottle figure. The global embedding of Diffusion Maps best approximates it up to affine transformations, while the proposed local method follows, however up to orthogonal transformations (see also table 3). As a result, the proposed local method follows the global embedding of IsoMap in the Euclidean approximation of the original manifold's geodesic distances, while Diffusion Maps distort them.

The RNA-seq embedding experiment has also been performed, using as intrinsic dimension estimate the mean of the GRIDE estimates, which are provided in figure 5 and their mean is 22.8196 (App. E.2). It should be stressed that the obtained embeddings do not coincide with the projection of the 30-dimensional embeddings to their 23 principal components, since the intrinsic dimension also determines the number of neighbors (via the minimal neighborhood with metric of dimension at least as the intrinsic one). The capability of these lower dimensional embeddings to capture the diagnoses' discrimination is summarized in table 4. Most embeddings' performance is slightly improved, with t-SNE and PHATE showing the largest improvement. The proposed method's embedding shows borderline improvement in the classification accuracy, however with a slight decline in the evaluation of the embedding's geometric consistency. As explained in the appendix E.2, the discrepancies may result from the fact that the RNA-seq data points form high dimensional clusters, globally arranged in a lower dimensional frame.

E.6.2 Results on the FMNIST

Comparative results, analogous to the ones of the RNA-seq, are met in the evaluation of the computed embeddings' capability to discriminate the categories of FMNIST data set. Collective results of the performed 10-fold experiment are provided in Table 5, which is analogous to Table 2 that summarizes the diagnoses' discrimination in the RNA-seq data set. In Table 5, the mean $\pm$ standard deviation of the measures' values across the 10 independent experiments are registered. A more detailed picture of the evaluation of the computed embeddings across these experiments is provided in the box plots of figure 11. These results indicate that the proposed method and t-SNE offer embeddings that simultaneously preserve nearest neighbors and their distances and mark the top evaluation in the FMNIST classes' discrimination. UMAP and IsoMap, though succeed to preserve nearest neighbors better than tSNE, they slightly distort local distances, bringing closer neighbors of different classes. On the other hand, LLE though succeeds to preserve local distances better than tSNE,it

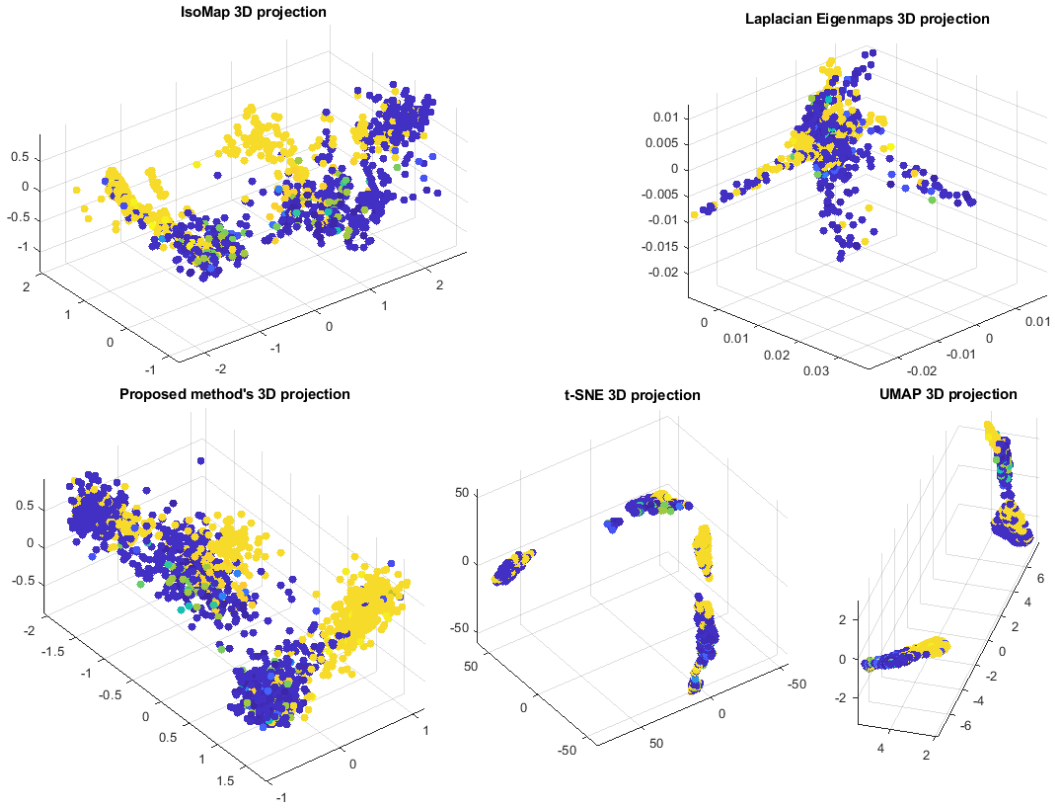

Figure 10: Visualization of the RNA-seq emedding. 3D projection of the computed 30-dimensional embeddings. Coloring is according to the lung cancer diagnosis. The proposed algorithm respects the global distances recovered by IsoMap and enhanses the data discrimination (table 2) by retaining logal structures, up to integrability.

Table 5: Evaluation of the clusters' representation for the FMNIST dataset

	FMNIST classification within embeddings' neighborhoods						
	LLE	LTSA	LplcMps	IsoMap	t-SNE	UMAP	proposed
ACC ↑	0.7120 ± 0.0091	0.6189 ± 0.0184	0.6967 ± 0.0125	0.7105 ± 0.0169	0.7458 ± 0.0082	0.7317 ± 0.0163	0.7525 ± 0.0130
NMI ↑	0.6290 ± 0.0101	0.5132 ± 0.0196	0.6127 ± 0.0120	0.6204 ± 0.0139	0.6643 ± 0.0076	0.6467 ± 0.0157	0.6595 ± 0.0093
ARI ↑	0.5280 ± 0.0124	0.3916 ± 0.0222	0.5082 ± 0.0187	0.5158 ± 0.0231	0.5715 ± 0.0132	0.5507 ± 0.0238	0.5721 ± 0.0176
	Local/global geometric consistency						
Lcl-F1 ↑	0.6780 ± 0.0067	0.2888 ± 0.0055	0.6755 ± 0.0057	0.7160 ± 0.0108	0.7064 ± 0.0045	0.7215 ± 0.0058	0.7852 ± 0.0091
Lcl-mtrc ↓	0.1429 ± 0.0041	0.1429 ± 0.0041	0.1429 ± 0.0041	0.1763 ± 0.0087	0.1390 ± 0.0040	0.1515 ± 0.0047	0.1194 ± 0.0042
Glbl-mtrc ↓	0.9993 ± <1e-4	0.9993 ± <1e-4	0.9999 ± <1e-4	0.1045 ± 0.0016	0.6703 ± 0.0054	0.9745 ± 0.0069	0.2638 ± 0.0189

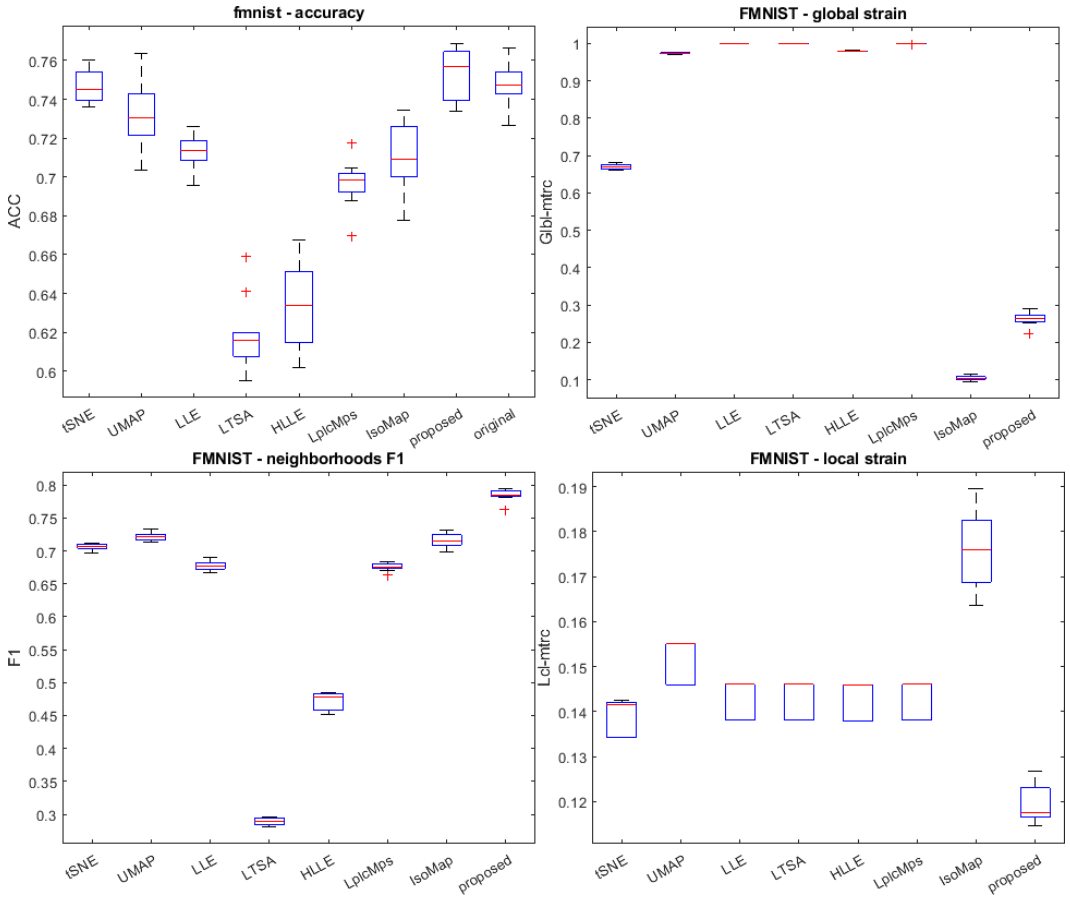

Figure 11: FMNIST dataset 10-fold experiment. The measures' box plots depict the maximum/minimum boundaries at whiskers, the boxes of the populations' 25th-75th percantile and the median values in red. Possible outliers are marked by red crosses. The proposed method best preserves neighboring relations and together with t-SNE, it best represents datapoints' clusters, slightly enhancing the clustering according to the "original" graph's neighborhoods. The performance of UMAP and IsoMap that preserve neighborhoods but not local distances, indicates that the original neighborhoods contain data points of mixed labels, however ordered according to data points' distances

imposes more neighborhood violations, bringing outliers of different classes in distances comparative to the ones of the neighbors of the same class. Again, it should be noted that for the proposed method, the preservation of the data local structure that facilitates discriminability is combined with a faithful, approximately isometric reconstruction of their global structures, since the proposed method is the only one achieving global stress, comparative to the minimal one of IsoMap.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the Introduction state the paper's scope: a variational, representation-free method for approximate Euclidean embedding of data and its distance graph-based implementation (Sections 2–4). The theoretical claims are supported by Sections 2–4 and Appendices A–D; Introduction's claim on convergence to the set of stationary Euclidean embeddings is consistent with Theorem 1: asymptotic embedding stationarity is established unconditionally, whereas the accumulation-point existence relies on the compactness and almost-everywhere full-rank assumptions stated in Section 3.2 and Appendix D.The empirical claims summarize the method's evaluation on synthetic manifolds and real datasets, provided in Section 5 and Appendix E.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the main limitations: reliance on symmetric local neighbor relations, sensitivity of graph differential-operator estimates to outliers, computational cost of the iterative sparse solves and the non-parametric nature of out-of-sample extension. Also, Sections 5 and Appendices E.2–E.6 examine method's sensitivity to neighborhood size, intrinsic dimension and the absence of smooth low-dimensional Euclidean isometries.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Sections 2–3 state the continuum variational formulation and the Euler–Lagrange stationarity equations, whose derivation is provided in Appendix C. Appendix A provides the differential 1-form identities, necessary in the proposed method's derivations and implementation. The graph realizations of the differential operators used by the implemented algorithm are derived in Appendix B. Appendix D proves approximation error descent and vanishing successive changes of the embedding's differential, characterizes asymptotic Euler–Lagrange stationarity and gives the conditional LaSalle limit-set result, under compactness and almost-everywhere full rank of the embedding differentials; the assumptions are explicitly stated in Theorem 1.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sections 4 and 5 describe the algorithmic steps, graph construction, intrinsic-dimension estimation, convergence tolerance, evaluation metrics, datasets and baseline methods. The experiments use standard synthetic manifolds and publicly available real datasets and the baseline implementations are identified by citation. Appendix E.1 reports the timing protocol and computing environment and E.2–E.6 specify intrinsic-dimension choices, metric definitions, and additional results. The method's public MATLAB repository and its linked experimental-results/data archive provide scripts, processed data documentation and the complete iteration-timing records for the method's timing experiment.

Guidelines:

- The answer [Yes] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification:The MATLAB implementation and experiment scripts are publicly accessible in the EucApprox repository (`https://github.com/DArabadjis/EucApprox`). Its README links the experimental data/results archive, including MNIST/FMNIST test-set material and the processed open-access GDC lung RNA-seq data and documents the dependencies and main scripts.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 specifies the construction of the common distance graph, the intrinsic-dimension estimator, convergence tolerances, evaluation metrics, datasets, compared baselines and repeated-sampling protocols. Section 4 gives the implementation details of the proposed graph algorithm. Appendix E.1 documents the execution-time experiment and parallelization rule, while E.2–E.6 define the evaluation metrics and provide sensitivity and additional experimental results. Raw timing records are supplied in the "experimental results" folder of the linked archive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:The neighborhood-size experiments use 15 independent random samplings per setting, with box plots for the difficult manifold and Swiss roll (Figures 1 and 7). The MNIST and FMNIST test sets are each partitioned into ten disjoint groups; Figures 2 and 11 show their evaluation distributions. Table 5 reports FMNIST means and standard deviations across those ten groups. For the RNA-seq experiment, results are reported as deterministic evaluations on fixed datasets and graph constructions, rather than as repeated stochastic trials.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E.1 specifies the timing platform (AMD EPYC 9655P CPU, 128 GB RAM), software (MATLAB R2025b and the Parallel Computing Toolbox), thread-based execution and the condition for embedding's component-wise parallelization. It reports actual per-iteration wall-clock measurements versus dataset size and embedding dimension (Figure 3). Complete timing records for that experiment are available in the "experimental results" folder of the linked archive, permitting reconstruction of the per-run execution times. Overall solver time additionally depends on the number of iterations to the convergence target, which is data- and problem-dependent; per-iteration measurements are therefore used to evaluate algorithmic scaling.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work introduces a general geometric embedding method and evaluates it on synthetic data, standard image benchmarks and previously collected open-access NCI GDC data. It does not involve participant recruitment, intervention, new collection of private patient data or deployment of high-risk models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The paper is a foundational work on representation-free geometric embedding and does not target a specific deployed application. Method's application to lung RNA-seq and image benchmarks is only methodological. Potential downstream impacts would depend on the domain in which the embedding is applied, especially if used for clustering, visualization, or analysis of sensitive data, but the current manuscript does not include a dedicated broader-impact discussion.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk pretrained models, generative systems, scraped datasets or other assets with substantial dual-use or misuse risk. The biological experiment uses previously collected open-access GDC data.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Original publications and implementations for the baselines and datasets are credited in the paper and the documentation in the method's repository. The repository records attribution, provenance and applicable license notices for reused or redistributed assets, including DRToolbox-related routines, HYPRE, Fashion-MNIST and the GDC data source. PHATE is an external MATLAB baseline: its implementation is cited separately from the original PHATE paper and neither its code nor its binaries are redistributed. Third-party rights and the license of the author's own code are distinguished in the method's repository notices.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer :[Yes]

    Justification: The method's MATLAB implementation is released at `https://github.com/DArabadjis/EucApprox`, with a README describing algorithm entry points, experiments, dependencies and the separately downloadable results/data archive. The repository includes citation information, copyright and third-party notices and documentation of the author's processed lung RNA-seq dataset. No trained predictive model is released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing, user study, participant recruitment or direct human-subject experiment was performed. The lung RNA-seq evaluation uses previously collected open-access NCI GDC records.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study does not recruit or intervene with human subjects or collect new individual-level clinical data. The lung RNA-seq experiment uses previously released open-access GDC data.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs or any other automation tool is not exploited at any stage of the development or the programming of the proposed methodology. The technical derivations and the implementation of the resulting method are made manually by the author.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.