

École polytechnique de Louvain

Quality assessment of dimensionality reduction: which criterion for which purpose?

Author: **Lina BOURAS**
Supervisor: **John LEE**
Readers: **Edouard COUPLET, Michel VERLEYSEN**
Academic year 2025–2026
Master [120] in Data Sciences Engineering

Contents

1	Introduction	1
2	Embedding techniques	3
2.1	Principal Component Analysis	3
2.1.1	Theoretical Foundations: The Dual Perspectives	3
2.1.2	Mathematical Formulation and Steps	3
2.1.3	Applications and Interpretability	4
2.1.4	Limitations and Extensions	4
2.2	Metric Multidimensional Scaling	5
2.2.1	Theoretical Foundation: Stress Minimization	5
2.2.2	Algorithmic Steps	5
2.2.3	Weighting Schemes and Variants	5
2.2.4	Limitations and the Curse of Dimensionality	6
2.3	Isometric Feature Mapping	7
2.3.1	Theoretical Foundation: Geodesic Distances and Graph Approximation	7
2.3.2	The algorithm steps	7
2.3.3	Hyper-parameters	8
2.3.4	Strength and limitations	8
2.4	Uniform Manifold Approximation and Projection	9
2.4.1	Theoretical Foundations: Riemannian Geometry and Fuzzy Topology	9
2.4.2	The Algorithm: A Two-Phase Process	9
2.4.3	Hyper-parameters	10
2.4.4	Strengths and Limitations	10
2.5	Stochastic neighbour Embedding and Fast Fourier Acceleration	11
2.5.1	The Theoretical Foundation: SNE and t-SNE	11
2.5.2	The Need for Multi-scale Preservation	12
2.5.3	The Challenge of Accelerating Multi-scale Embeddings	12
2.5.4	The Solution: Fast Multi-scale neighbour Embedding (FM _s TSNE and FM _s SNE)	12
3	Quality metrics	14
3.1	Silhouette Coefficient	14
3.1.1	Theoretical Foundation: Tightness and Separation	14
3.1.2	Mathematical Formulation	14
3.1.3	Methodology: Comparing HD and LD Profiles	15
3.1.4	Strength and limitations	15
3.2	Pearson correlation between dimensions	16
3.2.1	Theoretical Foundation: Standardized Covariance	16
3.2.2	Application	16
3.2.3	Spearsman's correlation ρ	16
3.3	Rank-based Criteria: The $R_{NX}(K)$ and Co-ranking Matrix Framework	18

3.3.1	Theoretical Foundation: The Co-ranking Matrix, $Q_{NX}(K)$	18
3.3.2	Limitation of $Q_{NX}(k)$ and the Need for Rescaling	18
3.3.3	The Solution: The $R_{NX}(k)$ Metric	18
3.4	ZADU: A Comprehensive Tool for Evaluating Distortions	20
3.4.1	Theoretical foundations: Structural levels of Distortions	20
3.4.2	Local Measures: neighbourhood and Rank-Based Topology	20
3.4.3	Cluster-Level Measures: Structural Fidelity and Validation	20
3.4.4	Global Measures: Geometric and Statistical Consistency	21
3.4.5	Visual Analytics Integration	21
4	Results and analysis	23
4.1	Experimental setup	23
4.1.1	Pre-processing	23
4.1.2	Datasets description	24
4.1.3	Summary	25
4.1.4	Framework	25
4.1.5	Packages	25
4.2	Visual Analysis of the Embeddings	26
4.2.1	Evaluation	26
4.2.2	COIL-20 Dataset	27
4.2.3	MNIST Dataset	30
4.2.4	Phoneme Dataset	33
4.2.5	Isolet Dataset	36
4.2.6	CoverType Dataset	39
4.2.7	PBMC3K Dataset	42
4.3	Quantitative Metric Analysis	45
4.3.1	Evaluation of cluster integrity metrics	45
4.3.2	$R_{NX}(K)$ curves and AUC:	46
4.4	Concordance Analysis: $R_{NX}(K)$ vs. ZADU Metrics	49
4.4.1	Correlation analysis	49
4.4.2	Comprehensive Discussion Across All Datasets	49
4.5	Concordance Analysis of Embedding method and $R_{NX}(K)$ per ZADU criteria	54
4.5.1	Overview of the Data Structure	54
4.5.2	Local measures	54
4.5.3	Cluster level measures	56
4.5.4	Global measures	58
5	Conclusion	61
5.0.1	Use of Generative AI tools	62
	Bibliography	63

Abstract

Dimensionality reduction (DR) is essential for visualising and interpreting high-dimensional data, yet every embedding introduces distortions. This study provides a systematic evaluation of six DR methods (PCA, MDS, Isomap, UMAP, FMsSNE, and FMsTSNE) across six datasets with diverse structures (COIL-20, MNIST, Phoneme, Isolet, CoverType, and PBMC3K). The analysis combines visual diagnostics (ZADUVis), quantitative metrics (silhouette, label correlation, rank-based criteria $R_{NX}(K)$ and AUC), and the 17-measure ZADU framework to characterise local, cluster-level and global fidelity. A concordance study quantifies how neighbourhood preservation at different scales relates to specific distortion types.

Results show that no DR method is universally optimal: each exhibits strengths and weaknesses depending on data topology and preservation goals. The study highlights the risk of visually convincing but structurally misleading embeddings and demonstrates that reliable DR assessment requires integrating visual inspection with scale-aware rank-based and cluster-level metrics. Together, these findings provide a clearer understanding of how DR methods behave across manifold types and offer practical guidance for selecting an embedding strategy aligned with the structural properties of the data.

Chapter 1

Introduction

Introduction

In an age where data has become the engine of scientific progress, researchers are increasingly confronted with information that is vast, intricate, and overwhelmingly high-dimensional. The challenge is no longer collecting information but understanding it. Across disciplines as varied as image processing, genomics, neuroscience, or the social sciences, modern datasets are often of unprecedented scale (sometimes millions of variables), complexity, and dimensionality. Making sense of such data is far from trivial: patterns become obscured, distances lose meaning, and traditional analytical tools begin to break down under the weight of dimensionality.

To navigate this complexity, dimensionality reduction (DR) has emerged as a cornerstone technique. By projecting high-dimensional (HD) data into more manageable low-dimensional (LD) spaces, typically two or three dimensions, DR enables intuitive visual exploration, reveals latent structure, and mitigates the notorious “curse of dimensionality” that hampers many machine-learning algorithms. In essence, DR transforms overwhelming data landscapes into interpretable geometric representations, making hidden relationships visible and actionable.

However, the transition from a high-dimensional manifold to a plane is inherently imperfect, as different algorithms prioritize different aspects of the data structure. This work addresses the critical challenge of evaluating the reliability of these projections. While linear methods like Principal Component Analysis (PCA) excel at capturing global variance, they often fail to resolve complex non-linear structures. Conversely, neighbourhood-based methods like Uniform Manifold Approximation and Projection (UMAP) and Stochastic neighbour Embedding (SNE) provide superior local cluster separation but may introduce distortions such as the “crowding problem” [15] or lose global geometric context.

Consequently, several fundamental questions arise: How do single-scale embeddings compare to multi-scale approaches in resolving diverse manifolds? Can we rigorously quantify structural distortions using a multi-perspective framework that bridges local, global, and cluster-level fidelity? And finally, is there a statistical concordance between established rank-based quality curves (R_{NX}) and modern pointwise distortion measures?

This work contributes to the field of dimensionality reduction through a systematic and multi-perspective evaluation of embedding reliability. Specifically, it provides a rigorous comparative analysis between traditional linear baselines like PCA, iterative non-linear methods such as Metric MDS, geodesic-preserving techniques such as Isomap, and modern multi-scale neighbour embeddings including UMAP, and advanced multi-scale neighbour embeddings like FMsTSNE. Central to this work is the application of a comprehensive quality framework that categorizes seventeen distortion measures into local fidelity, global preservation, and cluster integrity levels to identify

structural discrepancies. Furthermore, this study leverages advanced visual diagnostics through the ZADUVis module, utilizing CheckViz Voronoi cells and Reliability Maps to precisely localize and interpret geometric failures such as manifold tearing or crushing within the latent space, comparison between clusters in LD and HD and ground truth label distribution. Finally, the study introduces a novel concordance analysis that quantifies the statistical agreement between the rank-based ($R_{NX}(K)$) curves and the ZADU metric suite along two complementary axes. In the first axis, we calculate Pearson correlations for each dataset between $R_{NX}(K)$ (at every neighbourhood size K) and each ZADU measure across the six dimensionality reduction methods (PCA, UMAP, MDS, Isomap, FMsSNE, and FMsTSNE). This analysis identifies which distortion measures are effectively tracked by R_{NX} at local, medium or macroscopic scales, while exposing method-specific deviation. In the second axis, we compute these correlations for each DR method across the six datasets. This allows us to assess the robustness and consistency of a specific algorithm's R_{NX} curve, determining how reliably it aligns with specific structural distortions across a variety of different data topologies.

Chapter 2

Embedding techniques

The landscape of dimensionality reduction encompasses a wide spectrum of mathematical approaches, each rooted in distinct theoretical principles ranging from linear variance-maximization techniques to non-linear methods that preserve geometric or topological structure. In this chapter, we introduce several representative algorithms: Principal Component Analysis, Metric Multidimensional Scaling, Isometric Feature Mapping, Uniform Manifold Approximation and Projection, and Fast Multi-scale Stochastic neighbour Embeddings.

2.1 Principal Component Analysis

Principal Component Analysis (PCA) serves as the fundamental linear approach to dimensionality reduction. The core objective of PCA is to transform a high-dimensional dataset into a lower-dimensional space while preserving the most significant patterns of variability.

2.1.1 Theoretical Foundations: The Dual Perspectives

The mathematical beauty of PCA lies in the fact that it can be derived from two seemingly different but equivalent optimization problems: maximizing the variance of the projected data and minimizing the mean squared reconstruction error.[11]

- **Maximum Variance Perspective:** From this viewpoint, we seek a set of orthogonal axes—known as principal components—such that the variance of the data coordinates when projected onto these axes is maximized. The first principal component is the direction in the data space along which the observations show the greatest spread. Each subsequent component must be orthogonal to the previous ones and capture the maximum remaining variance.
- **Minimum Reconstruction Error Perspective:** PCA also finds the best low-rank approximation to the centred data matrix $\mathbf{X}^*$. For a target dimension $q < \text{rank}(\mathbf{X}^*)$, the optimal rank- q matrix $\mathbf{X}_q^*$ that minimises the sum of squared Euclidean distances between the original $\mathbf{X}$ points and their orthogonal projections onto a q -dimensional subspace $\mathbf{Y}$ is given by the singular value decomposition (SVD) of $\mathbf{X}^*$. The rows of $\mathbf{X}_q^*$ are the projections, and the residual sum of squares equals the sum of the discarded eigenvalues. This geometric interpretation dates to Pearson (1901).

2.1.2 Mathematical Formulation and Steps

The standard procedure for computing PCA involves several rigorous mathematical steps: [11][10]

1. **Data Centering:** A critical first step is to center the data by subtracting the mean of each variable. If variables have different units or scales, they are often standardized (divided by their standard deviation) to prevent variables with large scales from dominating the analysis.
2. **The Covariance Matrix:** Let $\mathbf{X}^*$ be the centered $N \times M$ data matrix. PCA performs an eigendecomposition of the sample covariance matrix $\mathbf{S} = \frac{1}{N-1}\{\mathbf{X}^{*T}\mathbf{X}^*\}$. This matrix $\mathbf{S}$ captures all linear dependencies between pairs of variables.
3. **Eigendecomposition:** The spectral decomposition of $\mathbf{S}$ yields a set of eigenvalues λ_k and their corresponding eigenvectors v_k . The eigenvalues represent the amount of variance captured by each eigenvector.
4. **Projection:** To reduce the data to P dimensions, we select the P eigenvectors associated with the largest eigenvalues to form a projection matrix V_P . The low-dimensional coordinates are then computed as $\mathbf{Y} = \mathbf{X}\mathbf{V}_P$.

Alternatively, modern implementations often use Singular Value Decomposition (SVD) directly on the data matrix $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. In this framework, the right singular vectors V provide the principal components, and the singular values in $\mathbf{\Sigma}$ are related to the standard deviations of the data along these axes.[10]

Components are derived by maximizing the remaining variance while enforcing orthogonality with previously extracted components. The first principal component is defined by the weight vector $\mathbf{v}_1$ that maximizes:[11]

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|=1} \text{Var}(\mathbf{X}\mathbf{v}) = \arg \max_{\|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{S} \mathbf{v} \quad (2.1)$$

2.1.3 Applications and Interpretability

PCA is utilized in two primary contexts: preprocessing and visualization. In preprocessing, PCA is often used to reduce data from thousands of features to remove statistical noise and alleviate the "curse of dimensionality" for downstream tasks like clustering or classification. In visualization, the first two or three principal components are plotted to provide an interpretable 'map' of the data.

2.1.4 Limitations and Extensions

As a linear method, PCA is restricted to capturing global, linear structures. It often fails to resolve complex non-linear manifolds, such as the "Swiss Roll" structure, where local neighbourhood relationships are more important than global variance. In these cases, PCA may project distant points into the same region of the embedding, causing "crowding" and loss of structural detail.

Furthermore, PCA is highly sensitive to outliers because it relies on squared errors. To address these weaknesses, several extensions have been developed, including Kernel PCA (which uses the "kernel trick" to perform PCA in a non-linear feature space), Sparse PCA (for better interpretability), and Robust PCA. Despite its age, PCA remains a vital baseline; for instance, classical metric Multidimensional Scaling (MDS) applied to Euclidean distances is mathematically equivalent to PCA, highlighting its role as the foundation for much of the field. [11][10]

2.2 Metric Multidimensional Scaling

Multidimensional Scaling (MDS) encompasses a family of dimensionality reduction techniques that aim to reproduce, in a low-dimensional representation, the pairwise proximities (similarities or dissimilarities) observed in high-dimensional data. Unlike PCA, which focuses on preserving variance between variables, MDS operates directly on the relationships between observations. While some variants are linear, the broader application of MDS through iterative optimization classifies it as a fundamental non-linear embedding approach. [10]

2.2.1 Theoretical Foundation: Stress Minimization

While Classical MDS relies on spectral decomposition (eigenvalues/eigenvectors), Metric MDS (mMDS)—also known as stress-based MDS—generalized the approach in the 1960s to allow for non-linear mappings. The core principle of mMDS is to find a configuration of points $Y = (y_1, y_2, \dots, y_N)$ in a low-dimensional space such that the Euclidean distances between them ($\|y_i - y_j\|$) approximate the original high-dimensional dissimilarities (d_{ij}) as closely as possible. This approximation is achieved by minimizing an objective function called Stress (σ), introduced by Kruskal in 1964:

$$\sigma(Y) = \sum_{i < j} w_{ij} (d_{ij} - \|y_i - y_j\|)^2 \quad (2.2)$$

Where w_{ij} are weights that can be used to modulate the importance of specific pairs. Because the stress function involves the Euclidean distance (the square root of a sum of squares), its optimization is non-convex and typically requires iterative gradient-based algorithms rather than a direct spectral solution.[10] [4]

2.2.2 Algorithmic Steps

The implementation of mMDS generally follows these steps: [4] [10]

1. **Proximity Matrix Construction:** Compute (or provide) an $N \times N$ matrix of dissimilarities D between all pairs of high-dimensional points.
2. **Initialization:** Assign initial coordinates to the points in the low-dimensional space, often using the results of Classical MDS to provide a reasonable starting configuration.
3. **Stress Evaluation:** Calculate the current stress value using the formula above to measure the discrepancy between high-dimensional and low-dimensional distances.
4. **Optimization and convergence:** Update the low-dimensional coordinates iteratively (e.g., using gradient descent) to reduce the stress. Stop the algorithm the change in stress falls below a predefined threshold.

2.2.3 Weighting Schemes and Variants

A significant advantage of mMDS over PCA is its flexibility through weighting schemes. By adjusting w_{ij} , practitioners can prioritize the preservation of certain distances over others. [10]

- **Sammon Mapping:** Introduced by Sammon in 1969, this variant uses weights $w_{ij} = 1/d_{ij}$. This explicitly grants more importance to preserving short distances, making it more effective at resolving local cluster structures than standard MDS.
- **Curvilinear Component Analysis (CCA):** This method further refines the weighting by making it dependent on the distances in the low-dimensional space, allowing the manifold to "unfold" more effectively by tolerating stretching of long distances.

2.2.4 Limitations and the Curse of Dimensionality

Despite its theoretical elegance, mMDS faces a major challenge known as norm concentration. In very high-dimensional spaces, the "curse of dimensionality" causes the distances between points to become increasingly similar (the ratio between the farthest and nearest points approaches 1). When mMDS tries to preserve these nearly uniform distances in a 2D or 3D space, the resulting embedding often appears as a featureless "cloud" or disc, failing to reveal meaningful structures like `Clusters_kmeans` or manifolds. This limitation eventually led to the development of neighbourhood-based methods like SNE and UMAP, which prioritize local rankings over absolute distance values.

2.3 Isometric Feature Mapping

Isometric Feature Mapping (Isomap) is a non-linear dimensionality reduction method that extends classical Multidimensional Scaling (MDS) by replacing Euclidean distances with geodesic distances measured along the data manifold. This allows Isomap to unfold curved, low-dimensional structures embedded in a high-dimensional space, such as a “Swiss roll” or a set of face images taken under varying poses.

2.3.1 Theoretical Foundation: Geodesic Distances and Graph Approximation

The core motivation behind Isomap is to uncover the true, low-dimensional structure hidden within complex, high-dimensional data. Traditional techniques like Principal Component Analysis (PCA) or standard Multidimensional Scaling (MDS) evaluate relationships by measuring straight-line Euclidean distances between points. However, if the data lies on a curved, non-linear surface (a manifold), a straight line might cut through empty space, entirely misrepresenting how the points are actually related.

Isomap solves this by measuring “geodesic distance”, the distance traveled strictly along the curved surface of the data manifold. By doing this, it successfully “unrolls” complex structures (like the famous Swiss roll example) into a flat, interpretable space. [21]

2.3.2 The algorithm steps

Isomap computes the lower-dimensional mapping through a strict, three-step procedure:[21]

- **Step 1: Constructing the Local Graph.** The algorithm starts by identifying which data points are immediate neighbours. It assumes that on a very small, local scale, straight-line distances are a safe approximation of the curved manifold. It connects neighbouring points to form a web or graph, while leaving distant points unconnected. Mathematically, the distance between neighbouring points i and j in the original high-dimensional space is defined by the standard Euclidean metric $d_X(i, j)$, subject to either a K -nearest neighbour or a fixed ϵ -radius constraint.
- **Step 2: Estimating Geodesic Distances.** To figure out the distance between any two points on the manifold, Isomap calculates the shortest path through the web created. Using standard graph search algorithms (like Dijkstra’s or Floyd’s), it hops from neighbour to neighbour, accumulating the distances to estimate the true global path along the surface. This is achieved by initializing graph distances $D_G(i, j) = d_X(i, j)$ for connected neighbours, and iteratively updating them using the shortest-path equation:

$$D_G(i, j) = \min\{D_G(i, j), D_G(i, k) + D_G(k, j)\} \quad (2.3)$$

- **Step 3: Dimensionality Reduction via MDS.** Finally, Isomap takes this newly calculated matrix of shortest-path distances and feeds it into classical MDS. MDS mathematically translates these distances into a low-dimensional coordinate system (e.g., 2D or 3D) that best preserves the distances measured in Step 2. This optimization minimizes the cost function E , which measures the L^2 matrix norm difference between the graph distances and the new lower-dimensional Euclidean distances D_Y :

$$E = \|\tau(D_G) - \tau(D_Y)\|_{L^2}. \quad (2.4)$$

where τ represents the mathematical operator that converts these distance matrices into inner-product (Gram) matrices.

2.3.3 Hyper-parameters

The effectiveness of Isomap is highly sensitive to the construction of its neighbourhood graph. To define which points qualify as neighbours, Isomap offers two main strategies for doing so: **K -Isomap** and **ϵ -Isomap**. K -Isomap consists of connecting a point to its K nearest neighbours while ϵ -Isomap connects a point to all other points within a fixed radius of ϵ .

The chosen parameter size is highly sensitive. If K or ϵ is set too small, the graph might fracture into disconnected islands, making it impossible to compute global paths. Conversely, if the parameter is set too large, the algorithm might mistakenly connect points that are far apart on the manifold but physically close in space. This creates "short circuits" that ruin the geodesic distance calculations. [21]

2.3.4 Strength and limitations

Isomap's primary advantage is that it combines the ability to handle highly non-linear data with the mathematical stability of linear tools. Unlike some artificial neural networks or iterative optimization techniques that can get trapped in "local minima" (suboptimal solutions), Isomap guarantees a globally optimal solution because it relies on standard eigenvalue math (via MDS) and shortest-path algorithms. It is also non-iterative and computationally predictable.[21]

Isomap operates on the strict assumption that the data manifold is well-sampled and continuous. If the dataset has massive gaps, is extremely noisy, or features multiple disconnected structures, the shortest-path graph will fail to accurately represent the surface geometry. It struggles if the intrinsic curvature is too extreme compared to the density of the data points.[21]

2.4 Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) is a manifold learning technique for dimension reduction constructed from a theoretical framework based on Riemannian geometry and algebraic topology. UMAP's design decisions are grounded in a solid theoretic foundation, specifically leveraging category-theoretic approaches to the geometric realization of fuzzy simplicial sets.[4] [16]

2.4.1 Theoretical Foundations: Riemannian Geometry and Fuzzy Topology

The UMAP algorithm is built upon three core assumptions about the structure of the data. First, it assumes the existence of an underlying manifold on which the data are uniformly distributed. Second, this manifold is presumed to be locally connected, ensuring that nearby points reflect meaningful local relationships. Finally, UMAP is designed with the objective of preserving the topological structure of this manifold as faithfully as possible, guiding the construction of its low-dimensional embedding. To address the reality that finite real-world data is rarely uniformly distributed, UMAP assumes that the manifold possesses a Riemannian metric under which the data is approximately uniform. This leads to the creation of a custom local distance for each point X_i , normalizing distances based on the distance to its k^{th} nearest neighbour. This process effectively patches together incompatible local metric spaces into a consistent global structure using fuzzy simplicial sets, a topological structure where connections between points are not binary but probabilistic. [16] [4]

2.4.2 The Algorithm: A Two-Phase Process

The UMAP algorithm proceeds in two primary phases: the construction of a high-dimensional fuzzy topological representation and the optimization of a low-dimensional layout that mirrors this topology.

Phase 1: High-Dimensional Graph Construction

UMAP first identifies the local neighbourhood for each point using a weighted k-nearest neighbour (kNN) graph. For each point x_i , UMAP defines a local distance scale σ_i and a connectivity parameter ρ_i (the distance to the nearest neighbour). The similarity v_{ij} between points i and j in the high-dimensional space is defined as:

$$v_{ij} = \exp \left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i} \right) \quad (2.5)$$

This formula ensures that every point is connected to at least its nearest neighbour with a weight of 1, effectively "normalizing" local density variations. To create a globally consistent structure, UMAP symmetrizes these weights, ensuring that the relationship between i and j is the same as between j and i . [16]

Phase 2: Low-Dimensional Optimization

In the second phase, UMAP initializes a low-dimensional layout, often using Laplacian Eigenmaps (a spectral method) to provide a global initialization that reflects the manifold's coarse geometry. The optimization seeks to minimize the fuzzy set cross-entropy (C) between the high-dimensional

fuzzy graph (v) and the low-dimensional equivalent (w): [16]

$$C = \sum_{i \neq j} \left[v_{ij} \log \left(\frac{v_{ij}}{w_{ij}} \right) + (1 - v_{ij}) \log \left(\frac{1 - v_{ij}}{1 - w_{ij}} \right) \right]. \quad (2.6)$$

Attraction-Repulsion Dynamics

A useful way to interpret UMAP’s behavior is through a physical “attraction–repulsion” analogy, in which the cross-entropy objective (Eq.2.6) combines an attraction term that pulls neighbouring points together in the embedding space and a repulsion term that pushes non-neighbours apart.

Compared to t-SNE, UMAP applies stronger attractive forces. This leads to more compact and pronounced Clusters_kmeans, making it highly effective for identifying distinct groups in datasets. [4]

2.4.3 Hyper-parameters

UMAP’s behavior is primarily governed by two hyper-parameters that balance the trade-off between local and global structure: [16]

- **Number of neighbours (n):** This controls the local scale at which the manifold is approximated as flat. Small values capture fine-grained detail, while large values capture the "big picture" of the manifold at the expense of local detail.
- **Minimum Distance (min-dist):** An aesthetic parameter that determines how closely points can be packed in the low-dimensional representation. Low values faithfully represent manifold structure but may lead to overplotting, while higher values spread points out for better visualization.

2.4.4 Strengths and Limitations

UMAP’s primary strengths lie in its superior computational scalability and its ability to preserve global structure through a cross-entropy objective that explicitly pushes dissimilar points apart. Grounded in category theory, it scales near-linearly, making it efficient for millions of samples where methods like t-SNE struggle. However, its non-linear nature means axes lack direct interpretability, and the assumption of manifold structure can lead to spurious "constellations" in noisy data. Furthermore, as a single-scale method, it may fail to resolve nested hierarchical structures better captured by multi-scale approaches. Finally, distances between its characteristic compact clusters are not strictly metric, precluding absolute comparisons in the low-dimensional space.

2.5 Stochastic neighbour Embedding and Fast Fourier Acceleration

Stochastic neighbour Embedding (SNE) is a non-linear dimensionality-reduction technique designed to preserve local neighbourhood relationships by matching probability distributions defined in the high- and low-dimensional spaces.

2.5.1 The Theoretical Foundation: SNE and t-SNE

Stochastic neighbour Embedding

The fundamental principle of Stochastic neighbour Embedding (SNE) is to convert pairwise distances into probability distributions that represent neighbourhood similarities.

For a dataset of N points in a high-dimensional space, SNE defines the conditional probability $p_{j|i}$ that point $\mathbf{x}_i$ would pick $\mathbf{x}_j$ as its neighbour as:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}. \quad (2.7)$$

Where the variance σ_i^2 is adapted to ensure the distribution matches a user-specified perplexity, K , which acts as a soft measure of the effective number of local neighbours.

While groundbreaking, standard SNE suffers from the *crowding problem*. This phenomenon describes the geometric difficulty of mapping data from large, high-dimensional volumes into exponentially smaller low-dimensional spaces, causing distinct clusters to crush together in the center of the embedding. [8] [15]

t-Distributed Stochastic neighbour Embedding

t-Distributed Stochastic neighbour Embedding (t-SNE) was introduced specifically to circumvent this crowding problem. Instead of modeling the LD space with a Gaussian distribution, t-SNE uses a heavy-tailed Student t-distribution with one degree of freedom to define the low-dimensional similarities q_{ij} :

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}} \quad (2.8)$$

where $\mathbf{y}_i$ and $\mathbf{y}_j$ represent the coordinates of the data points in the low-dimensional embedding. This heavy tail allows moderately distant HD points to be pushed much further apart in the LD space, leaving sufficient room for local clusters to form cleanly. Furthermore, t-SNE symmetrizes the HD similarities ($p_{ij} = (p_{j|i} + p_{i|j})/2N$) and minimizes a symmetric KL divergence cost function. Despite its excellent visualization quality, standard t-SNE requires evaluating all pairwise similarities, resulting in an unaffordable $O(N^2)$ computational time complexity for large datasets.[3] [15]

Kullback-Leibler divergence

To optimize the low-dimensional coordinates $\mathbf{y}_i$, the algorithm minimizes the Kullback-Leibler (KL) divergence between the probability distributions of the high-dimensional space (P) and the low-dimensional space (Q). The overall cost function C is mathematically defined as:

$$C = KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (2.9)$$

Because the KL divergence is asymmetric, this specific cost function heavily penalizes placing widely separated low-dimensional points to represent close high-dimensional neighbours (i.e., when

p_{ij} is large but q_{ij} is small). This asymmetry perfectly aligns with the algorithm’s primary goal: preserving local neighbourhood structures above all else.

2.5.2 The Need for Multi-scale Preservation

A structural limitation of both SNE and t-SNE is their "single-scale" nature. Because they rely on a single, user-defined perplexity parameter, the neighbourhood probability distributions forcefully focus on a fixed local resolution. Most HD similarities to distant points become infinitesimal and are ignored by the cost function, meaning the algorithm beautifully captures local interactions but largely neglects mid-range and global data structures.

To resolve this, Multi-scale SNE and Multi-scale t-SNE were developed. Instead of a single perplexity, multi-scale methods systematically combine neighbourhood probabilities tuned across a wide hierarchy of data scales. They utilize exponentially growing perplexities (e.g., $K_h = 2^{h-1}K_1$) spanning from a small local base up to the macroscopic scale of the entire dataset ($H = \lfloor \log_2(N/K_1) \rfloor$). The algorithm simply averages these single-scale similarities:

$$P_{ij} = \frac{1}{H} \sum_{h=1}^H p_{ij}^{(h)}. \quad (2.10)$$

This meta-parameter-free approach successfully retrieves both fine-grained local topologies and global geometric trends simultaneously. However, this multi-scale summation introduces an extra logarithmic factor, pushing the computational complexity to $O(N^2 \log N)$, making it completely intractable for big data. [3]

2.5.3 The Challenge of Accelerating Multi-scale Embeddings

To accelerate single-scale t-SNE to $O(N \log N)$ time, researchers historically exploited the inherent sparsity of local HD similarities. Because single-scale probabilities vanish rapidly for distant points, they can be truncated to only the closest neighbours (typically $3K$ points). By using Vantage-Point (VP) trees to rapidly find these exact nearest neighbours, and a Barnes-Hut (BH) algorithm to approximate the repulsive LD forces using spatial quadtrees, single-scale t-SNE runs highly efficiently.[14]

Unfortunately, this exact sparsification trick fails entirely for multi-scale methods. Multi-scale HD similarities are intrinsically dense because the largest scale perplexities encompass the entire dataset ($O(N)$ neighbours). Because the method is designed to account for long-range global structures, truncating the distributions to only local neighbours would destroy the multi-scale theory.[3]

2.5.4 The Solution: Fast Multi-scale neighbour Embedding (FMsTSNE and FMsSNE)

To break this computational bottleneck, Fast Multi-scale neighbour Embedding introduces a highly sophisticated randomized subsampling acceleration that achieves $O(N \log^2 N)$ complexity without losing global structural integrity.

The core intuition is that while modeling fine-grained local details requires precisely identifying exact nearest neighbours, modeling macroscopic global trends only requires a rough, representative summary of the overall data cloud. FMsTSNE operationalizes this by defining a hierarchy of randomly subsampled HD datasets at every scale, denoted as $\mathcal{X}_h$. The base scale ($\mathcal{X}_1$) contains the full dataset, but each subsequent scale halves the number of points ($\lfloor 2^{1-h}N \rfloor$).

The algorithm follows these theoretical steps:

1. **Hierarchical Vantage-Point Trees:** Instead of searching the full dataset for global neighbours, VP trees are constructed on each subsampled dataset $\mathcal{X}_h$.
2. **Targeted neighbour Sets:** For each data point, the VP trees identify small, relevant neighbour sets $\mathcal{N}_h(i)$ across all scales. At the smallest scale, it accurately targets the exact local neighbours. As the scale grows, the subsampling forces the tree to retrieve points that are increasingly dispersed across the high-dimensional space, perfectly emulating a low-resolution global view.
3. **Sparse Multi-scale Similarities:** The extracted sets from all scales are merged into a single sparse subset $\mathcal{N}_h(i)$ of size $O(\log N)$ for each data point. The multi-scale HD similarities are only computed over these targeted subsets.
4. **Barnes-Hut LD Approximation:** Finally, to prevent "LD crushing", when low-dimensional points collapse onto each other due to the artificially sparse HD attractions, a modified Barnes-Hut algorithm evaluates the divergences between the sparse HD similarities and the continuous LD distributions.

By relying on randomized multi-scale subsets rather than dense $O(N)$ matrices, FM_sTSNE explicitly preserves both local fluctuations and long-range structural consensus in $O(N \log^2 N)$ time. Empirical testing demonstrates that this method statistically significantly outperforms non-accelerated t-SNE in terms of neighbourhood preservation, while running orders of magnitude faster, unlocking multi-scale manifold extraction for massive datasets. [3]

Chapter 3

Quality metrics

To move beyond subjective visual assessment, it is essential to rely on a robust suite of quantitative measures capable of mathematically detecting structural discrepancies between the original space and its low-dimensional embedding. These quality metrics provide objective evidence of how faithfully a method preserves neighbourhood relations, global geometry, and cluster structure. In this chapter, we develop four complementary evaluation tools: the silhouette coefficient, correlation-based measures, a rank-based criterion, and finally ZADU, a new comprehensive framework for distortion analysis.

3.1 Silhouette Coefficient

The Silhouette Coefficient is a classic internal clustering validation measure, first introduced by Peter Rousseeuw in 1987, that has become a fundamental tool for assessing the quality of dimensionality reduction embeddings. In the taxonomy of quality measurements, the Silhouette Coefficient belongs to the "Compact" semantic class, characterized by metrics that compare the arrangement of all data points in terms of distance, specifically evaluating the relationship between inter-cluster and intra-cluster distances.

3.1.1 Theoretical Foundation: Tightness and Separation

The construction of silhouettes assumes that the proximities (dissimilarities or similarities) are on a ratio scale, such as Euclidean distances. The core objective is to determine whether the clusters in a given partition are high-quality, meaning that the "within" dissimilarities are small compared to the "between" dissimilarities. Each object i in the dataset is assigned a value $s(i)$, which reflects how well it matches its assigned cluster. [20]

3.1.2 Mathematical Formulation

The calculation of the Silhouette Coefficient for a single data point i involves two primary distance-based components: [20]

1. Average Intra-cluster Distance ($a(i)$):

This value represents the average distance between point i and all other points in the same cluster, it represents the "tightness" or cohesion of the cluster around object i . If cluster A contains other objects besides i , $a(i)$ is defined as the average dissimilarity (noted $d(i, j)$ for the dissimilarity between points i and j) of i to all other objects in its own cluster A :

$$a(i) = \text{average dissimilarity of } i \text{ to all other objects of } A = \frac{1}{|A| - 1} \sum_{\substack{j \in A \\ j \neq i}} d(i, j) \quad (3.1)$$

A small $a(i)$ indicates that the point is well-assigned to its cluster.

2. Minimum Average Inter-cluster Distance ($b(i)$):

This value represents the average distance from point i to all points in the nearest neighbouring cluster. For any cluster C different from A , we first compute the average dissimilarity of i to all objects in C , denoted as $d(i, C)$. We then define $b(i)$ as the minimum of these average dissimilarities across all competing clusters:

$$b(i) = \min_{C \neq A} d(i, C). \quad (3.2)$$

A large $b(i)$ indicates that the point is far from the next closest cluster.

3. The Silhouette Coefficient ($s(i)$):

The final coefficient for point i is calculated by normalizing the difference between separation and cohesion:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \quad (3.3)$$

Result interpretation: The silhouette coefficient $s(i)$ provides a normalized measure of clustering quality ranging from -1 to $+1$. A value of $s(i) \approx 1$ indicates that a point is very well-clustered and positioned far from other groups. Conversely, $s(i) \approx 0$ suggests the point lies on the border between two clusters, while a score of $s(i) \approx -1$ signifies that the point is likely assigned to the wrong cluster.

The Global Silhouette Score for a dataset is simply the mean $s(i)$ across all N points.

3.1.3 Methodology: Comparing HD and LD Profiles

In this study, we compute the Silhouette score in both the high-dimensional input space (S_{HD}) and the two-dimensional output space (S_{LD}) using a shared set of cluster labels. The comparison between these two scores provides deep insight into the behaviour of the DR algorithm:

- $Sil_{LD} \approx Sil_{HD}$: Indicates a faithful preservation of cluster separability.
- $Sil_{LD} \ll Sil_{HD}$: Suggests the DR method has introduced significant Backward Projection Error (BPE), where distant HD points are forced together in LD.
- $Sil_{LD} \gg Sil_{HD}$: Often observed in UMAP or t-SNE due to "early exaggeration" or strong attractive forces. While this makes the visualization cleaner, it may indicate that the algorithm is "inventing" cluster separation that is less pronounced in the original data.

3.1.4 Strength and limitations

Silhouette is intuitive and does not require an explicit geometric center (unlike the Calinski-Harabasz index), making it compatible with non-convex clusters found in manifold learning. However, it is computationally heavy ($O(N^2)$) as it requires all-pairs distance calculations, although libraries like ZADU provide optimized implementations to mitigate this. Furthermore, it is sensitive to the "curse of dimensionality" in HD, where norm concentration can cause HD Silhouette scores to be very low even for structured data.

3.2 Pearson correlation between dimensions

The Pearson product-moment correlation coefficient (r) is the standard measure of the linear relationship between two continuous variables.

In the taxonomy of quality measurements, the Pearson correlation belongs to the "Compact" semantic class. Compact measurements compare the arrangement of all data points across the entire metric space by evaluating the preservation of pairwise distances or similarities.

The core theoretical motivation for using Pearson correlation in DR is to assess Global Preservation. A high-quality embedding should not only for local preservation but should also ensure that distant points remain distant, and that the relative proportions of these distances are maintained. Within this framework, Pearson correlation acts as a linear indicator of how well the output distances d_{ij} correlate with the input distances δ_{ij} . [16]

3.2.1 Theoretical Foundation: Standardized Covariance

The Pearson correlation coefficient r is mathematically defined as the ratio between the covariance of two variables and the product of their standard deviations. In the specific context of DR assessment, we compare a set of N pairwise distances or proximity values extracted from both spaces.

Given a set of paired observations (x_k, y_k) representing distances in HD and LD respectively:

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (3.4)$$

Where:

- n is the total number of paired distances (for N points, $n = \frac{N(N-1)}{2}$).
- x_k is the distance between two entities (points or cluster centroids) in the high-dimensional space.
- y_k is the corresponding distance in the low-dimensional embedding. $\bar{x}$ and $\bar{y}$ are the sample means of the high- and low-dimensional distance sets respectively.

The coefficient r ranges from -1 to 1, where $r = 1$ signifies a perfect linear relationship (perfect distance preservation), $r = 0$ indicates no linear relationship (structural collapse), and $r = -1$ suggests an inversion of the distances. [16]

3.2.2 Application

In this work, two measures of correlation are studied, one is the standard correlation of distance matrix between LD and HD in ZADU package, and the other is extended to Cluster-level Granularity, that will be called '*correletaion of labels*'. The objective is to verify that the clusters identified in the original high-dimensional space maintain their relative global organization in the low-dimensional visualization by matching labels between the two dimensions. This is particularly critical because visual analysis often relies on the proximity of clusters to infer biological or semantic relationships.

3.2.3 Spearsman's correlation ρ

The Pearson correlation coefficient, being inherently linear, is unable to identify strong but non-linear associations within the data. Patterns such as U-shaped, polynomial, or exponential relationships

are therefore misrepresented. In addition, Pearson's coefficient is highly sensitive to outliers as it relies on the mean and standard deviation of the row data, making it easy for extreme observations to skew the correlation estimate.

To overcome these limitations, Spearman's rank correlation coefficient, denoted as Spearman's ρ (rho), offers a powerful non-parametric alternative. The core of the measure lies in its method: raw data values are transformed into their relative ranks before performing any computation. This ranking process makes the measure far more resistant to outliers, since such an extreme value that would normally pull the Pearson's correlation merely receives a rank slightly higher than the next largest observation. Moreover, because this method does not impose assumptions about the form of the associations, it remains flexible and highly effective for capturing non-Linear relationships. [2]

3.3 Rank-based Criteria: The $R_{NX}(K)$ and Co-ranking Matrix Framework

Rank-based criteria evaluate the performance of a dimensionality reduction (DR) technique by comparing the neighbourhood of each point in the high-dimensional (HD) space with its neighbourhood in the low-dimensional (LD) space. A fundamental advantage of these criteria is their invariance to any monotonic transformation of distances, which ensures that the quality assessment is robust regardless of the specific metrics used or the non-linear warping applied by embeddings like t-SNE or UMAP. [13]

3.3.1 Theoretical Foundation: The Co-ranking Matrix, $Q_{NX}(K)$

The mathematical framework for rank-based quality assessment is built upon the comparison of proximity orderings in the high-dimensional (HD) and low-dimensional (LD) spaces. For each data point i , a rank ρ_{ij} is assigned to every other point j based on its sorted distance in the original space, with a corresponding rank r_{ij} determined within the embedding. These rankings are aggregated into a co-ranking matrix Q , where each element $Q_{kl} = |\{(i, j) : \rho_{ij} = k \text{ and } r_{ij} = l\}|$ serves as a joint histogram of HD and LD ranks. From this matrix, the mean relative rank agreement $Q_{NX}(K)$ is derived to quantify the preservation of K -ary neighbourhoods by calculating the average fraction of neighbours shared between the HD set ν_i^K and the LD set n_i^K , expressed as: [10]

$$Q_{NX}(K) = \frac{1}{KN} \sum_{i=1}^N |\nu_i^K \cap n_i^K|. \quad (3.5)$$

The $Q_{NX}(K)$ value strictly falls within the interval. An output of 0 indicates an empty intersection (complete loss of neighbourhood structure), while an output of 1 indicates perfect agreement (the exact same K neighbours surround every point in both spaces) [3]

3.3.2 Limitation of $Q_{NX}(k)$ and the Need for Rescaling

While intuitive, the $Q_{NX}(k)$ metric suffers from a critical mathematical flaw when used for comparative analysis: its baseline expectation is highly sensitive to the chosen neighbourhood size K . As K grows and neighbourhoods encompass a larger fraction of the dataset, the probability of points randomly overlapping in both sets naturally increases. Specifically, if an algorithm were to completely fail and position points randomly in the LD space, the expected overlap size is not zero. For a purely random embedding, the expected value of $Q_{NX}(k)$ is approximately $\frac{K}{N-1}$. This shifting baseline makes it extremely difficult to fairly compare the preservation of small, fine-grained local neighbourhoods (small K) against the preservation of large, macroscopic global structures (large K) using $Q_{NX}(k)$ alone. [10] [3]

3.3.3 The Solution: The $R_{NX}(k)$ Metric

To resolve the scale-dependency bias inherent in $Q_{NX}(K)$, the metric was mathematically adjusted to account for the baseline expected by chance. This led to the creation of the Rescaled Average Agreement Rate (R_{NX}). The $R_{NX}(K)$ metric subtracts the random expectation from the measured overlap and rescales the result, yielding a standardized metric across all possible neighbourhood sizes K (where $1 \leq K \leq N-2$). The transformation is formulated as: [3]

$$R_{NX}(k) = \frac{(N-1)Q_{NX}(k) - k}{N-1-k}. \quad (3.6)$$

Interpretation: A perfect neighbourhood preservation for a given k corresponds to $R_{NX}(k) = 1$ while a random embedding corresponds to $R_{NX}(k) = 0$.

Scale Sensitivity: Plotting $R_{NX}(k)$ against k on a logarithmic scale allows for the evaluation of local preservation (small k) versus global preservation (large k).

AUC Evaluation: To provide a single summary statistic, the Area Under the $R_{NX}(k)$ Curve (AUC) is calculated:

$$AUC = \frac{\sum_{k=1}^{N-2} R_{NX}(k)/k}{\sum_{k=1}^{N-2} 1/k}. \quad (3.7)$$

This weighted AUC puts more emphasis on the preservation of small neighbourhoods, which is often the primary goal of non-linear DR.

3.4 ZADU: A Comprehensive Tool for Evaluating Distortions

ZADU is a comprehensive Python library specifically designed to evaluate the reliability of Dimensionality Reduction (DR) embeddings by mathematically quantifying the structural distortions that occur when mapping high-dimensional data to a low-dimensional space. Historically, assessing these distortions required analysts to rely on scattered, fragmented codebases or general-purpose machine learning libraries that offered very limited support for specialized evaluation metrics.

The core concept behind ZADU is to centralize and optimize this evaluation process through a unified framework built on three primary conceptual and technical pillars: comprehensive measure integration, automated execution optimization, and the extraction of local pointwise distortions. [9]

3.4.1 Theoretical foundations: Structural levels of Distortions

Because different DR algorithms distort data in different ways, ZADU categorizes its distortion measures into three levels based on their target structural granularity: local, cluster-level, and global. Employing multiple metrics with varying granularity is essential for a comprehensive evaluation. [9]

3.4.2 Local Measures: neighbourhood and Rank-Based Topology

Local measures evaluate the preservation of the neighbourhood structure from the high-dimensional space X into the embedding Y . These metrics primarily assess the integrity of the k -nearest neighbours and the relative rankings of data points. [9]

Rank-Based neighbourhood Assessment

The core of local evaluation within ZADU rests on rank-based preservation principles. **Trustworthiness and Continuity (tnc)** and **Mean Relative Rank Errors (mrre)** are mathematically similar in their focus on neighbour rankings. Trustworthiness penalizes "intrusions" where false neighbours appear in the 2D plane, while Continuity penalizes "extrusions" where original neighbours are torn apart. Similarly, the **Local Continuity Meta Criteria (lcmc)** and **Class-Aware Trustworthiness and Continuity (ca_tnc)** leverage these ranking foundations to ensure that the local manifold structure remains intact across projections.

Graph and Proximity Metrics

Beyond ranking, ZADU incorporates measures that utilize graph theory and linear alignment. **neighbour Dissimilarity (nd)** evaluates the Shared-Nearest neighbour (SNN) graph structure, while **neighbourhood Hit (nh)** measures the class-homogeneity of local neighbourhoods. The **Procrustes Measure (proc)** assesses local geometric fidelity through linear transformation alignment between X and Y .

3.4.3 Cluster-Level Measures: Structural Fidelity and Validation

Cluster-level measures assess the reliability of broader groupings, ensuring that visual clusters in the embedding reflect genuine mathematical structures. [9]

Inter-Cluster Reliability (S&C)

The primary diagnostic in this category is **Steadiness and Cohesiveness (snc)**. Steadiness quantifies the absence of intrusions, ensuring unrelated clusters do not overlap, while Cohesiveness measures the absence of extrusions, ensuring that a single manifold is not fragmented]. These metrics are particularly effective when utilized with density-based clustering like HDBSCAN or kNN-based assignments.

Alignment and Internal Validation

To verify that the 2D layout allows for the recovery of original groupings, ZADU includes **Distance Consistency (dsc)** and **Label Trustworthiness and Continuity (l_tnc)**, which assess the consistency of class-level distances. Furthermore, **Internal Validation Measures (ivm)** such as the Silhouette score and **External Validation Measures (c_evm)** like the Adjusted Rand Index (ARI) via K-Means are used to confirm the alignment between unsupervised groupings in Y and high-dimensional ground truth.

3.4.4 Global Measures: Geometric and Statistical Consistency

Global measures evaluate whether large-scale geometric relationships and point-pairwise distances are maintained across the entire projection.[9]

The Stress Family and Divergence Principal

ZADU provides a suite of metrics focused on total geometric divergence. **Stress (stress)**, **Non-Metric Stress (nm_stress)**, and **Scale-Normalized Stress (sn_stress)** all share the principle of measuring the aggregate variance between distance matrices. Similarly, **Kullback-Leibler Divergence (kl_div)** and **Distance to Measure (dtm)** evaluate the global distribution of data density to prevent "manifold crushing" during the reduction process.

Statistical Correlation and Global Mapping

To assess monotonic and linear consistency, ZADU employs **Pearson's r (pr)** and **Spearman's ρ (srho)**. Both share the principle of quantifying how the ranking of all point-pairs varies between X and Y . Finally, the **Topographic Product (topo)** specifically identifies global topological errors such as "folding" or "stretching" of the high-dimensional skeleton.

3.4.5 Visual Analytics Integration

The mathematical isolation of local pointwise distortions naturally extends into visual analytics. ZADU includes a complementary module called **ZADUVis**, which takes the local distortion values computed by ZADU and maps them directly onto the low-dimensional scatterplot. This allows practitioners to see where and how severely the embedding is distorted, rather than relying on a single global score.

ZADUVis implements two state-of-the-art distortion visualisation techniques: CheckViz and Reliability Map. [9]

CheckViz

CheckViz decomposes the 2D embedding space into a Voronoi diagram. Each data point is represented by its Voronoi cell, and the cell is coloured according to the local distortion score (e.g., Steadiness or Trustworthiness) of that point. The colour map follows the 2D scheme proposed by

Lespinats and Aupetit (a hue-saturation-lightness map that encodes two distortion components simultaneously).

- Pale or uncoloured cells indicate high steadiness / low distortion.
- Dark or saturated cells highlight regions where false neighbours (intrusions) or torn manifolds (extrusions) are concentrated.

CheckViz is particularly effective at localising intrusions (points far apart in the original space become neighbours in the low-dimensional map) because the Voronoi cell reflects the immediate spatial neighbourhood. [9]

Reliability Map

The Reliability Map takes a different approach: it constructs the k-nearest neighbour graph in the embedded space by drawing edges between points that are neighbours in the 2D projection. Each edge is then coloured by the distortion magnitude of its two endpoints.

- Faint, discrete edges indicate that the local neighbourhood structure is intact, points that are neighbours in the embedding are also true neighbours in the high-dimensional space.
- Dark, dense webs of edges reveal extrusions: the manifold has been torn, and points that should be far apart are forced into the same region.

Reliability Maps are complementary to CheckViz: whereas CheckViz highlights intrusions via cell colour, Reliability Maps highlight extrusions via edge colour and density.

Chapter 4

Results and analysis

4.1 Experimental setup

This section details the data acquisition, pre-processing pipeline, and the multi-perspective evaluation framework employed to assess the quality of various dimensionality reduction (DR) embeddings. To evaluate DR performance across diverse topological structures, 6 benchmark datasets were selected based on their distinct geometric properties: two image datasets COIL-20 and MNIST, two phonetic datasets Isolet and Phoneme, a tabular with binary features CoverType and a biological dataset PBMC3K.

4.1.1 Pre-processing

Mild PCA: Variance-Preserving

Directly processing flattened image vectors introduced significant computational overhead due to the high initial feature count (i.e: $D = 16,384$ for COIL-20). To mitigate this, a "mild" Principal Component Analysis (PCA) keeping 99% variance was applied as a de-noising and reduction step. Data centering, is a critical first step in Principal Component Analysis to ensure that the analysis focuses correctly on the variance (section 2.1). However, this adjustment is often deemed unnecessary when all features in a dataset already share a common scale. Within this experimental setup, CoverType is the primary dataset requiring such treatment, as its 54 features encompass a diverse range of measurements including continuous variables like elevation and aspect alongside binary soil type indicators.

Dataset sampling

In addition to feature reduction, the sample size (N) of the large datasets was addressed to accommodate the high computational complexity of non-linear embedding techniques. More specifically Metric MDS (and subsequently Isomap), which typically scales with $O(N^2)$ in memory and $O(N^3)$ in time. The datasets were sampled using a stratified methodology to preserve the original class proportions, ensuring reliable results while enabling the high-fidelity computation of exact pairwise distance matrices and $R_{NX}(K)$ curves within a feasible timeframe.

4.1.2 Datasets description

Dataset	Samples (N)	Dims (D)	Classes	Description
COIL-20	1,440	16,384	20	Multi-object rotational manifolds
MNIST	60,000	784	10	Handwritten digits (0–9)
Phoneme	150	150	5	Log-periodogram speech spectra
CoverType	581,012	54	7	Forest cover cartographic features
Isolet	7,797	617	26	Isolated letter speech recordings
PBMC3k	2,638	1,838	8	PBMC single-cell RNA-seq (10X)

Table 4.1: *Experimental datasets description*

COIL-20 (Columbia Object Image Library)

Consists of 1,440 grayscale images of 20 physical objects, each photographed in-plane at 72 equally-spaced rotations (one frame every 5°). The images are resized to 128×128 pixels, yielding 16,384 features per sample. This dataset is widely used for evaluating dimensionality reduction on visual manifolds, as the rotational progression induces a smooth, continuous structure in pixel space. It contains 20 balanced classes. [17]

MNIST

A benchmark dataset of 60,000 handwritten digit images (0–9), each a 28×28 grayscale image (784 pixels). It is a standard reference for comparing classification and embedding methods due to its moderate dimensionality, large sample size, and clearly separated classes with real-world variability. For computational tractability, a stratified subset of 3,000 samples is used, and PCA reduces the 784 features to 331 components while retaining 99% variance.[12] [18]

Phoneme

Comprises 4,509 speech recordings from the TIMIT corpus, each represented as a 256-dimensional functional data curve capturing frequency profiles over time. The samples belong to five phoneme classes and are provided via the *scikit-fda* library function *fetch_phoneme()*. This dataset tests the ability of DR methods to preserve structure in smooth, functional data with class-specific spectral patterns. [19][7]

Isolet (Isolated Letter Speech Recognition)

Consists of 7,797 recordings of spoken letters (A–Z) produced by 150 speakers, each uttering each letter twice. The 617 acoustic features per sample include spectral coefficients, pitch, and sonorant information. This dataset provides a balanced 26-class problem with moderate-to-high dimensionality, suitable for assessing how well embeddings separate acoustically-similar phoneme groups.[5]

CoverType

Contains 581,012 cartographic observations from the Roosevelt National Forest, each described by 54 features (elevation, aspect, slope, soil types, wilderness area indicators, etc.) predicting one of seven forest cover types. The features include highly-correlated binary and continuous variables, PCA at 99% variance reduces dimensions from 54 to only 49 components.[1]

PBMC3k (Peripheral Blood Mononuclear Cells)

It's a single-cell RNA sequencing dataset of Peripheral Blood Mononuclear Cells from 10X Genomics, containing approximately 2,700 cells after quality filtering. It is loaded via scanpy's built-in `pbmc3k_processed()` function, which returns log-normalized expression values for 1,838 highly variable genes across 2,638 cells. The dataset includes 8 Louvain cell-type cluster labels representing major immune cell populations. In the pipeline, mild PCA at 99% variance retained 1,556 components, meaning the data's effective rank is very close to its original dimensionality, this is typical for scRNA-seq data where many independent gene expression programs contribute to the biological variation. [23][22]

4.1.3 Summary

Although the CoverType and PBMC3k datasets exhibit significant class imbalance, K -means clustering was purposefully selected as the uniform grouping method across all datasets to ensure a standardized and fixed cluster size for the ZADU evaluation framework. This methodological choice facilitated a direct comparison between Ward's hierarchical clustering and K -means through concordance heatmap analysis, aiming to identify which clustering strategy better aligns with the $R_{NX}(K)$ preservation curve. The findings revealed that measures derived from K -means fundamentally dominated those based on Ward clustering within the small to medium K -neighbourhood ranges, providing a more reliable signal of local and mesoscopic structural integrity. While Ward's method demonstrated a marginal advantage (more than 0.1 in absolute difference) in macroscopic, large-scale neighbourhoods for a select subset of scores (specifically 5 correlation pairs only), the overall results confirm that K means is the superior choice for tracking embedding fidelity across most neighbourhood scales.

	COIL-20	MNIST	Phoneme	Isolet	CoverType	PBMC3K
Samples (n)	1,440	3,000	3,000	3,000	3,000	2,638
Dim (d)	580	331	210	386	49	1,556
Number of clusters	20	10	5	26	7	8
Class distribution	balanced	balanced	balanced	balanced	imbalanced	imbalanced
Cluster sizes	72	300	600	115	428	337
Cluster size (%)	5%	10%	20%	3.8%	14.2%	12.5%
Data density	Very dense	Dense	Sparse	Dense	Sparse	Very dense

Table 4.2: *Summary of Datasets after Pre-processing*

4.1.4 Framework

4.1.5 Packages

In this study, two particular Python packages were utilized to provide the computational and evaluative framework. The first is "*Fast_Multi-scale_NE*" (https://github.com/cdebodt/Fast_Multi-scale_NE/tree/master), which provides the implementation of accelerated multi-scale dimensionality reduction techniques (FM_sSNE and FM_sTSNE) designed to preserve both local and global structures efficiently and `eval_dr_quality()` to compute the quality metrics; RNX curve and AUC. The second is "*ZADU*" (<https://github.com/hj-n/zadu>), a comprehensive library of dimensionality reduction distortion measures used to calculate the seventeenth quality metrics required for the concordance analysis. [9][6] For reproducibility, the `random_seed` was fixed at 42.

Dimensionality Reduction Framework: Six DR algorithms were selected to represent linear baselines, modern manifold learning, and distance-based optimization. All methods were tasked with projecting the pre-processed data into a 2D latent space ($d = 2$).

Method	Initialization	Parameters
PCA	SVD (deterministic)	$n_{comp} = 2$
UMAP	Spectral	$k = 15, \delta = 0.1$
Metric MDS	Classical MDS	$n_{init} = 1$
Issomap	-	$n_{comp} = 2, n_{neighbours} = 10$
FMsSNE	PCA	Multi-scale
FMsTSNE	PCA	Multi-scale

Table 4.3: *Dimensionality Reduction Methods and Configurations*

Multi-Perspective Quality Framework: To rigorously quantify the quality of the 2D embeddings, various metrics were employed. These measures are organized into three thematic categories that capture distinct aspects of structural preservation: local fidelity, global preservation, and cluster integrity defined in Section 3.4. Afterward a comparison of difference in heatmap’s correlations with ward linkage was made

4.2 Visual Analysis of the Embeddings

The visual interpretation of dimensionality reduction (DR) embeddings is often hindered by the subjective nature of scatter plot layouts. To achieve a more rigorous assessment, this chapter employs a multi-layered diagnostic framework that synthesizes qualitative visual patterns with quantitative geometric metrics. The primary challenge in projecting high-dimensional data into a 2D plane is the ‘crowding problem’, where the limited spatial capacity of the lower dimension inevitably introduces distortions.

Our approach begins by evaluating semantic fidelity through a comparison of ground-truth labels and unsupervised k-means clusters. This is followed by a structural validation that compares high-dimensional (HD) clusters, representing the mathematical ‘truth’ of the original space, against their low-dimensional (LD) projections. Finally, these observations are cross-referenced with ZADUVis diagnostics, specifically CheckViz and Reliability Maps, to localize topological errors such as intrusions and extrusions that might otherwise lead to data misinterpretation.

4.2.1 Evaluation

The evaluation of DR embeddings in this study is anchored by the dual pillars of semantic alignment and topological reliability. While semantic validation qualitatively assesses an algorithm’s ability to “untangle” high-dimensional feature spaces into recognizable class regions, the primary benchmark for success is the preservation of structural fidelity. To mitigate the inherent topographic bias toward spherical and convex Clusters in K-means clustering, the alignment between mathematical groupings identified in the original high-dimensional space (HD-Clusters) and their visual counterparts in the embedding (LD-Clusters) is prioritized.

The objective is to verify that the DR algorithm has maintained the Steadiness and Cohesiveness of the data manifold, ensuring that observed clusters are faithful representations of intrinsic structures rather than artifacts of the projection or “topographic hallucinations” of the clustering tool. This structural assessment is further localized by employing ZADUVis spatial diagnostics to measure localized geometric distortions as explained in Section 3.4.5.

4.2.2 COIL-20 Dataset

The COIL-20 dataset features 1,440 images with an extremely high raw dimensionality ($D = 16,384$), which was reduced to $d = 580$ through PCA to maintain 99% variance. Theoretically, this dataset captures object manifolds formed by 360-degree rotations, requiring an algorithm that can "untangle" these non-linear loops in 2D space without improperly tearing or crushing the neighbourhood structures. A successful mapping must minimize both intrusions (false neighbours) and extrusions (torn manifolds) to ensure that the 20 distinct object classes remain isolated and mathematically coherent.

UMAP: The Standard for Local Topology

Visually UMAP provides the most structurally sound projection for COIL-20. The alignment between the Ground Truth (a) and clusters LD (b) (low-dimensional k-means) plots shows that the unsupervised algorithm perfectly isolated the true classes with virtually no overlap. Crucially, the clusters HD (c) (high-dimensional labels projected into 2D) plot perfectly mirrors the LD clusters, confirming that the 2D layout faithfully represents the original high-dimensional groupings without "hallucinatory" clustering. The ZADU diagnostics provide empirical proof of UMAP's structural success. The CheckViz plot (d) consists of predominantly pale or white Voronoi cells, which indicates that the local topology is intact and free from significant distortion. Its Reliability Map (e) displays only isolated, discrete rings with faint edges, proving that the neighbourhood structure is highly reliable and that the 360-degree loops were successfully unfolded without neighbourhood intrusions

FMsTSNE and FMsSNE: Successful Separation with local distortion and neighbourhood instability

When comparing the fast multi-scale methods, both FMsTSNE and FMsSNE successfully separate the data into discrete islands, but they exhibit distinct localized distortions and global arrangements. Regarding their layouts, FMsTSNE tends to produce more discrete and separated islands with clear "white space" between them, whereas FMsSNE results in a more centrally-clustered and interconnected layout where digit manifolds often appear as "fingers" extending from a common mass. The FMsTSNE CheckViz plot reveals localized drops in Steadiness, evidenced by distinct patches of dark purple where "false neighbours" have intruded into specific digit subgroups. In contrast, FMsSNE maintains a visually consistent separation, but its Reliability Map reveals a denser web of colored edges compared to UMAP, signaling a higher level of neighbourhood instability where k-nearest neighbours in 2D do not perfectly mirror the original 331-dimensional space. While both methods avoid the "crushing" failures of linear models, FMsTSNE suffers from more severe local intrusions, while FMsSNE exhibits broader neighbourhood instability across the entire embedding.

Conclusion

In stark contrast to the manifold learning techniques, PCA produces compressed clouds where classes blur together, whilst methods like MDS, and Isomap fail to clearly separate the classes in the low-dimensional space because the underlying manifolds are not well separated and exhibit substantial overlap. Thus resulting in a catastrophic failure to find a separating axis in the high-dimensional space. Their CheckViz plots (d, i, n) exhibit heavy dark shading, and their Reliability Maps (e, j, o) show highly dense, interconnected webs of dark edges. These dark markers confirm that these methods "severely crushed" the 20 distinct image classes together, resulting in highly unreliable and uninterpretable local neighbourhoods.

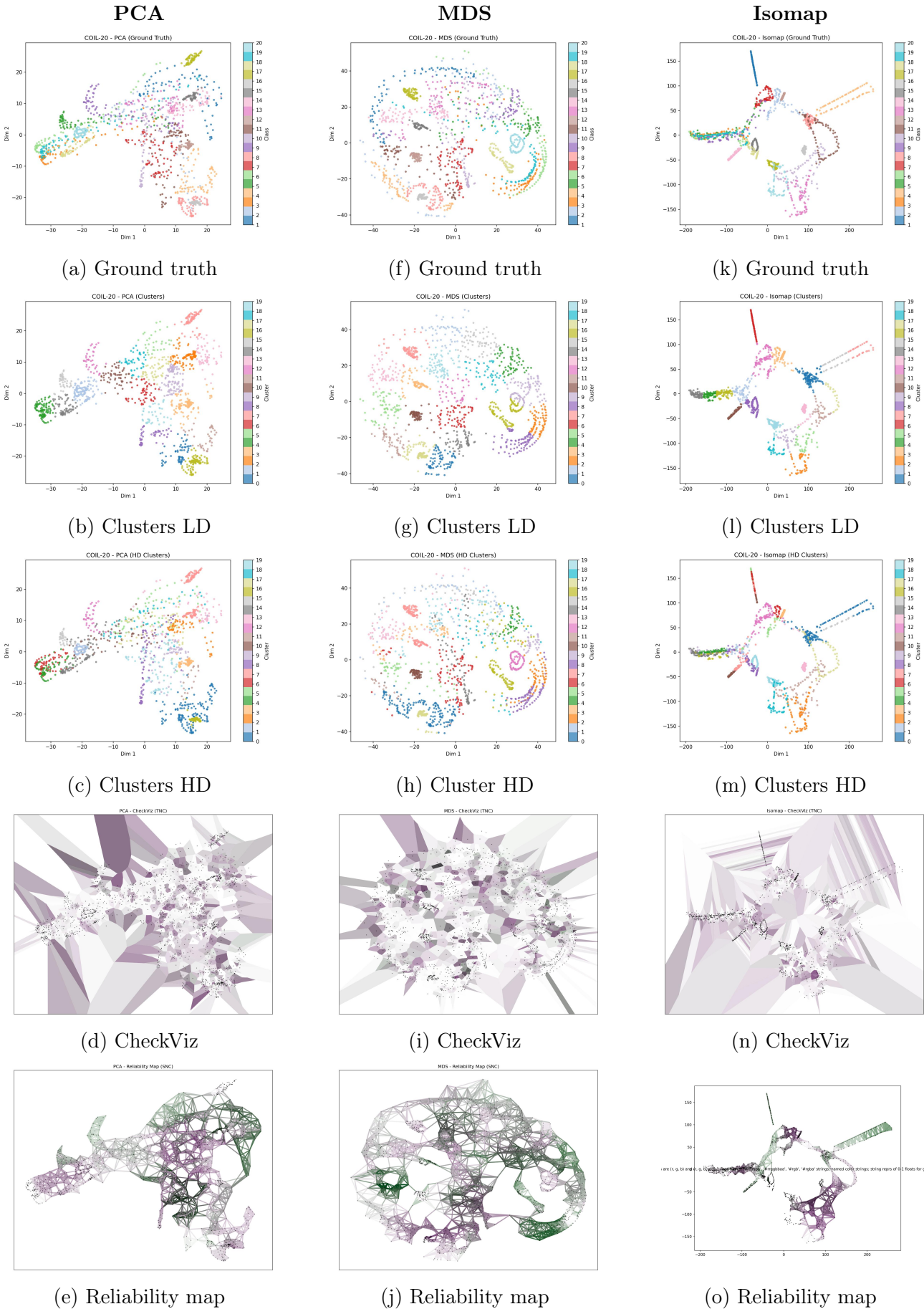


Figure 4.1: From left to right: PCA, MDS, and Isomap embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the COIL-20 dataset.

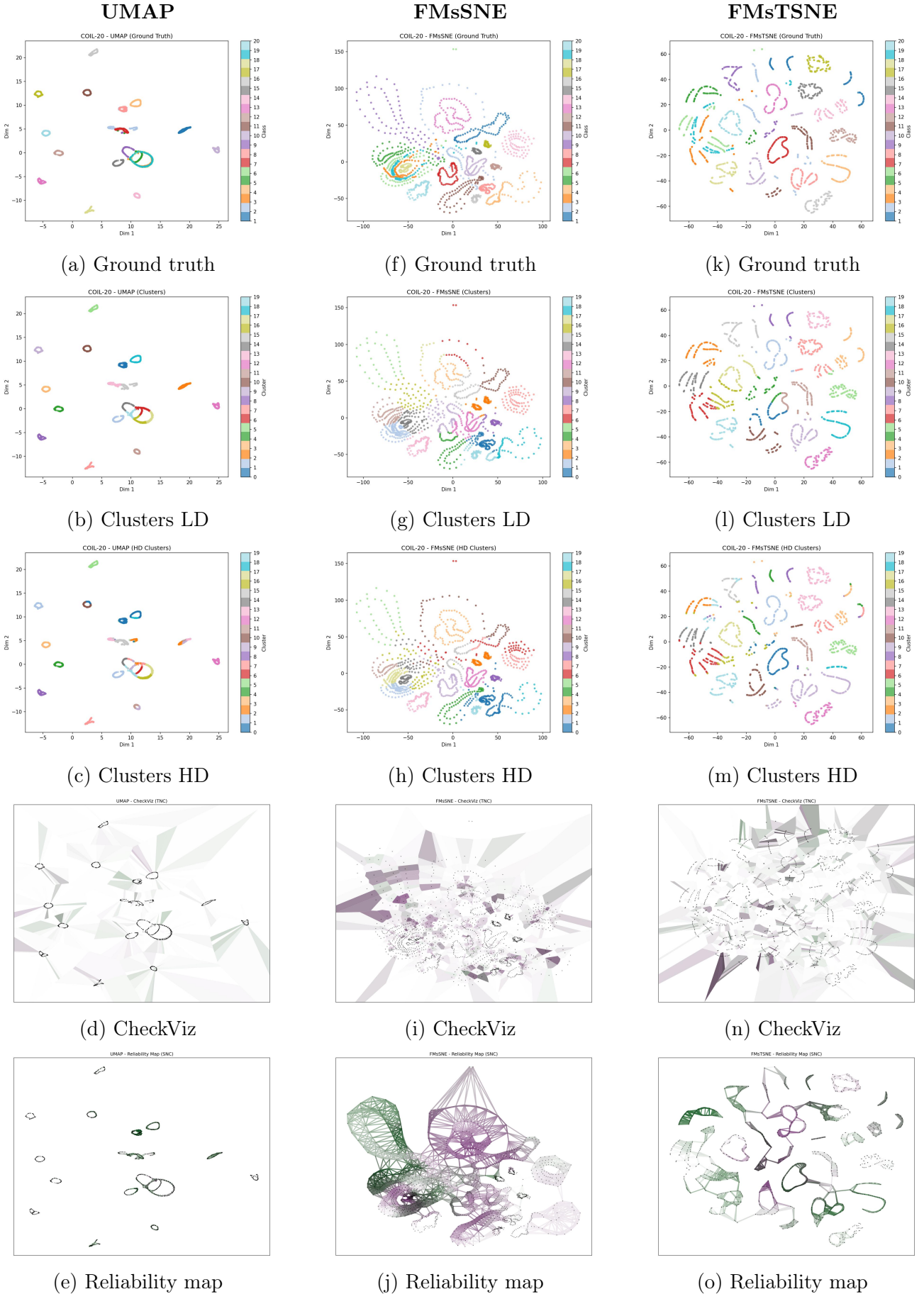


Figure 4.2: From left to right: UMAP, FMsSNE, and FMsTSNE embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the COIL-20 dataset.

4.2.3 MNIST Dataset

The analysis of the MNIST dataset embeddings is conducted on a sampled set of 3,000 digits, where the initial 784 raw dimensions were reduced to 331 through PCA to preserve 99% of the variance. Theoretically, the MNIST manifold is composed of ten non-linear structures that often share high-dimensional similarities, such as the structural overlap between '3' and '8'. A successful embedding must effectively navigate the "crowding problem" to untangle these manifolds into a 2D plane without improperly "tearing" the digit structures (low cohesiveness) or forcing unrelated digits into the same visual space (low steadiness).

UMAP: The Best Available, Though Imperfect, Embedding

While UMAP is not an ideal projection for this specific sample of MNIST, it remains the most viable embedding in the study. Visually, the class separation is not as distinct as seen in other benchmarks, with some digit clusters appearing crowded or lacking clear boundaries. However, the structural integrity of the mapping is salvaged by the significant alignment between the LD-Clusters and HD-Clusters. This consistency proves that even if the visual separation is weak, the 2D layout manages to reflect the mathematical groupings present in the original 331-dimensional space.

The ZADUVis diagnostics provide a nuanced view of this "best-available" status. Despite distortions revealed by localized dark and grey patches in the CheckViz plot and a relatively dense web of coloured edges in the Reliability Map, UMAP's metrics remain significantly better than its competitors. These markers indicate that while neighbourhood intrusions and manifold "stretching" are present, they are managed more effectively here than in any other mapping, making UMAP the only method that maintains a functional level of topological reliability.

FMsSNE: The Salvageable Multi-scale Alternative

Though FMsSNE's embedding is more diffuse than UMAP's embedding and doesn't separate clusters clearly, it demonstrates faint localized distortion compared to other embeddings. Specifically, it lacks the intense dark purple patches in CheckViz that signal severe class intrusions in FMsTSNE. Although its Reliability Map reveals a denser web of coloured edges, indicating broader neighbourhood instability. It avoids the catastrophic manifold crushing seen in linear models, providing a mathematically viable, if noisy, representation of the digit manifolds.

Conclusion: Structural Fidelity in a Distorted Plane

In contrast to UMAP and FMsSNE, the remaining methods (FMsTSNE, PCA, MDS, and Isomap) fail to produce interpretable or reliable layouts. The linear and global methods suffer from a total collapse of neighbourhood integrity, appearing as "crushed" masses where the ZADU diagnostics reveal extreme levels of intrusion and manifold tearing. Ultimately, the MNIST analysis highlights a hierarchy of reliability: while UMAP and FMsSNE both show visible signs of distortion, they are the only methods capable of preserving enough structural density and neighbourhood consistency to be considered mathematically salvageable.

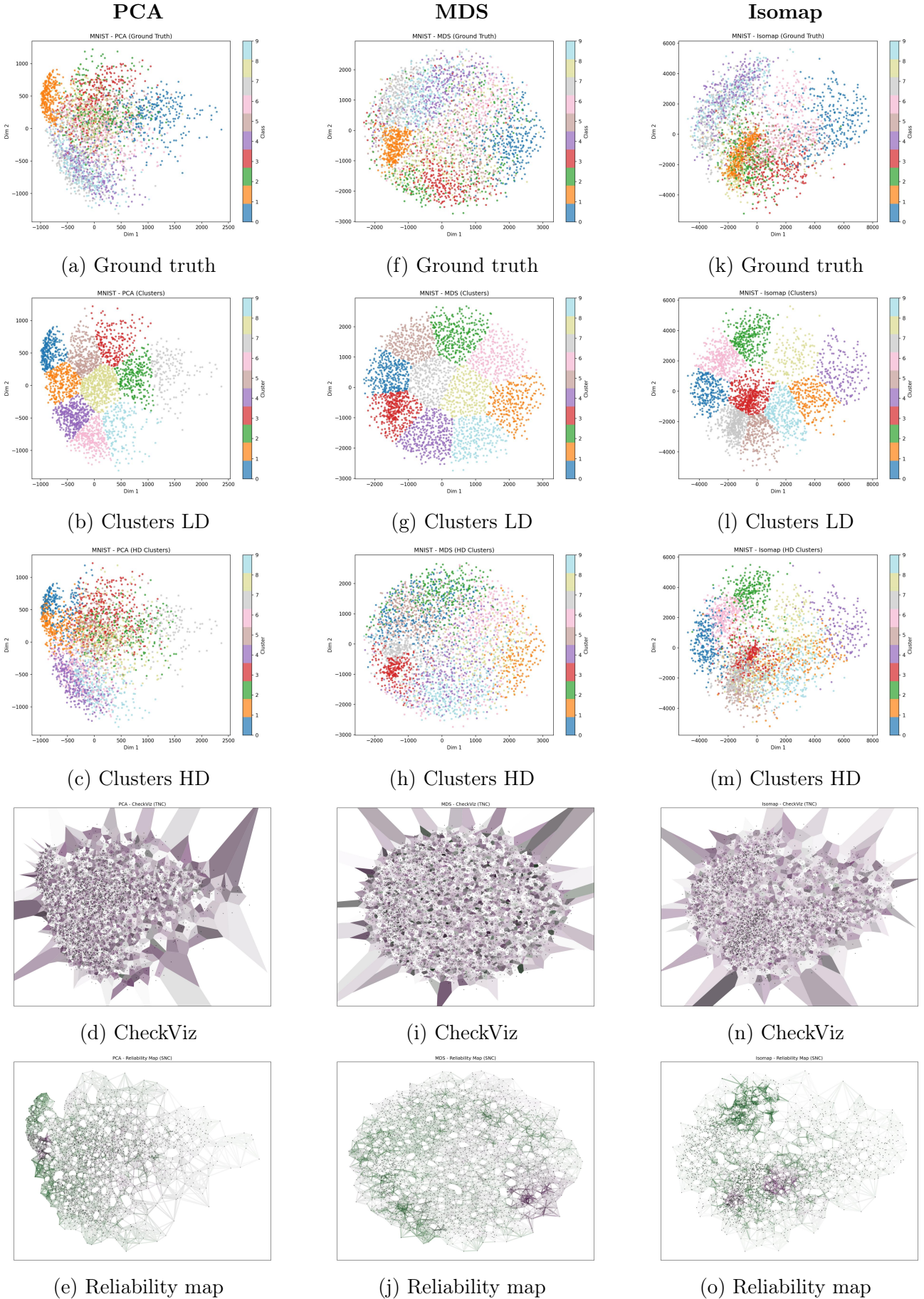


Figure 4.3: From left to right: PCA, MDS, and Isomap embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the MNIST dataset.

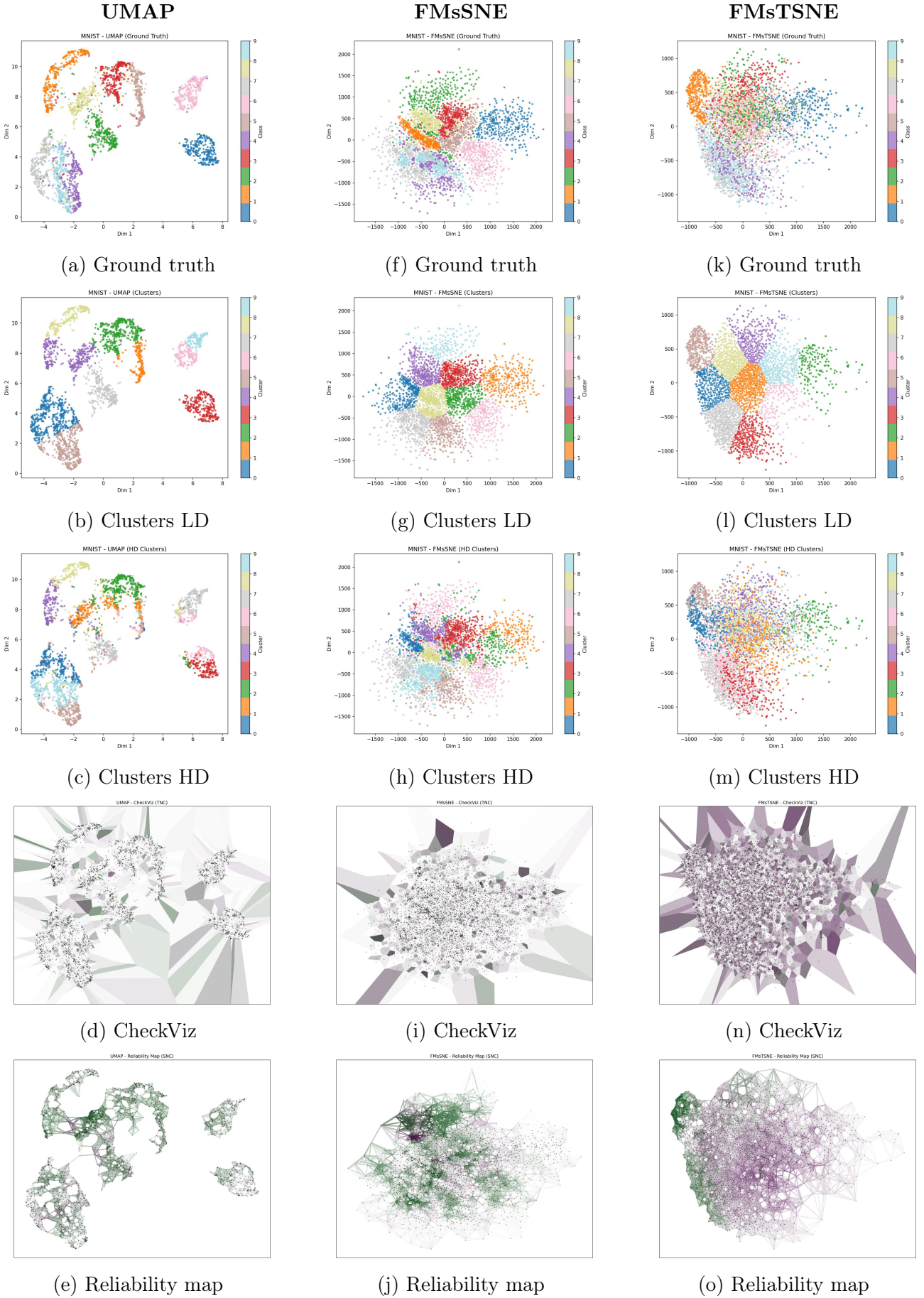


Figure 4.4: From left to right: UMAP, FMsSNE, and FMsTSNE embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the MNIST dataset.

4.2.4 Phoneme Dataset

The Phoneme dataset consists of 150 samples of spoken vowel sounds, reduced from raw time-series to 210 principal components (99% variance). The five phoneme classes represent different vowel sounds ('aa', 'ao', 'del', 'iy', 'sh') that lie on a continuous acoustic manifold rather than forming discrete islands. A successful embedding should preserve the local ordering and density of similar sounds while avoiding false neighbourhood intrusions across distinct phoneme categories.

UMAP: Aggressive Isolation

UMAP produces the most structured embedding. The ground-truth projection shows several distinct lobes corresponding to phoneme groups, though some classes remain adjacent or partially overlapping due to the faithful reflection of the continuous acoustic space rather than a failure. Critically, the HD cluster projection aligns well with the LD k-means Clusters, confirming that the 2D layout preserves the original mathematical groupings.

The CheckViz plot for UMAP displays mostly pale to light grey Voronoi cells, with only isolated dark patches at the boundaries between phoneme groups. This indicates high steadiness: points within a visual lobe are true neighbours in the original space. The Reliability Map shows faint, discrete edges within lobes and slightly darker edges only where two phoneme regions touch. This proves that UMAP has managed the crowding problem effectively, maintaining local cohesiveness despite the inherent class overlap.

FMsSNE and FMsTSNE: Preserving Transitions

The multi-scale methods successfully resolve the 5 classes but present a vastly different geometric topology than UMAP. Instead of totally disconnected islands, FMsSNE organizes the classes into a cohesive global structure with distinct lobes. This visual layout much more accurately reflects the continuous nature of acoustic data. The Reliability Map exhibits faint, manageable webbing at the boundaries between these lobes, indicating that while some neighbourhood stretching occurs, the global relationships and transition zones are far better preserved than in UMAP. FMsSNE preserves the global continuity of the phoneme manifold better than FMsTSNE but at the cost of some local neighbourhood accuracy.

PCA, MDS, and Isomap: Continuous Blurring

Visually, Isomap exhibits the clearest class separation, followed by MDS and then PCA. The latter shows an unstructured cloud where the phoneme's vowel classes are thoroughly intermixed in the ground-truth projections

The CheckViz plots (fourth row) for PCA, MDS, and Isomap are dominated by multiple grey shaded Voronoi cells, indicating widespread intrusions: points from different phonemes are forced into the same neighbourhoods. The Reliability Maps (bottom row) show extremely dense, nearly solid webs of dark edges, confirming that the k-NN graph in the embedding bears almost no relation to the original high-dimensional graph.

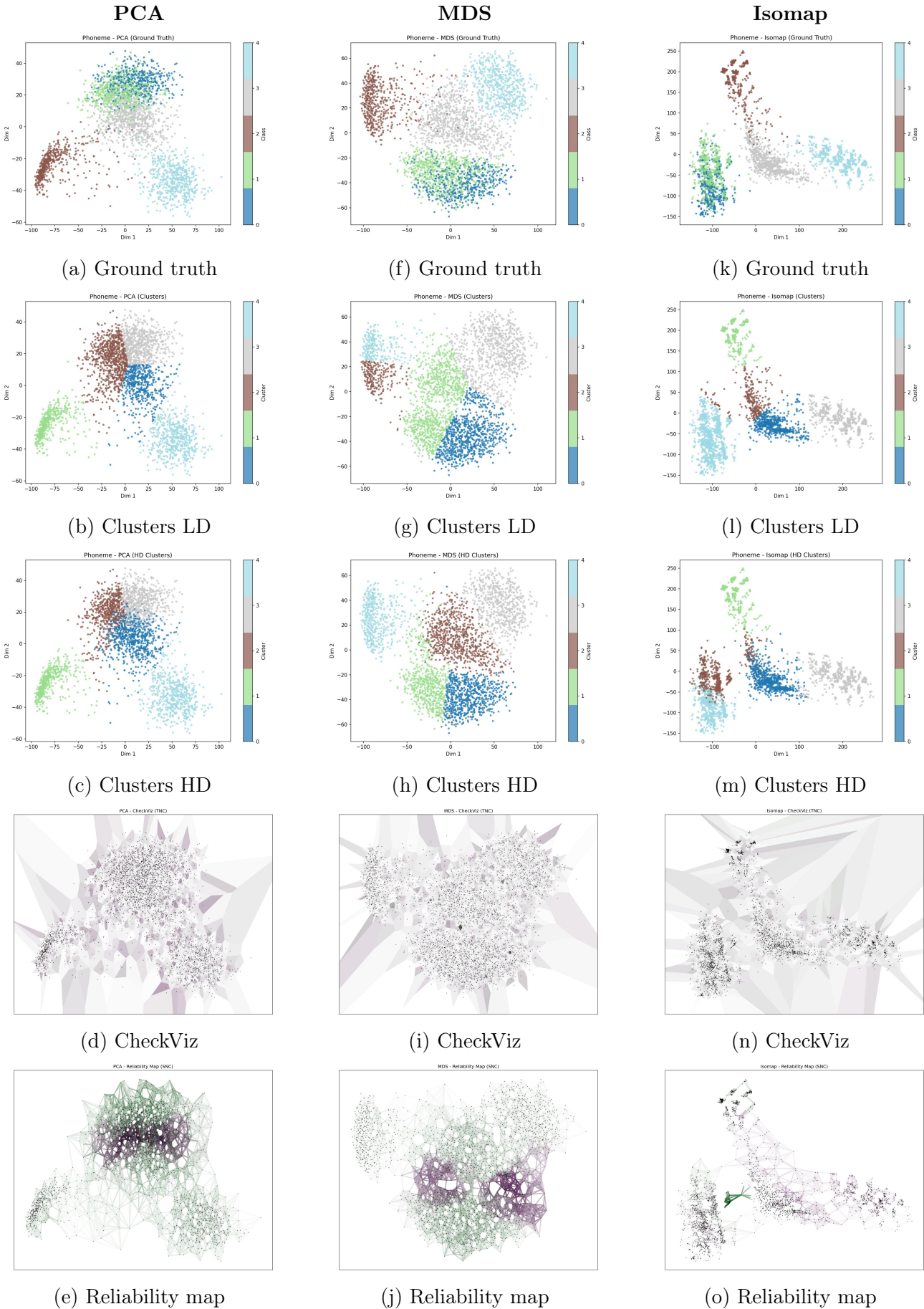


Figure 4.5: From left to right: PCA, MDS, and Isomap embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the Phoneme dataset.

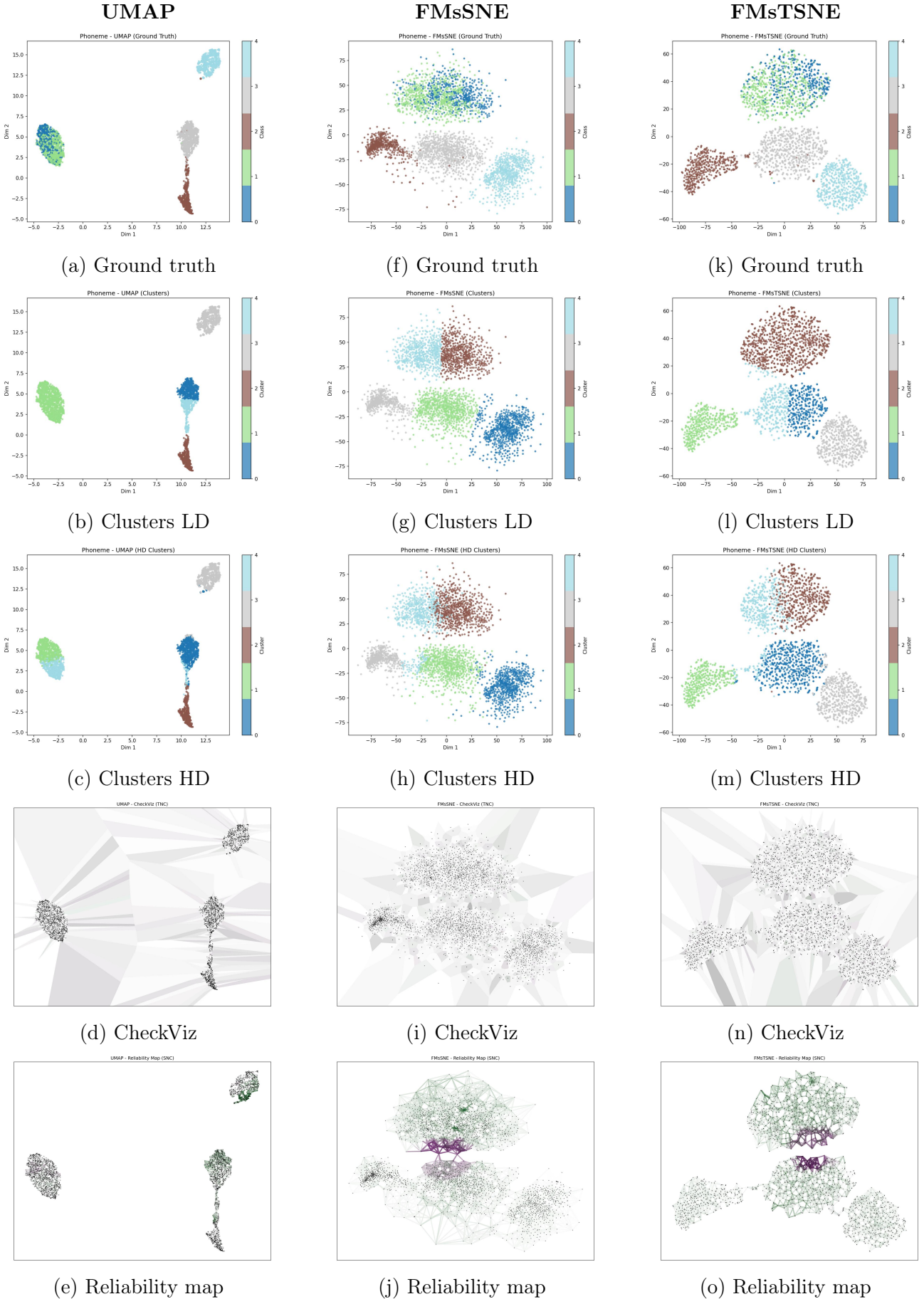


Figure 4.6: From left to right: UMAP, FMsSNE, and FMsTSNE embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the Phoneme dataset.

4.2.5 Isolet Dataset

The Isolet dataset contains 3,000 samples (subsampled from 7,797) of spoken letter names (26 classes), reduced to 386 principal components (99% variance). Each class corresponds to a different letter spoken by multiple subjects. The challenge is high class cardinality (26) with potential acoustic similarities (e.g., ‘B’ and ‘P’, ‘M’ and ‘N’). A successful embedding must separate letters into discernible Clusters while preserving the similarity structure among acoustically close letters.

UMAP: Constellation Formation

UMAP excels in this high-class-count environment. It aggressively shatters the crowded manifold, generating 26 distinct, tight constellations of points. Though we see that some classes overlap, this is due to human speech containing highly similar acoustic "families". The high semantic alignment between LD and HD Clusters proves that UMAP correctly identified the acoustic groupings. The CheckViz plot remains remarkably light, and the Reliability Map shows only isolated rings of connectivity, proving that UMAP navigated the crowding problem by forcefully maximizing inter-cluster repulsion.

FMSNE vs. FMTSNE

This dataset highlights a critical divergence between the multi-scale methods. FMSNE manages to carve out distinct local pockets of density while maintaining a connected, centralized "cloud" layout, offering a balanced, stable projection. In contrast, FMTSNE visibly struggles. Driven by its heavy-tailed Student-t distribution attempting to push 26 classes apart simultaneously, the global structure destabilizes. The Reliability Map for the multi-scale methods shows a chaotic, dense web of dark edges, indicating that in its attempt to separate the letters, they severely tore the local manifolds.

PCA and MDS: The Crowding Problem

PCA and MDS represent the textbook definition of the crowding problem. The 26 classes are entirely superimposed into a featureless, circular mass. The corresponding CheckViz plots are almost entirely dark, and the Reliability Maps are practically solid blocks of interconnected edges, proving that these global methods cannot resolve highly multi-class structures in two dimensions.

Conclusion

While UMAP does not deliver 26 perfectly isolated Clusters, it successfully identifies the distinct macro-families of acoustically similar letters and separates those families cleanly from one another. While the classes within those families still overlap heavily, UMAP is the only algorithm tested that successfully breaks the monolithic "hairball" apart without artificially tearing the underlying acoustic relationships.

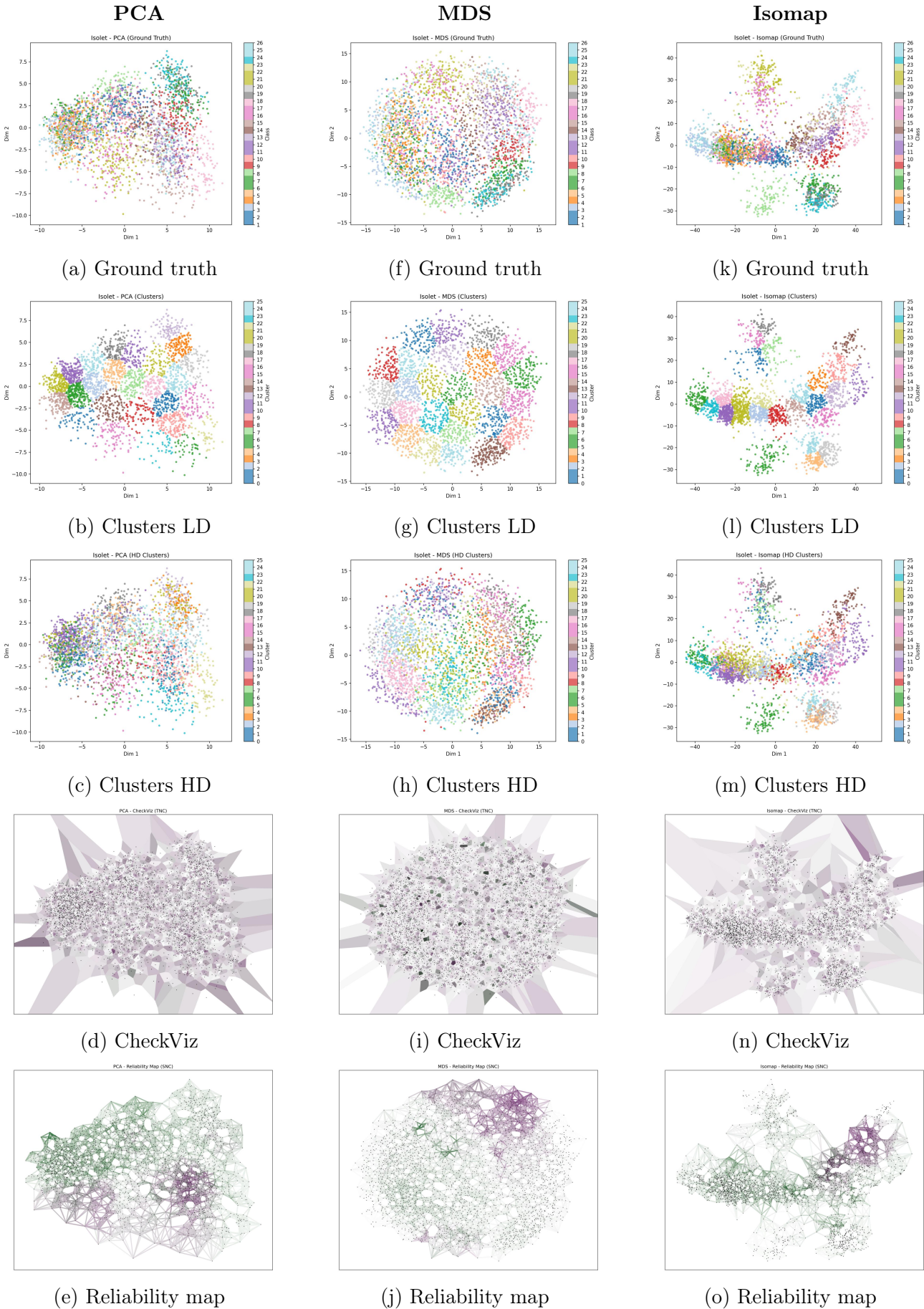


Figure 4.7: From left to right: PCA, MDS, and Isomap embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the Isolet dataset.



Figure 4.8: From left to right: UMAP, FMsSNE, and FMsTSNE embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the Isolet dataset.

4.2.6 CoverType Dataset

The dataset contains 3,000 samples of forest cover types (7 classes) with 49 variables. The space is moderately high-dimensional and likely contains noise and redundancy. The primary challenge is the crowding problem: many points in 49-D must be forced into 2D, inevitably creating false neighbourhoods.

UMAP: Artificial Tearing

Visually, UMAP appears to succeed by forming several distinct colour-coded islands. However, the spatial diagnostics reveal this to be a geometric illusion. Because the native ecological data is highly continuous, UMAP's forced separation creates massive structural extrusions. The Reliability Map is incredibly dense with dark edges crossing the "empty" space between islands, proving that UMAP has violently torn a continuous terrain into artificial fragments to satisfy its local attraction parameters.

FMsSNE vs. FMsTSNE: The Importance of the Kernel

The contrast between the two fast multi-scale methods on this tabular data is striking. FMsTSNE fails catastrophically, shattering into an uninterpretable, diffuse cloud because its heavy-tailed distribution amplifies the inherent noise of the raw features. However, FMsSNE manages to salvage a mathematically valid embedding. Relying on Gaussian distributions, FMsSNE produces a cohesive, centralized mass with distinct lobes. While its Reliability Map exhibits some localized webbing, it entirely avoids the structural explosion of FMsTSNE.

PCA, MDS, and Isomap: Honest Continuity vs Graph Disconnection and Noise

Ironically, these methods provide the most "honest" visual representation of the CoverType data. PCA and MDS display a single, highly continuous, overlapping gradient of points. While they fail to separate the 7 cartographic classes, their lack of artificial clustering accurately reflects the true overlapping nature of the raw ecological features.

Nevertheless, Isomap struggles severely, producing a compressed, fragmented layout with disconnected artifacts. Because the dataset is a massive continuum of overlapping ecological zones, Isomap's strict k -nearest neighbour graph captures noise rather than true manifold geometry. This leads to severe graph "short-circuiting," evidenced by heavy dark shading in its CheckViz plot.

In addition, the Reliability Maps show extremely dense, nearly solid webs of dark edges, confirming that the k -NN graph in the embedding bears almost no relation to the original high-dimensional graph.

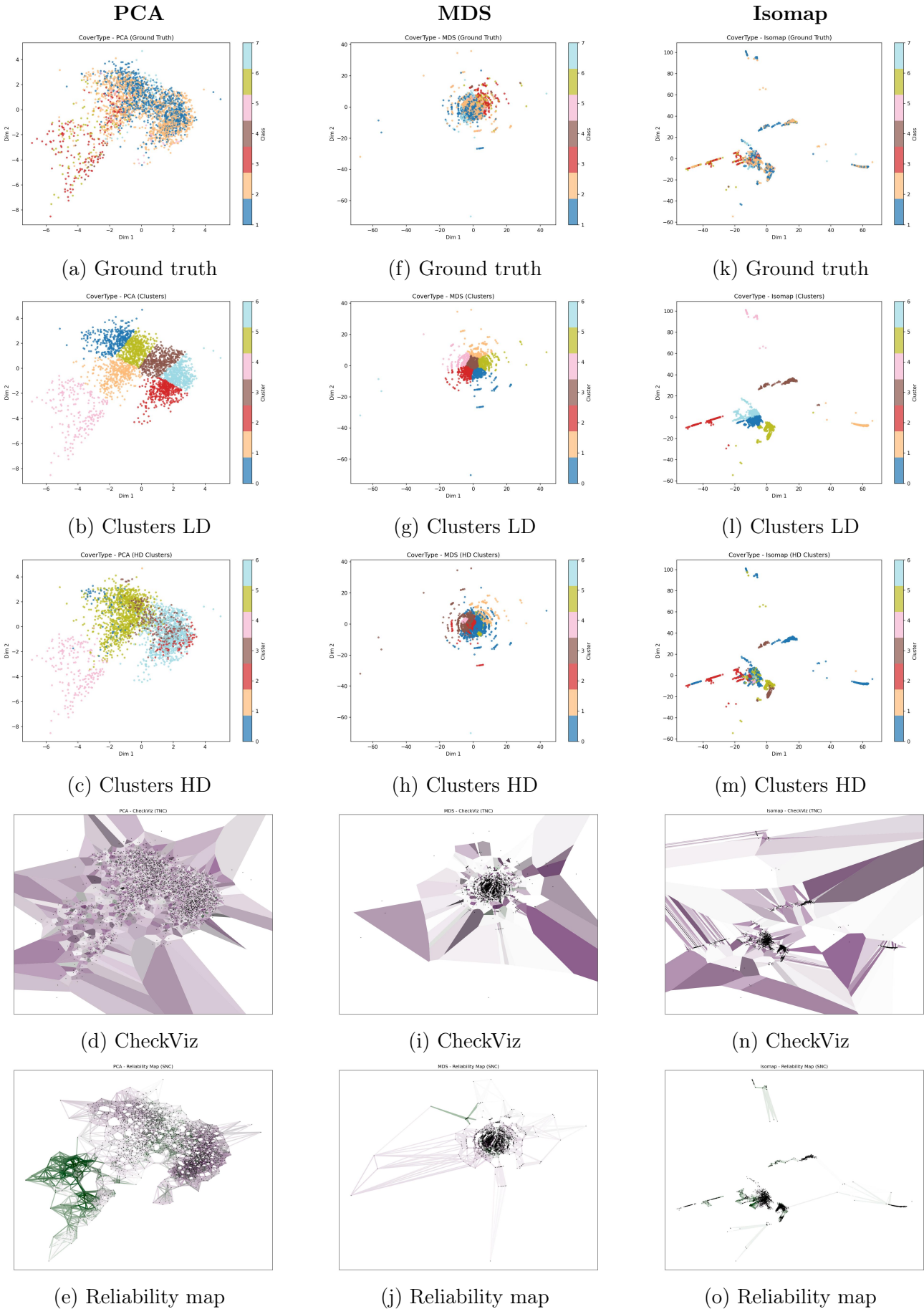


Figure 4.9: From left to right: PCA, MDS, and Isomap embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means Clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the CoverType dataset.

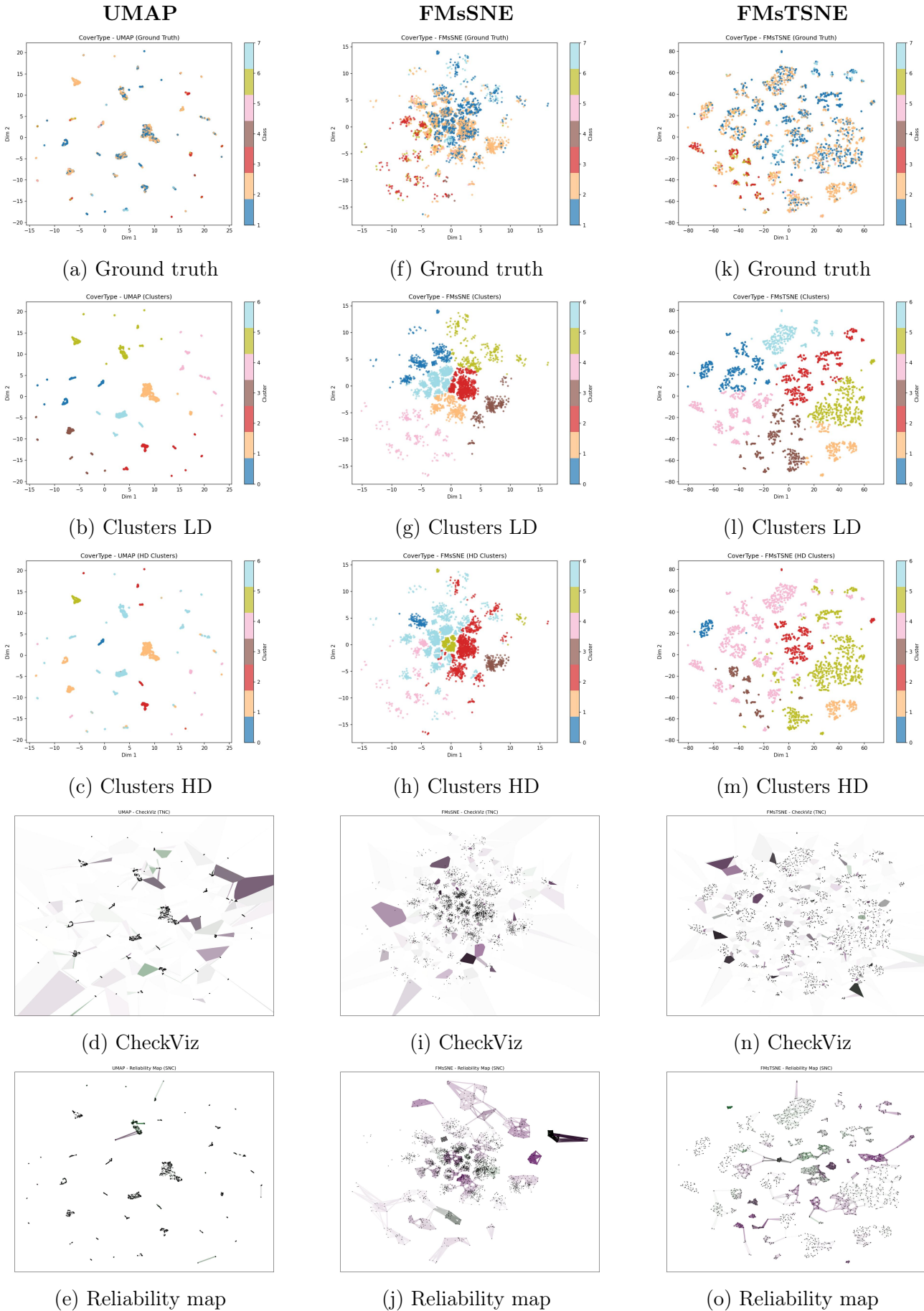


Figure 4.10: From left to right: UMAP, FMsSNE, and FMsTSNE embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means Clusters, high-dimensional Clusters, the CheckViz map, and the reliability map for the CoverType dataset.

4.2.7 PBMC3K Dataset

The PBMC3K (Peripheral Blood Mononuclear Cells, 3,000 cells) dataset is a single-cell RNA-sequencing (scRNA-seq) dataset. It contains 3,000 cells (samples) and thousands of gene expression features (typically reduced to a about 1500 principal components after filtering). The ground-truth labels correspond to different cell types (e.g., T cells, B cells, monocytes, NK cells). ScRNA-seq data represents a unique geometric challenge: it is notoriously high-dimensional, sparse, and exhibits complex, non-linear manifolds where different cell types form distinct but sometimes overlapping clusters. A successful embedding should separate cell types while preserving the continuous differentiation trajectories within each type.

PCA: Preserving Biological Lineages

Paradoxically, in this extremely high-noise environment, the global linear methods outperform the manifold learners in local stability. PCA projects the cells as continuous, branching trajectories. The CheckViz plots for the linear methods show surprisingly dark shading along these branches, indicating that the local cell-to-cell relationships is not intact, this is due to the continuous nature of the data. By focusing strictly on the axes of maximum global variance, PCA effectively filters out the transcriptomic noise, preserving the continuous cellular progression that neighbourhood methods artificially destroy.

Isomap: Sensitivity to Sparsity

Isomap's embedding yields the most severe visual artifacts in the study, projecting the cells into heavily distorted, "spiky" linear branches. The severe sparsity of single-cell data violently disrupts Isomap's geodesic assumptions. Shortest-path calculations are forced through narrow, noisy channels, stretching the low-dimensional mapping into unnatural streaks, leading to an almost entirely black CheckViz and Reliability Map.

MDS, UMAP, FMsSNE and FmsTSNE: the amorphous clouds

Contrary to its behavior on image or audio data, UMAP fails spectacularly on PBMC3k. Visually, it does not shatter the data into distinct cellular islands; instead, it crushes all 8 Louvain cell-type clusters into a single, dense, featureless circular mass.

For MDS, the data is projected as a perfectly uniform, featureless sphere. Its corresponding CheckViz and Reliability Maps are heavily darkened and extremely dense, confirming that macroscopic distance preservation completely failed to resolve the fine-grained biological nuances, crushing distinct cell types together.

The multi-scale methods, FMsSNE and FmsTSNE, fare similarly in terms of visual overlap, collapsing into tight, centralized clumps rather than massive spheres, however both multi-scale methods still show significant dark webbing and shading.

Ultimately, all four of these methods prove that probability-based neighbour and distance embeddings struggle to untangle the severe intrinsic dimensionality (1,556 effective components) and sparsity of scRNA-seq data. Forced to operate in an environment where local neighbourhoods are overwhelmed by transcriptomic noise, these algorithms crush continuous biological trajectories into uninterpretable, amorphous clouds.

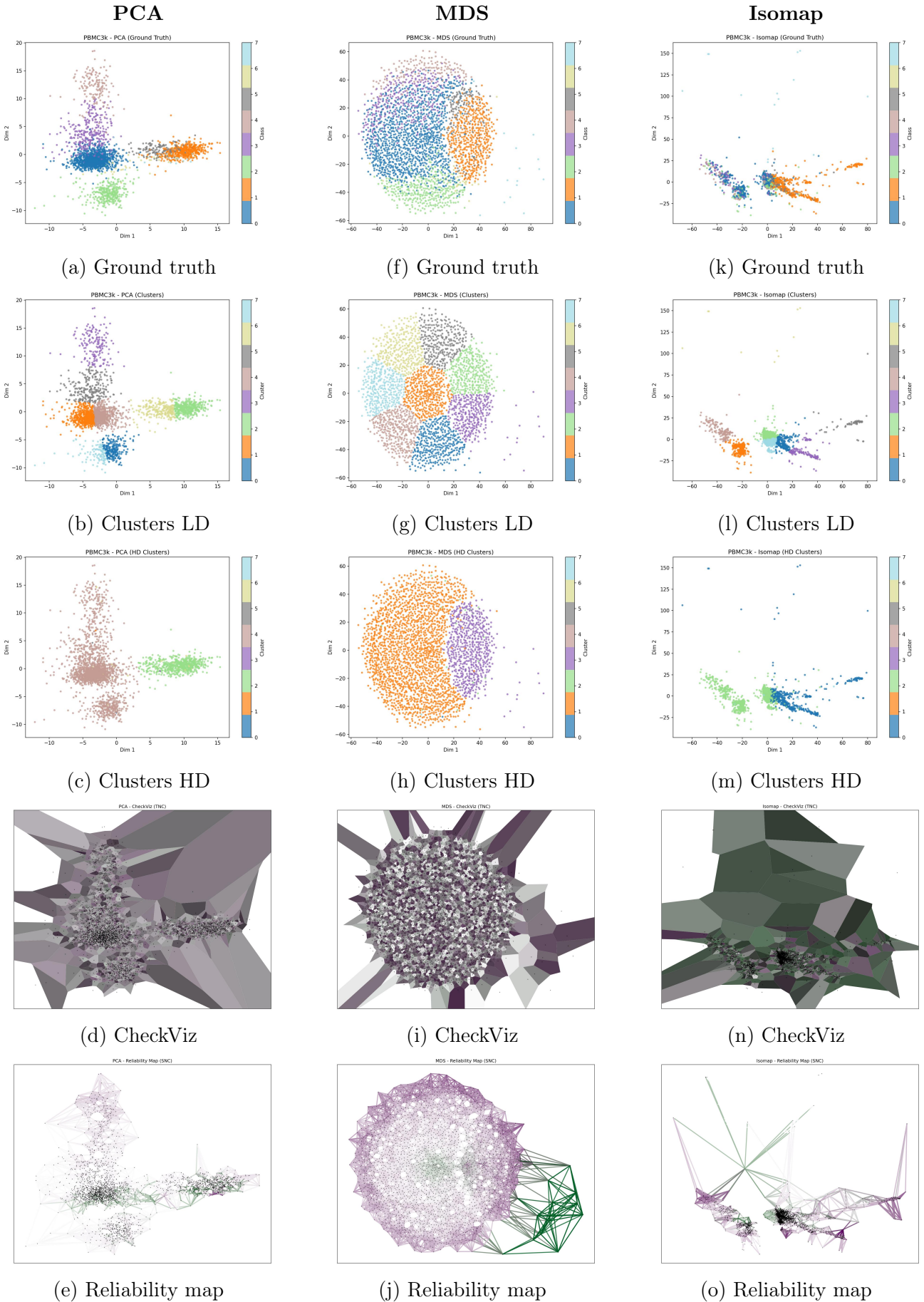


Figure 4.11: From left to right: PCA, MDS, and Isomap embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the PBMC3K dataset.

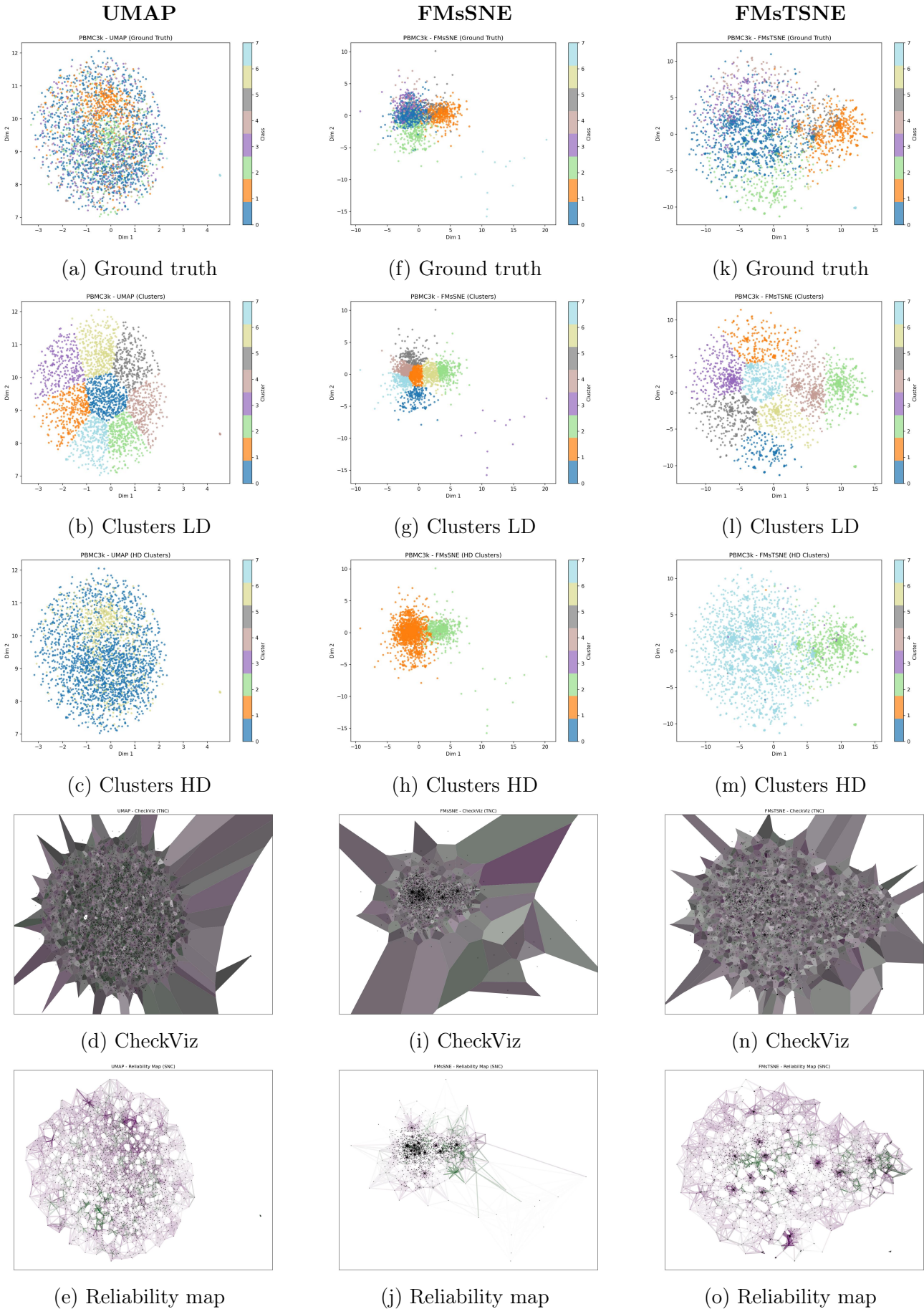


Figure 4.12: From left to right: UMAP, FMsSNE, and FMsTSNE embeddings, each displaying (from top to bottom) the ground-truth projections, low-dimensional k -means clusters, high-dimensional clusters, the CheckViz map, and the reliability map for the PBMC3K dataset.

4.3 Quantitative Metric Analysis

The visual observations are strongly backed by the quantitative metrics, specifically the R_{NX} curves and AUC scores, which measure neighbourhood preservation, and Silhouette scores, which measure cluster cohesion and separation. the Pearson correlation between corresponding clusters in LD and HD as well as zadu measures

4.3.1 Evaluation of cluster integrity metrics

Silhouette coefficient

Method	COIL-20	MNIST	Phoneme	CoverType	Isolet	PBMC3k
HD	0.219	0.052	0.263	0.092	0.089	-0.013
PCA	0.456	0.358	0.462	0.359	0.333	0.404
UMAP	0.717	0.491	0.745	0.517	0.612	0.325
MDS	0.448	0.350	0.388	0.302	0.333	0.359
Isomap	0.552	0.363	0.572	0.559	0.362	0.441
FMsSNE	0.41	0.350	0.537	0.366	0.380	0.376
FMsTSNE	0.456	0.35	0.462	0.352	0.419	0.372

Table 4.4: Silhouette across datasets.

The Silhouette Coefficient serves as a metric for evaluating how well a data point fits into its assigned cluster relative to its proximity to the next nearest cluster, with results ranging from -1 (poor assignment) to 1 (perfectly separated clusters). An analysis of the results across multiple datasets shows a consistent and significant discrepancy between high-dimensional (HD) and low-dimensional (LD) scores, where any difference exceeding 0.2 indicates a fundamental distortion of the dataset’s original geometry. While theoretically a reliable embedding should maintain scores approximately equal to the high-dimensional baseline to avoid structural fraud, only the Phoneme dataset under MDS and the COIL-20 dataset under FMsSNE approach this ideal.

In most other cases, particularly with non-linear methods like UMAP and Isomap, the silhouette scores are aggressively inflated; for example, UMAP increases the silhouette from 0.219 to 0.717 for COIL-20 and from 0.263 to 0.745 for Phoneme. Even in the PBMC3k dataset, where the original high-dimensional structure possesses a negative silhouette of 0.013, signalling a singular, indistinguishable mass, every dimensionality reduction technique produces scores above 0.3. This widespread inflation confirms that these techniques prioritize visual separation over structural truth, often "hallucinating" cluster distinctness to satisfy human visual preferences at the cost of the underlying mathematical relationships. It scores the highest among all datasets except PBMC3K, which is consistent with the observations during the visual analysis

Correlation of labels

Method	COIL-20	MNIST	Phoneme	CoverType	Isolet	PBMC3k
PCA	0.302	0.371	0.835	0.226	0.182	0.053
UMAP	0.701	0.491	0.705	0.304	0.607	-0.005
MDS	0.431	0.390	0.544	0.384	0.279	-0.095
Isomap	0.612	0.520	0.725	0.049	0.376	0.188
FMsSNE	0.636	0.483	0.899	0.456	0.537	0.520
FMsTSNE	0.601	0.362	0.706	0.414	0.641	0.369

Table 4.5: Label Correlation (Pearson) between dimensions across datasets.

The label correlation results reveal substantial differences in how effectively each dimensionality reduction method preserves class-related structure across datasets. UMAP performs particularly well on CoverType and Phoneme (0.7) while Isomap shows strong performance on MNIST (0.52) and Phoneme (0.725), indicating that preserving geodesic manifold distances can effectively retain semantic relationships in datasets with smooth intrinsic geometry. In contrast, MDS provides moderate but relatively stable correlations across most datasets, while exhibiting weaker performance on PBMC3k, where local neighbourhood preservation appears more important than global distance preservation.

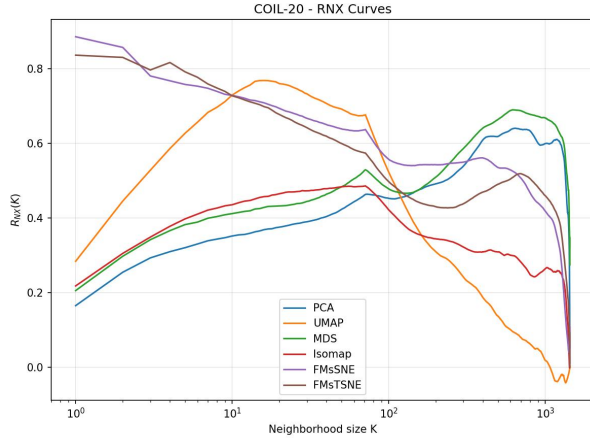
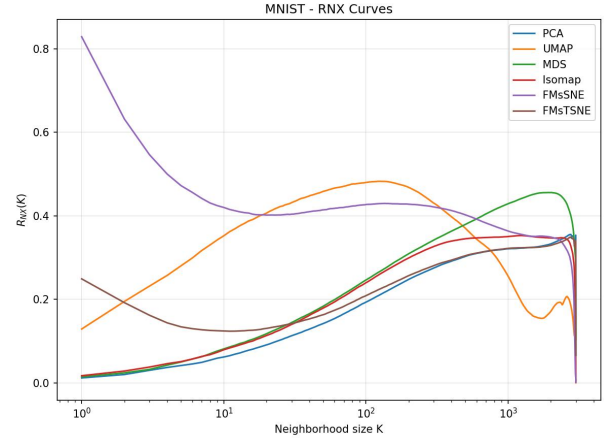
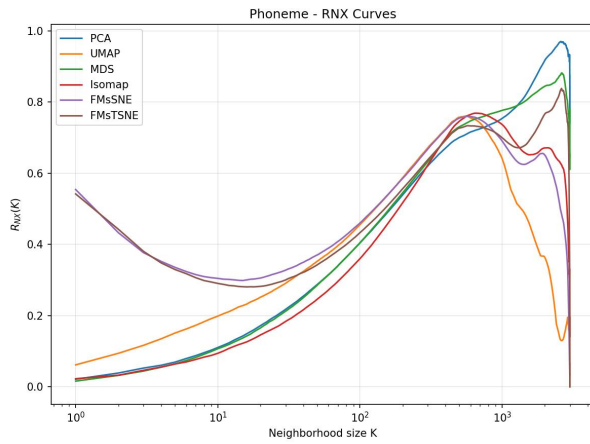
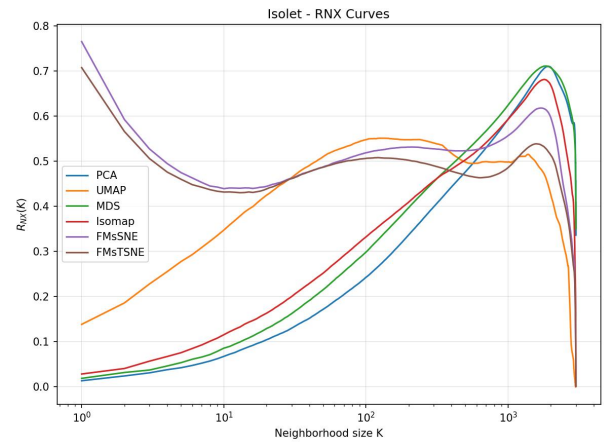
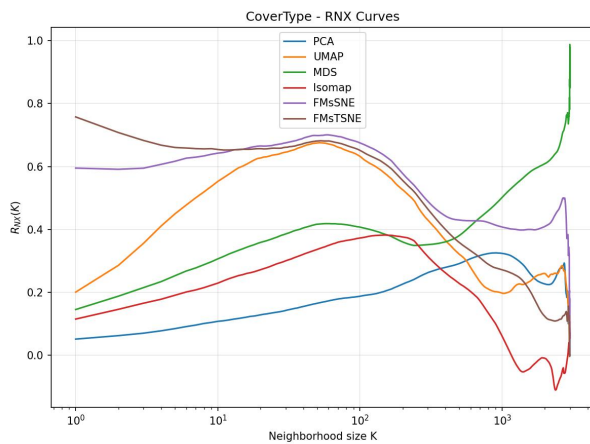
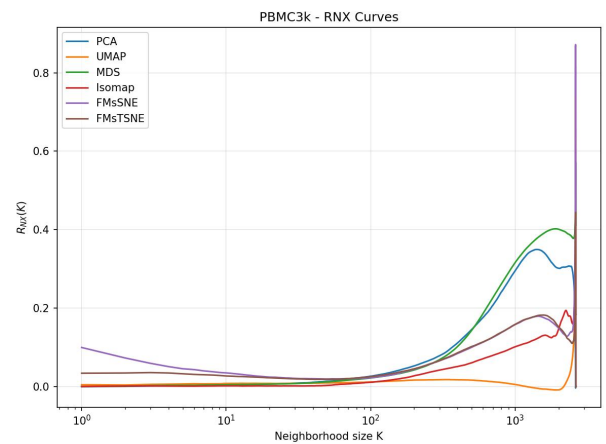
Interestingly, the Phoneme dataset achieves high results for PCA, UMAP FMsSNE and FMsTSNE methods. The strong performance of PCA (0.835) suggests that nonlinear dimensionality reduction may provide limited additional benefit for this dataset, as a substantial portion of the relevant structure appears to be preserved by a linear projection. The PBMC3k dataset appears to be the most challenging, with several methods achieving weak or even negative correlations, whereas FMsSNE remains relatively robust. These results indicate that nonlinear methods, particularly FMsSNE and FMsTSNE, are generally more effective at preserving label-related relationships across diverse datasets, although the extent of their advantage depends on the intrinsic structure of the data.

4.3.2 $R_{NX}(K)$ curves and AUC:

Although the Area Under the Curve (AUC) scores initially suggest that FMsTSNE and FMsSNE provide the optimal embeddings for the datasets, a detailed multi-scale examination of the $R_{NX}(K)$ curves clarifies the precise structural dynamics driving these values.

Method	COIL-20	MNIST	Phoneme	CoverType	Isolet	PBMC3k
PCA	0.399	0.157	0.354	0.166	0.241	0.070
UMAP	0.451	0.320	0.342	0.424	0.389	0.010
MDS	0.446	0.201	0.350	0.352	0.268	0.077
Isomap	0.359	0.181	0.317	0.207	0.275	0.030
FMsSNE	0.665	0.474	0.488	0.574	0.537	0.070
FMsTSNE	0.637	0.220	0.490	0.560	0.506	0.057

Table 4.6: AUC of $R_{NX}(K)$ curves per embedding

(a) *COIL-20*(b) *MNIST*(c) *Phoneme*(d) *Isolet*(e) *CoverType*(f) *PBMC3K*Figure 4.13: R_{NX} curves of embeddings on datasets

R_{NX} curve

For the COIL-20 dataset, FMsTSNE and FMsSNE strictly dominate the local scale (small K values), starting with exceptionally high scores above 0.8. This confirms their mathematical superiority in preserving the exact, immediate nearest neighbours of the high-dimensional data. Conversely, PCA and MDS dominate the macroscopic scale (large K values), reflecting their inher-

ent prioritization of large-scale global variances at the expense of local topologies. UMAP exhibits a uniquely specialized profile, peaking predominantly in the mid-range K values while maintaining relatively lower scores at the extreme local and global bounds.

These broad structural trends are largely mirrored in the Isolet and MNIST dataset, with the exception of FMsSNE in the latter, which demonstrates atypical behaviour by yielding similarly low preservation scores at both the smallest and largest neighbourhood scales. Furthermore, the notably low overall AUC scores observed for MNIST specifically for PCA (0.16), MDS (0.2), and FMsTSNE (0.22). This strongly indicate the presence of severe topological distortions, which fundamentally compromise the spatial integrity of those embeddings. A similar trend is observed for the CoverType dataset.

In contrast, for the Phoneme dataset, the fast multi-scale methods perform particularly well at preserving local neighbourhood structures for small values of (K). However, as the neighbourhood size increases, the remaining techniques exhibit competitive performance in capturing larger-scale relationships. A similar trend is observed for CoverType.

The behaviour of PBMC3k is markedly different and appears to reflect the effects of the “curse of dimensionality.” As shown in the curves, most methods exhibit near-zero slopes over a large range of neighbourhood sizes, followed by an abrupt asymptotic increase for large (K)-neighbourhoods, likely corresponding to transitions occurring after the formation of major clusters.

The Limitation of Relying Solely on AUC

Relying solely on AUC metric can be deceptive if not contextualized by the curves and visual diagnostics. A naive reading of the table would conclude that the fast multi-scale methods are the best DR methods for visualization since they achieved the highest AUC for the datasets except PBMC3K. However, as established in the previous ZADUVis mappings and scatter plot analyses, FMsTSNE often fails to produce coherent visual clusters for COIL-20 for example, generating artificially compressed bands instead. UMAP, despite having a lower overall AUC on COIL-20, is actually the optimal visual mapping.

The discrepancy is explained by the R_{NX} curves as UMAP sacrifices perfect immediate-neighbour preservation (which inflates the AUC of FMsTSNE) in order to peak at the mid-range scale. This specific mid-range preservation corresponds to the clusters sizes, yielding the tightest, most visually distinct, and reliable clusters without the severe geometric distortions introduced by the multi-scale methods.

4.4 Concordance Analysis: $R_{NX}(K)$ vs. ZADU Metrics

A critical step in validating quality assessment frameworks is to determine the statistical agreement between different measurement systems across varying neighbourhood scales.

To evaluate the degree to which neighbourhood preservation at different scales predicts specific structural distortions, we performed a cross-method correlation analysis. Unlike a standard evaluation table, which reports the performance of individual embeddings, this analysis quantifies the statistical agreement between two different quality assessment frameworks— $R_{NX}(K)$ and the ZADU suite—across the five dimensionality reduction methods.

4.4.1 Correlation analysis

Methodology: For each neighbourhood scale $K \in [1, N]$ and each ZADU measure M , we define two vectors, $\mathbf{r}_K$ and $\mathbf{m}$, both of length six (corresponding to PCA, MDS, Isomap, UMAP, FMsSNE, and FMsTSNE): $\mathbf{r}_K = [R_{NX}(K)_1, \dots, R_{NX}(K)_6]$ and $\mathbf{m} = [M_1, \dots, M_6]$. The Pearson correlation coefficient ρ is then calculated for each pair (K, M) . Each cell in the resulting heatmap represents the correlation across our five DR methods, answering the research question: How well does performance in neighbourhood preservation at scale K predict the score of a given ZADU metric?

Interpretation: A high absolute correlation $|\rho| \approx 1$ indicates that the ranking of DR methods provided by $R_{NX}(K)$ is highly consistent with the ranking provided by the ZADU measure.

- **Positive Correlation:** Methods that achieve high neighbourhood preservation also achieve high scores in the corresponding ZADU metric (e.g., Trustworthiness).
- **Negative Correlation:** High neighbourhood preservation is associated with low values in error-based ZADU metrics (e.g., Stress).

The heatmap serves as a meta-evaluation tool to identify which ZADU metrics are redundant with $R_{NX}(K)$ at specific scales and which metrics capture unique structural information that is not reflected in simple neighbourhood preservation curves.

4.4.2 Comprehensive Discussion Across All Datasets

The synthesis of correlation heatmaps across the diverse suite of datasets (COIL-20, MNIST, Phoneme, Isolet, CoverType, and PBMC3k) reveals deeply consistent patterns in metric behavior that transcend the specific geometric characteristics of each manifold. By analyzing how the $R_{NX}(K)$ curve correlates with the ZADU metrics across these datasets, we can identify universal laws of dimensionality reduction evaluation.

Local fidelity and Cluster Integrity Concordance

Across the evaluated datasets, local and cluster integrity metrics generally exhibit high concordance with $R_{NX}(K)$ at small neighbourhood scales, typically manifesting as a strong positive correlation. This relationship fundamentally inverts as the neighbourhood size increases, resulting in negative correlations at larger K scales. A notable exception within the local measures is neighbour_dissimilarity, which remains highly correlated with larger neighbourhood sizes, a behavior consistent with its underlying distance based nature. Similarly, within the cluster integrity category, label_continuity demonstrates a unique scale dependency. At extremely small K sizes, the limited number of samples within the immediate neighbourhood makes it statistically difficult to establish a reliable continuity of labels, leading to lower initial concordance compared to other cluster metrics.

Scale-Dependent Redundancy in Global Measures

Global structural measures exhibit a characteristic scale-dependent redundancy when compared to R_{NX} curves. Specifically, error-based ZADU metrics, such as Stress, show high redundancy at small K values but shift to negative correlations at higher neighbourhood sizes. The logic for this trend lies in the local topology; at small scales, the sum of squared differences between pairwise distances is minimized as the high-dimensional proximities are preserved in the 2D embedding. As the neighbourhood expands, these cumulative errors naturally grow. In contrast, divergence-based measures follow an inverse pattern, showing significant divergence at local scales that progressively minimizes as the neighbourhood size increases.

The Crossover Point and Dataset Scaling

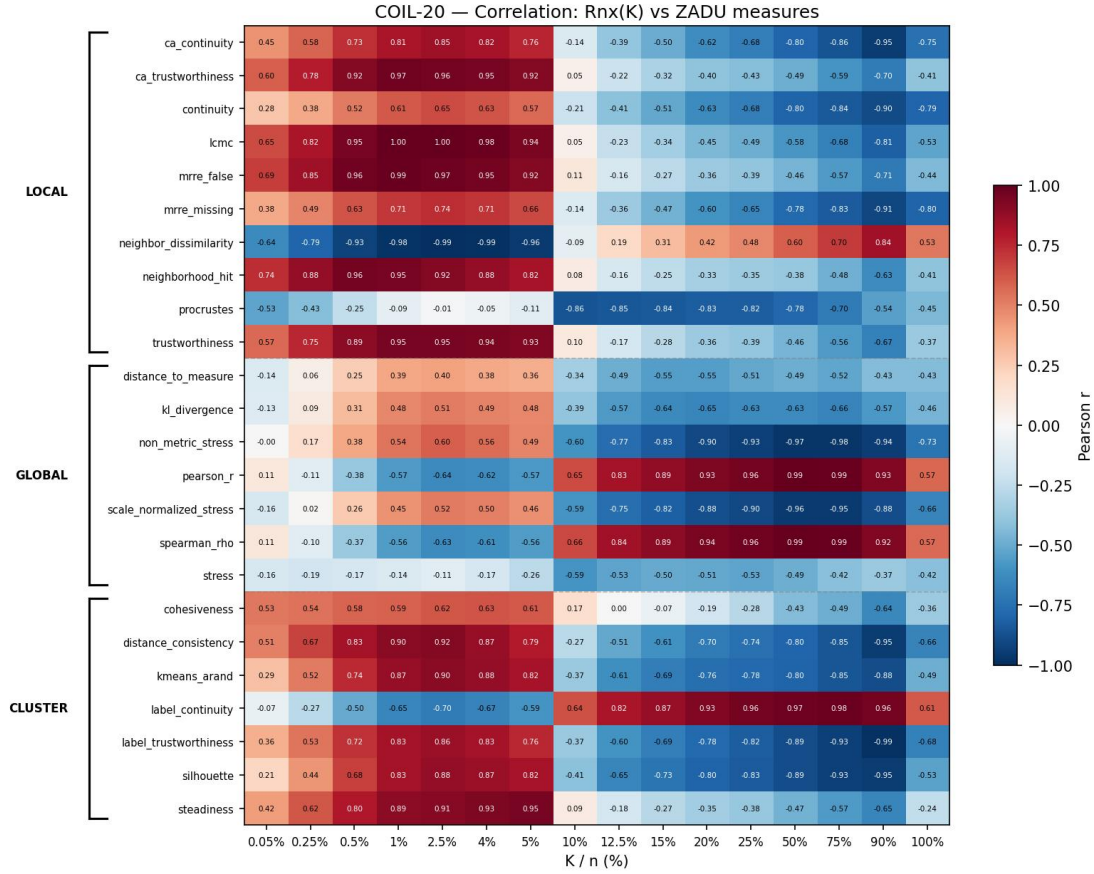
A critical observation in this meta-evaluation is the existence of a crossover neighbourhood size corresponding to the cluster size, where global measures undergo a sign-reversal in their correlation with $R_{NX}(K)$, transitioning from strong positive to strong negative concordance (or vice versa). This transition point effectively marks the boundary where the metric's scale-redundant nature shifts its alignment from local to global structural priorities.

The exceptions: CoverType and PBMC3K.

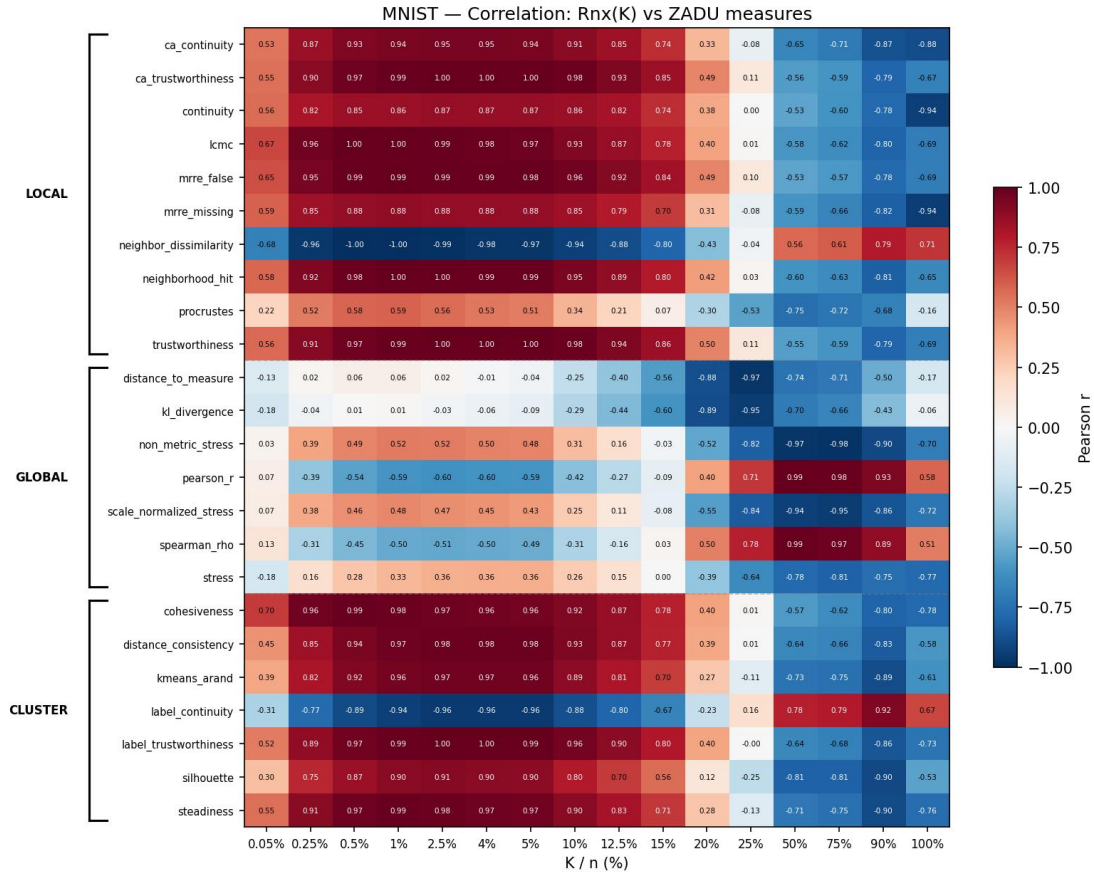
The difference arises from the intrinsic topology of the datasets as they do not contain clearly separated classes. The correlation heatmaps for CoverType indicate that structural fidelity is governed by two different regimes: local neighbourhood preservation dominates the micro-scale assessment, while global geometric and cluster-level consistency define the macroscopic reliability of the embedding. This transition occurs at a "crossover point" corresponding to the intrinsic scale of the data's groupings.

The early correlations for PBMC3K are notably low and fragmented. Because of the "curse of dimensionality," the ambient distances between immediate true biological neighbours are overwhelmed by transcriptomic noise. Consequently, preserving the exact k -nearest neighbours at minimal scales does not yield a statistically reliable local topology. The absence of relevant scores for cluster measure for neighbourhood below the cluster size indicates that the internal structure of the clusters is highly disordered, while the global distribution of the data outside the clusters behaves reasonably well. This is due to the highly imbalanced nature of the datasets. The absence of relevant information on cluster's fidelity measures on smaller is due to the highly class-imbalanced dataset.

A profound structural shift occurs as the neighbourhood size approaches approximately 12% of the dataset. This specific scale is not arbitrary; it corresponds exactly to the average size of the 8 Louvain cell-type clusters present in the data. Once the neighbourhood radius K expands enough to encompass a full biological cluster, the transcriptomic noise is effectively absorbed, and the curse of dimensionality disappears. Afterword, the heatmap reveals a massive synchronized block of redundant behaviour. A sudden high positive correlation across local measures, cluster measures, and global correlation measures simultaneously.



(a) COIL-20



(b) MNIST

Figure 4.14: Correlation heatmaps of ZADU measures and Rank-based metric's neighbourhood size K in size percentage for COIL-20 (top) and MNIST (bottom) datasets

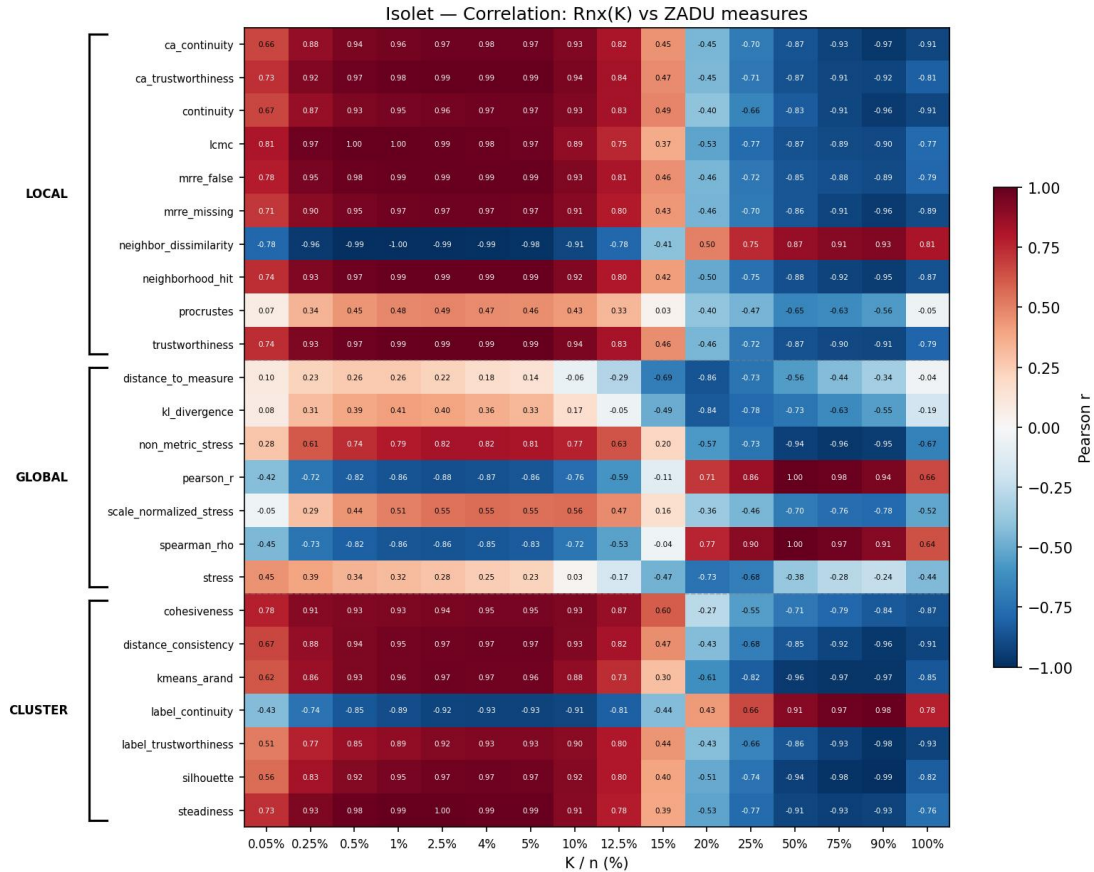
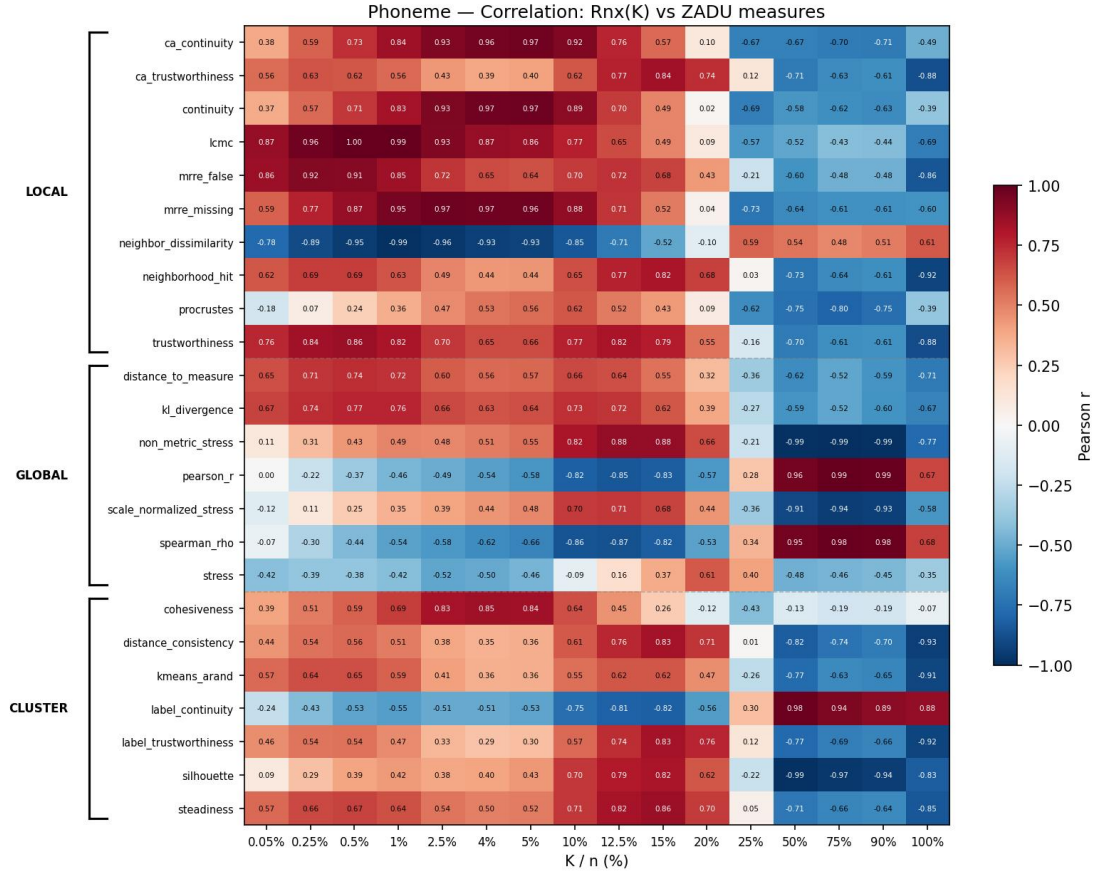
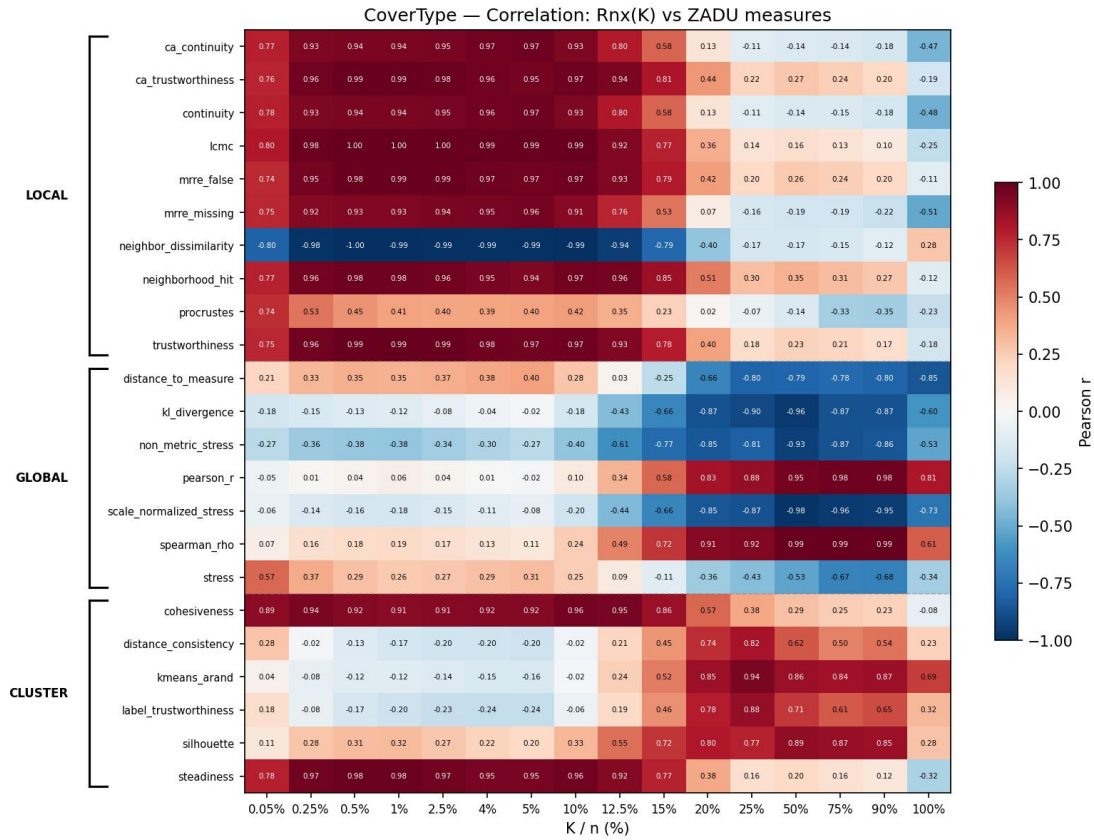
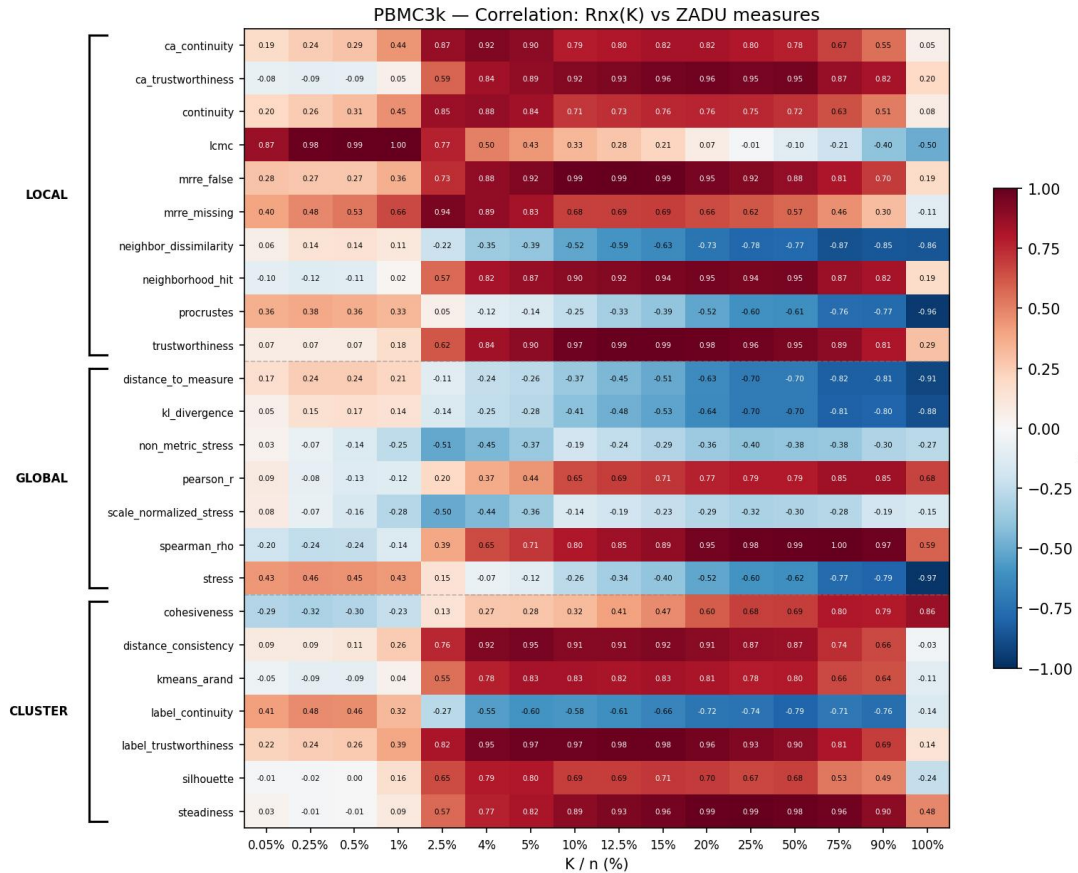


Figure 4.15: Correlation heatmaps of ZADU measures and Rank-based metric's neighbourhood size K in size percentage for Phoneme (top) and Isolet (bottom) datasets



(a) CoverType



(b) PBMC3K

Figure 4.16: Correlation heatmaps of ZADU measures and Rank-based metric's neighbourhood size K in size percentage for CoverType (top) and PBMC3K (bottom) datasets

4.5 Concordance Analysis of Embedding method and $R_{NX}(K)$ per ZADU criteria

This section presents a concordance analysis of the embedding methods by examining their neighbourhood-preservation behavior through the $R_{NX}(K)$ metric, evaluated within the ZADU framework to quantify how consistently each embedding maintains local and global structural relationships

4.5.1 Overview of the Data Structure

For each ZADU measure, a matrix is presented with:

- **Rows:** Six different DR methods in order: PCA, UMAP, MDS, Isomap, FMsSNE and FMsTSNE.
- **Columns:** neighbourhood sizes K (percentiles, from very small to N-2 the maximum K neighbourhood)
- **Cell values:** Pearson correlation between that measure (computed on the embedding) and $R_{NX}(K)$ across the 6 datasets.

A high positive correlation means that as the measure increases, $R_{NX}(K)$ also increases, indicating a positive relation between the two, the measure agrees with R_{NX} . A strong negative correlation indicates an inverse relationship. Magnitude near 0 means no linear association.

Because the correlations are averaged over methods and datasets, they reveal general tendencies rather than method-specific behaviour.

4.5.2 Local measures

Local distortion measures evaluate the fidelity of a dimensionality reduction embedding by quantifying how well the k-nearest neighbour (kNN) structure of the high-dimensional space is preserved in the low-dimensional projection. The correlation analysis between local ZADU measures and the rank-based quality curve ($R_{NX}(K)$) reveals several key insights. The correlation analysis between these global ZADU measures and the rank-based quality curve $R_{NX}(K)$ (Pearson's r across 6 datasets and 6 DR methods) provides insights into how well RN X captures local fidelity at small neighbourhood scales.

First, neighbourhood perservation measures such as Trustworthiness, Continuity and class aware-TNC variants show very strong positive correlations with $R_{NX}(K)$ at small to medium neighbourhood sizes ($2.5\% \leq K \leq 15\%$ of N) across all embedding methods. This positive alignment confirms that for these "higher-is-better" metrics, a high ZADU score is a reliable predictor of high neighbourhood preservation. In contrast, LCMC is strictly focused on micro-scale neighbourhoods ($K \leq 2.5\%$); its correlation drops significantly beyond $K = 10\%$, identifying it as a specialized ultra-local metric. MRRE is split into two variants MRRE_failed to detect false neighbours (close in LD but not in HD) and MRRE_missing for missing neighbours (close in HD space but not in LD). We can see that all methods create new neighbours in LD for K mainly between 2.5% and 25%.

Second, the framework incorporates error-based metrics which require an inverse interpretation. The Procrustes Measure and Neighbour Dissimilarity are mathematically defined such that a lower value represents higher embedding quality. Consequently, when analysed against $R_{NX}(K)$, we observe and search for a negative correlation where successful neighbourhood preservation is statistically tied to a reduction in geometric and graph-based errors. The Procrustes Measure provides a strong negative correlation signature for linear and Gaussian-based methods (PCA, MDS, and FMsSNE).

The scale-dependent predictive power of Neighbour Dissimilarity varies significantly according to the chosen algorithm. FMsSNE exhibits universal consistency for this metric, maintaining a reliable negative correlation until K reaches approximately 90% of the dataset size N . In contrast, MDS demonstrates an extended macroscopic reach with strong negative concordance up to medium neighbourhood scale (50%). For the remaining methods, including PCA, Isomap, UMAP and FMsTSNE, the predictive power is more restricted, with reliable negative correlations typically concentrated at small neighbourhood scales (between 1% and 15%), with UMAP additionally focused ultra-local neighbourhoods below 1%; highlighting the critical importance of matching ZADU error metrics to the specific structural priorities and scales of each embedding technique.

Finally, The sharp correlation change at 100%, particularly pronounced for Isomap, is a consequence of the fundamental conflict between manifold unrolling and global Euclidean preservation. The $R_{NX}(K)$ metric at $K = 100\%$ effectively measures how well the embedding preserves the entire dataset's neighbourhood structure, which is essentially an assessment of global Euclidean consistency. A 'perfect' Isomap embedding that identifies the true geodesic manifold will appear "wrong" to a global Euclidean.

These findings confirm that ZADU local metrics, when matched to the appropriate scale K , provide a comprehensive and multi-perspective audit of local embedding quality that simple aggregated scores cannot capture. FMsSNE proved to be an all-rounder for local measures.

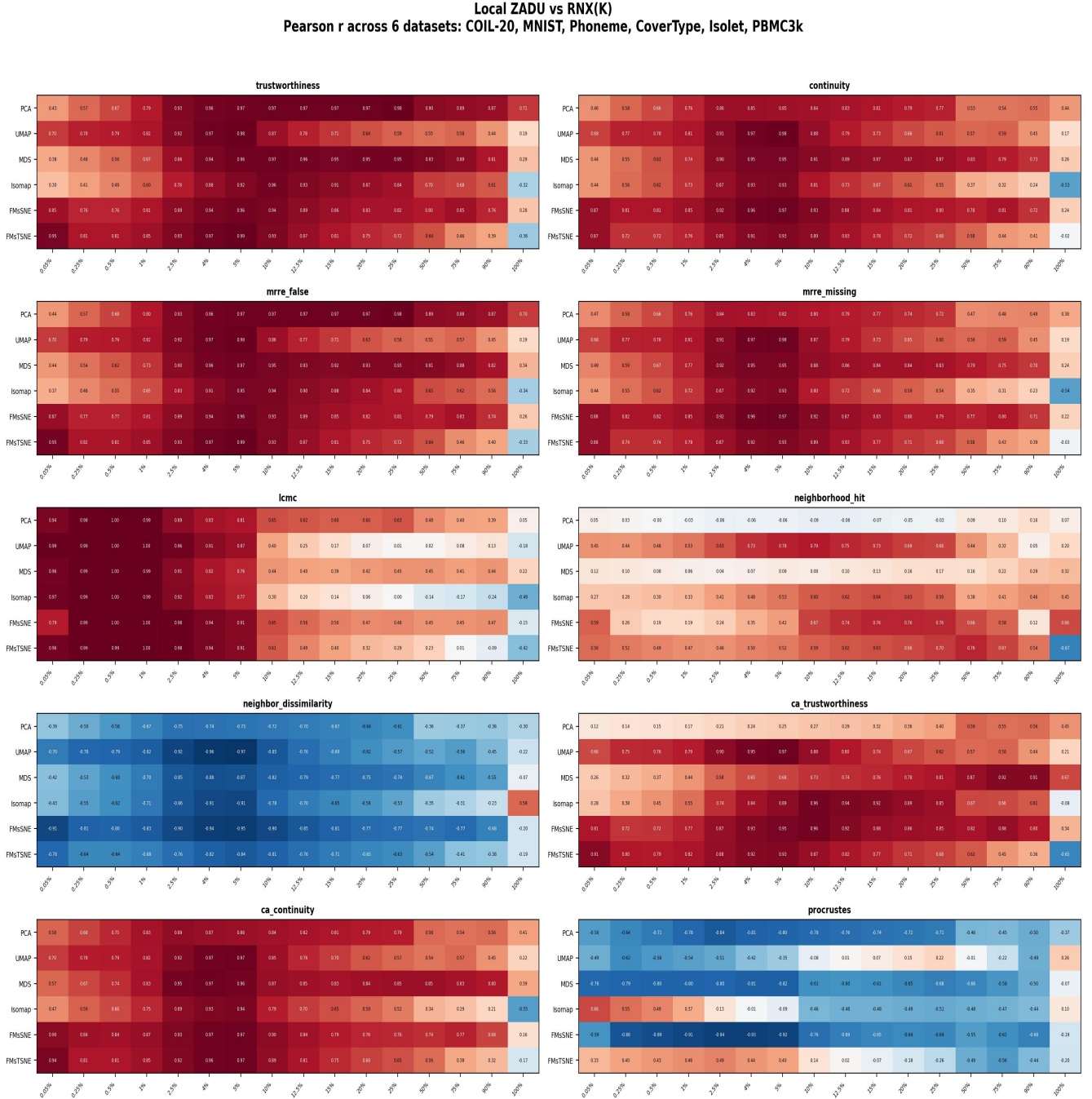


Figure 4.17: Correlation of DR methods (in order from top: PCA, UMAP, MDS, Isomap, FMsSNE and FMsTSNE) with R_{NX} across datasets per ZADU local measure starting from top to bottom then left to right we have: trustworthiness, mrre_false, lcmc, neighbour_dissimilarity, ca_continuity, continuity, mrre_missing, neighbourhood_hit, ca_trustworthiness, Procrustes

4.5.3 Cluster level measures

Cluster-level distortion measures evaluate how well the grouping structure of the high-dimensional space is preserved in the low-dimensional embedding. Unlike local measures that focus on point-wise neighbourhoods, cluster measures assess whether entire clusters remain separated (steadiness) and internally coherent (cohesiveness), and whether the relative order of cluster distances is maintained. ZADU implements several cluster measures, including Steadiness & Cohesiveness (S&C), Distance Consistency, Silhouette, k-means Adjusted Rand Index (ARI), and label-based Trustworthiness & Continuity. The correlation analysis between these cluster ZADU measures and the

rank-based quality curve $R_{NX}(K)$ (Pearson’s r) across 6 datasets and 6 DR methods) reveals how well R_{NX} captures cluster-level fidelity across different neighbourhood scales.

Geometric integrity metrics such as Silhouette and Distance Consistency provide critical insights into cluster compactness and separation. Silhouette scores show relatively strong positive correlations for non-linear methods including UMAP, FMsSNE, FMsTSNE, and Isomap within the range $10\% \leq K \leq 25\%$, with Isomap and FMsTSNE maintaining high correlation levels up to 90% of N . In contrast, PCA exhibits consistently weak or negative correlations, marking a significant failure of linear projections to track cluster compactness accurately. Similarly, Distance Consistency exhibits high positive correlations for UMAP and Isomap between 10% and $25\%K$, while Isomap sustains this alignment even at $90\%K$ due to its global distance preservation. While FMsTSNE shows strong correlations from 15% to $90\%K$, FMsSNE remains moderate, and both MDS and PCA show only weak correlations across all scales.

Label recovery metrics, such as the k -means Adjusted Rand Index (ARI) and label-based variants of Trustworthiness and Continuity, evaluate how well the embedding preserves the categorical structure of the data. The ARI follows a pattern similar to Distance Consistency but reveals that MDS can partially recover class labels with scores often above 0.5, suggesting that its global structure retains some semantic information despite local preservation failures. Label-based Trustworthiness shows that PCA’s global variance preservation aligns with class-aware rank order at macroscopic scales starting from $50\% K$. UMAP peaks earlier between 10% and 25% , while Isomap and FMsTSNE maintain relevance through the 15% to 90% range. Label-based Continuity yields unexpected positive correlations for PCA between 2.5% and $25\%K$, but for MDS, Isomap, and FMsTSNE, the correlations eventually become negative, indicating a fundamental clash between class-aware continuity and the rank-based criterion of R_{NX} .

The duality of cluster distortion is best captured by the distinct behaviors of Steadiness and Cohesiveness, which measure cluster-level intrusions and extrusions respectively. Steadiness shows highly negative correlations for very small neighbourhoods ($K \leq 5\%$) for MDS, Isomap, and FMsSNE. This counter-intuitive relationship, where lower steadiness appears associated with higher R_{NX} , arises because micro-scale neighbourhood preservation is dominated by point-wise relationships that are not yet affected by cluster-level intrusions. Conversely, Cohesiveness is highly positively correlated with R_{NX} up to $50\%K$ for most methods, confirming that fewer "tears" or extrusions generally lead to higher neighbourhood preservation. Notable exceptions include MDS and Isomap, where correlations turn negative at certain scales, while FMsSNE maintains high positive correlation up to $90\%K$.

In summary, cluster-level analysis demonstrates that $R_{NX}(K)$ at medium to large scales serves as a strong proxy for global cluster quality metrics like Silhouette and Distance Consistency, though the relationship remains highly method-specific. The stark differences between the correlations for Steadiness and Cohesiveness reinforce the necessity of treating intrusions and extrusions as distinct error types within the ZADU framework. Furthermore, the surprising performance of PCA and MDS on label-based measures underscores that even linear methods can preserve class-rank order at macroscopic scales despite crushing local structure. These findings validate the strategy of reporting R_{NX} curves across the full range of K and complementing them with diverse cluster-level measures for a comprehensive fidelity assessment.

Cluster ZADU vs RN(X)
Pearson r across 6 datasets: COIL-20, MNIST, Phoneme, CoverType, Isolet, PBMC3k

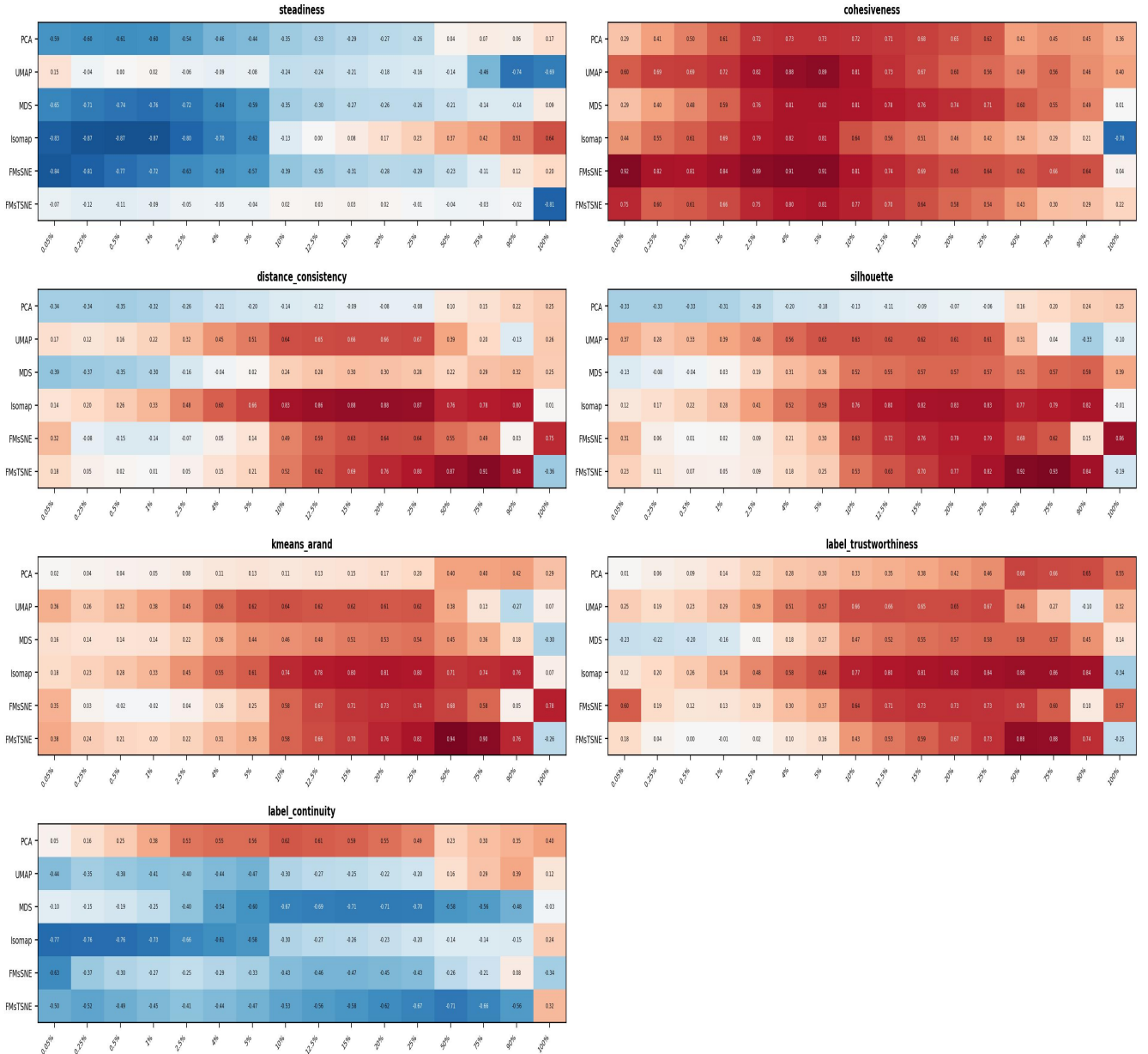


Figure 4.18: Correlation of DR methods (in order from top: PCA, UMAP, MDS, Isomap, FMsSNE and FMsTSNE) with R_{NX} across datasets per ZADU cluster measure starting from top to bottom then left to right we have: steadiness, distance_consistency, kmeans_arand, label_continuity, cohesiveness, silhouette, label_trustworthiness

4.5.4 Global measures

Global distortion measures assess the preservation of all pairwise distances or correlations across the entire dataset, answering the question: Does the overall shape and large-scale geometry of the data remain faithful? Unlike local measures, which focus on small neighbourhoods, global measures are sensitive to the “crowding problem” in high dimensions and can reveal whether an embedding maintains the relative ordering of distant points. ZADU implements several global measures, including Stress (and its variants), Kullback-Leibler (KL) divergence, Distance-to-Measure (DTM), Topographic Product, Pearson’s r and Spearman’s ρ on distance ranks. The correlation analysis

between these global ZADU measures and the rank-based quality curve $R_{NX}(K)$ (Pearson’s r across 6 datasets and 6 DR methods) provides insights into how well R_{NX} captures global fidelity at large neighbourhood scales.

First, Stress (Kruskal’s raw stress), Non-metric Stress and Scale-normalised Stress are error measures where lower values are better. For raw Stress only PCA, MDS, and FMsSNE exhibit the expected strong negative correlations, making it a reliable proxy for these methods. In contrast, Isomap and FMsTSNE often show highly positive or fragmented correlations at small neighbourhood scales, suggesting that raw stress is a poor or biased indicator of their structural truth. Scale-normalized Stress generally demonstrates improved stability across methods, though it notably fails for UMAP at small neighbourhood sizes ($K \leq 10\%$), where the correlation with structural preservation is inconsistent. Ultimately, Non-metric Stress emerges as the most robust and universally applicable stress-based criterion, providing consistent and strong concordance across all evaluated dimensionality reduction methods for neighbourhoods ranging from 10% to 90% of N . Furthermore, for algorithms like PCA, MDS, and FMsSNE, this reliability is sustained even at smaller neighbourhood scales starting from 2.5% of N , identifying Non-metric Stress as the superior diagnostic tool for evaluating global geometric fidelity.

Distribution preservation criteria show no consistent trend. Kullback-Leibler (KL) divergence (comparing high-dimensional and low-dimensional similarity distributions) shows weak to moderate negative correlations for most methods, with no clear pattern across K . This indicates that KL divergence captures a different aspect of distortion than R_{NX} , which is based on strict rank comparisons rather than probabilistic similarity matching. Strangely, FMsTSNE that explicitly optimizes a KL-based objective, remains notably weak across nearly all ranges, except the limit 100%. This lack of concordance likely stems from the fact that these heatmaps represent the Pearson r calculated across six diverse datasets. Distance-to-Measure (DTM) (a geometric measure comparing point cloud distributions) shows weak positive correlations for most methods and K , with no consistent trend. This suggests that DTM is largely independent of rank-based neighbourhood preservation, making it a complementary rather than a proxy metric.

Topographic Product (originally designed for self-organizing maps) exhibits very high positive correlations with $R_{NX}(K)$ across all K for MDS. The omission for the other methods is due to the measure’s applicability as Topographic Product assumes that the low-dimensional embedding has a continuous, well-defined “map” structure. Therefor, for methods producing disconnected clusters or non-continuous manifolds, the Topographic Product is undefined or meaningless.

Pearson’s r (distance matrix correlation) and Spearman’s ρ (rank correlation) emerge as highly informative indicators of structural fidelity, particularly as the neighbourhood scale K approaches the macroscopic range above 10% of the dataset. However, the concordance analysis reveals that the predictive power of these metrics is highly method-dependent. For linear and distance-based methods such as MDS and PCA, both correlation metrics exhibit strong positive concordance with structural preservation even at significantly smaller scales, starting as early as 1% of N . This early stability reflects the fundamental objective of these algorithms to preserve global distance relationships throughout the entire embedding process. A unique behavior is observed for FMsSNE in conjunction with Spearman’s ρ , which maintains a robust and high correlation starting from the smallest neighbourhoods ($K \approx 1\%$), proving its effectiveness as a universal audit for multi-scale Gaussian embeddings. Overall, Spearman’s ρ demonstrates superior performance across all methods and neighbourhood sizes, consistently yielding higher correlation values than Pearson’s r . This superiority is due to its rank-based mathematical formulation, which captures monotonic relationships and is thus more compatible with the rank-ordering logic of the $R_{NX}(K)$ curve.

Global ZADU vs $R_{NX}(K)$
Pearson r across 6 datasets: COIL-20, MNIST, Phoneme, CoverType, Isolet, PBMC3k

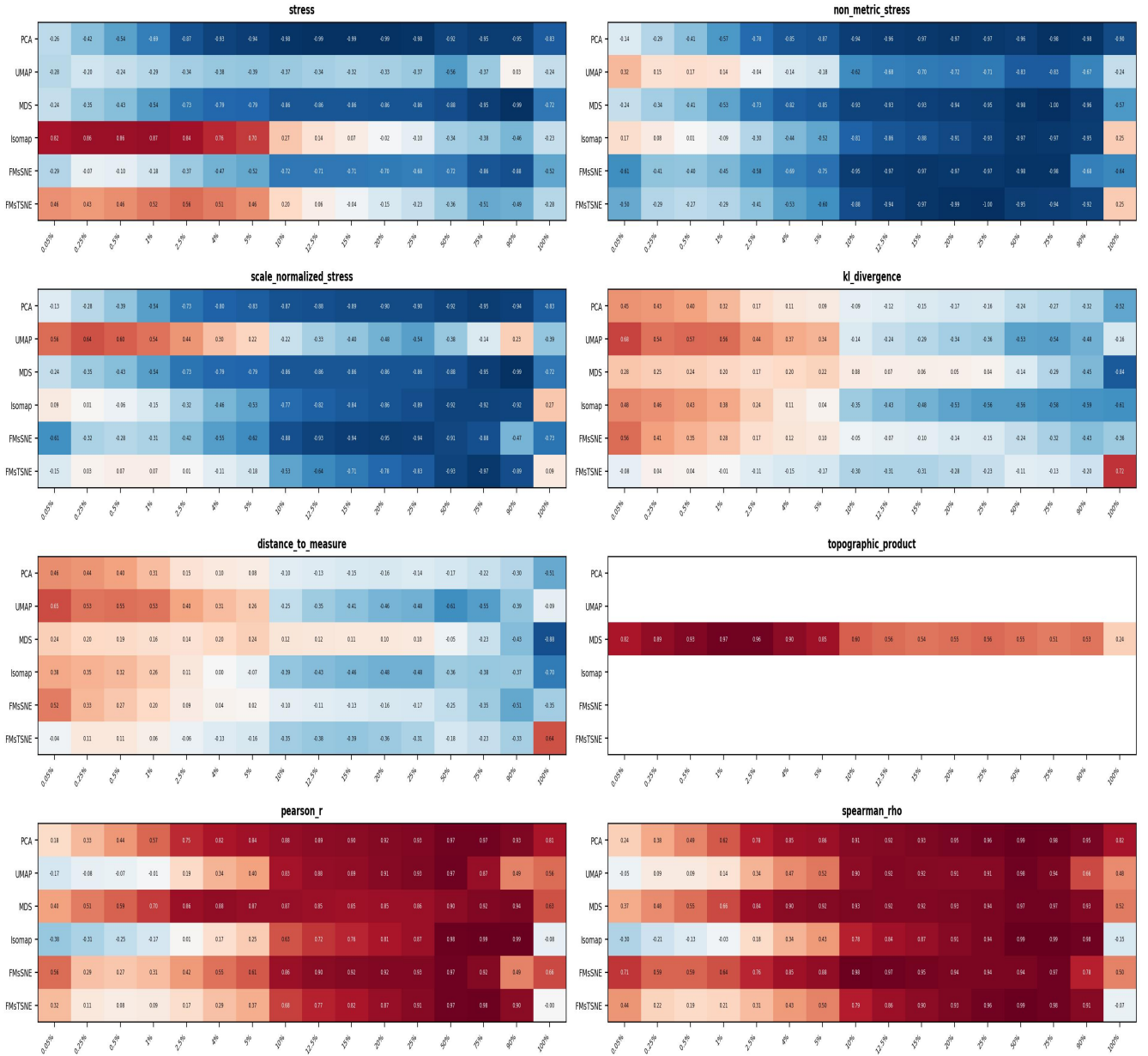


Figure 4.19: Correlation of DR methods (in order from top: PCA, UMAP, MDS, Isomap, FMsSNE and FMsTSNE) with R_{NX} across datasets per ZADU global measure starting from top to bottom then left to right we have: stress, scale_normalized_stress, distance_to_measure, Pearson, non_metric_stress, kl_divergence, topographic_product, spearman

Chapter 5

Conclusion

The study consists of a systematic comparative analysis of multiple dimensionality reduction methods (Principal Component Analysis (PCA), Multidimensional Scaling (MDS), and Isomap, Uniform Manifold Approximation and Projection (UMAP), FMsSNE, and FMsTSNE), across six benchmark datasets exhibiting diverse topological characteristics (COIL-20, MNIST, Phoneme, Isolet, CoverType, and PBMC3k).

A central component of this study was the application of the ZADU framework, which enabled the simultaneous assessment of 17 distortion metrics spanning local neighbourhood fidelity, cluster preservation, and global geometric consistency. Additional evaluation criteria included label-cluster correlation across embedding dimensions, rank-based measures, and silhouette analysis.

Beyond quantitative evaluation, qualitative visual inspection was also performed. Embeddings were first assessed through visual class separability, then further analysed using CheckViz Voronoi tessellations to localize intrusions and Reliability Maps to detect manifold tearing and extrusions. These analyses provided insight into where and why specific algorithms failed to preserve intrinsic data structures.

Finally, a novel concordance analysis was introduced. It quantifies the statistical agreement between established rank-based $R_{NX}(K)$ curves and point-wise ZADU metrics, revealing the characteristic behavioural signatures of each embedding technique, and the performance on each dataset.

The central finding of this study is that dimensionality reduction is inherently a lossy and imperfect process, in which each algorithm prioritizes different aspects of the underlying high-dimensional manifold. Although certain dimensionality reduction techniques may appear to produce well-separated classes, such separations can be misleading or even “hallucinatory” when contrasted with the evolution of silhouette scores across dimensions. Likewise, relying exclusively on aggregate measures such as AUC may produce inaccurate interpretations if the underlying curves are not jointly examined, since some methods primarily preserve local neighbourhood structures while others better retain global geometric organization.

Our results further demonstrate that there is no universally optimal dimensionality reduction technique. Instead, the effectiveness of a given method strongly depends on the intrinsic properties and topology of the dataset under study, with certain algorithms consistently performing better for specific data types. However, these apparent successes should not be accepted at face value: even when a method achieves strong visual separation or favourable numerical scores, its quality must still be validated through complementary metrics and cross-analysis. Altogether, this work highlights the necessity of combining multiple quantitative and qualitative evaluation criteria to obtain a reliable and nuanced assessment of dimensionality reduction methods.

Building upon these conclusions, several avenues for future research emerge to refine the validation and development of dimensionality reduction techniques. A primary focus involves a deeper

structural analysis of how individual ZADU distortion measures operate at a granular level; such an investigation could determine if an ensemble of specific metrics might be synthesized into a "unified reliability score" suitable for general-purpose applications. Furthermore, significant value lies in establishing a more rigorous mathematical relationship between the characteristic profiles of $R_{NX}(K)$ curves and specific qualitative distortion types, effectively translating abstract preservation values into actionable insights regarding manifold "tears" or "folds." Ultimately, these theoretical advancements should pave the way for the design of hybrid dimensionality reduction algorithms that empower practitioners to explicitly navigate the inherent trade-off between local and global structural preservation through a controllable, tunable parameter.

5.0.1 Use of Generative AI tools

Generative artificial intelligence tools were used during the preparation of this thesis to assist with text refinement, including improving clarity, grammar, and wording. An AI-assisted coding agent was also used to support software development tasks and code refinement.

All AI-generated text was reviewed, edited, and validated by the author before inclusion in the thesis. Likewise, all AI-assisted code was reviewed, modified where necessary, and tested by the author before being used in the research.

Bibliography

- [1] Jock Blackard. *Coverttype*. UCI Machine Learning Repository. 1998. DOI: 10.24432/C50K5N. URL: <https://doi.org/10.24432/C50K5N>.
- [2] Jan Bocianowski et al. “Comparison of Pearson’s and Spearman’s correlation coefficients for selected traits of *Pinus sylvestris* L.” In: *Biometrical Letters* 61.2 (2024), pp. 115–135. DOI: 10.2478/bile-2024-0008. URL: <https://doi.org/10.2478/bile-2024-0008>.
- [3] Cyril de Bodt et al. “Fast Multiscale Neighbor Embedding”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.4 (2022), pp. 1546–1560. DOI: 10.1109/TNNLS.2020.3042807.
- [4] Cyril de Bodt et al. *Low-dimensional embeddings of high-dimensional data*. 2025. arXiv: 2508.15929 [cs.LG]. URL: <https://arxiv.org/abs/2508.15929>.
- [5] Ron Cole and Mark Fanty. *ISOLET*. UCI Machine Learning Repository. 1991. DOI: 10.24432/C51G69. URL: <https://doi.org/10.24432/C51G69>.
- [6] C. de Bodt et al. “Fast Multiscale Neighbor Embedding”. In: *IEEE Trans. Neural Netw. Learn. Syst.* (2020), pp. 1–15. DOI: 10.1109/TNNLS.2020.3042807.
- [7] Trevor Hastie, Andreas Buja, and Robert Tibshirani. “Penalized Discriminant Analysis”. In: *The Annals of Statistics* 23.1 (1995), pp. 73–102. DOI: 10.1214/aos/1176324456. URL: <https://doi.org/10.1214/aos/1176324456>.
- [8] Geoffrey E. Hinton and Sam T. Roweis. “Stochastic Neighbor Embedding”. In: *Neural Information Processing Systems*. 2002. URL: <https://api.semanticscholar.org/CorpusID:20240>.
- [9] Hyeon Jeon et al. “ZADU: A Python Library for Evaluating the Reliability of Dimensionality Reduction Embeddings”. In: *2023 IEEE Visualization and Visual Analytics (VIS)*. 2023. URL: <https://www.hyeonjeon.com/assets/pdf/jeon23vis.pdf>.
- [10] Ludovic Journaux John Aldo Lee Cyril de Bodt and Lucile Sautot. “Proximities in dimensionality reduction”. English. In: *Handbook of Proximity Relations*. Publisher Copyright: © The Editors and Contributors Severally 2022. Edward Elgar Publishing Ltd., Jan. 2022, pp. 165–203. ISBN: 9781786434777.
- [11] Ian T. Jolliffe and Jorge Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (Apr. 2016), p. 20150202. ISSN: 1364-503X. DOI: 10.1098/rsta.2015.0202. eprint: <https://royalsocietypublishing.org/rsta/article-pdf/doi/10.1098/rsta.2015.0202/1381479/rsta.2015.0202.pdf>. URL: <https://doi.org/10.1098/rsta.2015.0202>.
- [12] Yann LeCun, Corinna Cortes, and CJ Burges. “MNIST handwritten digit database”. In: *ATT Labs [Online]* 2 (2010). URL: <http://yann.lecun.com/exdb/mnist>.

- [13] John A. Lee and Michel Verleysen. “Quality assessment of dimensionality reduction: Rank-based criteria”. In: *Neurocomputing* 72.7 (2009). Advances in Machine Learning and Computational Intelligence, pp. 1431–1443. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2008.12.017>. URL: <https://www.sciencedirect.com/science/article/pii/S0925231209000101>.
- [14] Laurens van der Maaten. “Accelerating t-SNE using Tree-Based Algorithms”. In: *Journal of Machine Learning Research* 15.93 (2014), pp. 3221–3245. URL: <http://jmlr.org/papers/v15/vandemaaten14a.html>.
- [15] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: <http://jmlr.org/papers/v9/vandemaaten08a.html>.
- [16] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML]. URL: <https://arxiv.org/abs/1802.03426>.
- [17] Samer A. Nene, Shree K. Nayar, and Hiroshi Murase. *Columbia Object Image Library (COIL-20)*. Tech. rep. CUCS-005-96. Department of Computer Science, Columbia University, Feb. 1996. URL: <https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php>.
- [18] “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *CoRR* abs/1912.01703 (2019). arXiv: 1912.01703. URL: <http://arxiv.org/abs/1912.01703>.
- [19] Carlos Ramos-Carreño et al. “scikit-fda: A Python Package for Functional Data Analysis”. In: *Journal of Statistical Software* 109.2 (2024). DOI: 10.18637/jss.v109.i02.
- [20] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427. URL: <https://www.sciencedirect.com/science/article/pii/0377042787901257>.
- [21] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. “A Global Geometric Framework for Nonlinear Dimensionality Reduction”. In: *Science* 290.5500 (2000), pp. 2319–2323. DOI: 10.1126/science.290.5500.2319. URL: <https://www.science.org/doi/abs/10.1126/science.290.5500.2319>.
- [22] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. “Scanpy: large-scale single-cell gene expression data analysis”. In: *Genome Biology* 19.1 (2018), p. 15. DOI: 10.1186/s13059-017-1382-0.
- [23] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, et al. “Massively parallel digital transcriptional profiling of single cells”. In: *Nature Communications* 8 (2017), p. 14049. DOI: 10.1038/ncomms14049.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl