Topological Singularity Detection at Multiple Scales

Julius von Rohrscheidt [1 2] **Bastian Rieck** [1 2]

Abstract

The manifold hypothesis, which assumes that data lies on or close to an unknown manifold of low intrinsic dimension, is a staple of modern machine learning research. However, recent work has shown that real-world data exhibits distinct non-manifold structures, i.e. *singularities*, that can lead to erroneous findings. Detecting such singularities is therefore crucial as a precursor to interpolation and inference tasks. We address this issue by developing a topological framework that (i) quantifies the local intrinsic dimension, and (ii) yields a *Euclidicity* score for assessing the 'manifoldness' of a point along multiple scales. Our approach identifies singularities of complex spaces, while also capturing singular structures and local geometric complexity in image data.

1. Introduction

The ever-increasing amount and complexity of real-world data necessitate the development of new methods to extract less complex but still *meaningful* representations of the underlying data. While numerous methods for approaching this representation learning problem exist, they all share a common assumption: the underlying data is supposed to be close to a manifold with small intrinsic dimension, i.e. while the input data may have a large ambient dimension N, there is an n-dimensional manifold with $n \ll N$ that best describes the data. For some data sets, such as natural images, this *manifold hypothesis* is appropriate (Carlsson, 2009). However, recent research shows evidence that the manifold hypothesis does not necessarily hold for complex data sets (Brown et al., 2023), and that manifold learning techniques tend to fail for non-manifold data (Rieck & Leitte, 2015; Scoccola & Perea, 2023). These failures are often the result of *singularities*, i.e. regions of a space that violate the properties of a manifold (see Section 2 for details).

[1]Helmholtz Munich [2]Technical University of Munich. Correspondence to: Bastian Rieck <bastian.rieck@helmholtz-munich.de>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

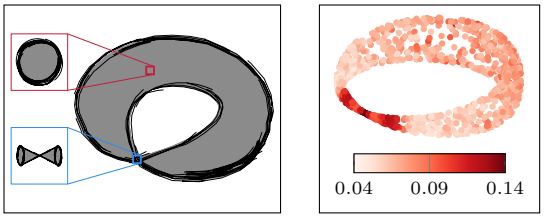


Figure 1: Overview of our method. Given a space with *singularities*, our *Euclidicity* score measures the deviation from a space to a Euclidean model space. Here, *Euclidicity* highlights the singularity at the 'pinch point.' Please refer to Section 4 for more details.

For example, the 'pinched torus,' an object obtained by compressing a random region of the torus (specifically, a meridian) to a single point, fails to satisfy the manifold hypothesis at the 'pinch point:' this point, unlike all other points, does *not* have a neighbourhood homeomorphic to $\mathbb{R}^2$ (see Fig. 1 for an illustration). Since singularities—in contrast to outliers arising from incorrect labels, for example—often carry relevant information (Jakubowski et al., 2020), new tools for detecting and handling non-manifold regions in spaces are needed.

Our contributions. We develop an *unsupervised representation learning framework for detecting singular regions in point cloud data*. Our framework is agnostic with respect to geometric or stochastic properties of the underlying data and only requires a notion of intrinsic dimension of neighbourhoods, which, crucially, is allowed to vary across different points. Our approach is based on a novel formulation of persistent local homology (PLH), a method for assessing the shape of neighbourhoods at multiple scales of locality. We use PLH to (i) estimate the intrinsic dimension of a point locally, and (ii) define *Euclidicity*, a novel quantity that measures the deviation of a point from being Euclidean. We also provide theoretical guarantees on the approximation quality for certain classes of spaces, including manifolds. *Euclidicity* yields a complementary perspective on data, highlighting regions where the manifold hypothesis breaks down. We show the utility of this perspective experimentally on several data sets, ranging from spaces with known singularities to high-dimensional image data sets.

2. Mathematical Background

We first provide an overview of persistent homology, and stratified spaces, as well as their relation to *local homology*. The former concept constitutes a generic framework for assessing complex data at multiple scales by measuring topological characteristics such as 'holes' and 'voids' (Edelsbrunner & Harer, 2010), while the latter serves as a general setting to describe singularities, in which our framework admits advantageous properties.

Persistent homology. Persistent homology is a method for computing topological features at different scales, capturing an intrinsic notion of relevance in terms of spatial scale parameters. Given a finite metric space $(\mathbb{X}, \mathrm{d})$, the *Vietoris–Rips complex* at step t is defined as the abstract simplicial complex $\mathcal{V}(\mathbb{X}, t)$, in which an abstract k-simplex $(x_0, \ldots, x_k)$ of points in $\mathbb{X}$ is spanned if and only if $\mathrm{d}(x_i, x_j) \leq t$ for all $0 \leq i \leq j \leq k$.[1] For $t_1 \leq t_2$, the inclusions $\mathcal{V}(\mathbb{X}, t_1) \hookrightarrow \mathcal{V}(\mathbb{X}, t_2)$ yield a filtration, i.e. a sequence of nested simplicial complexes, which we denote by $\mathcal{V}(\mathbb{X}, \bullet)$. Applying the ith homology functor to this collection of spaces and inclusions between them induces maps on the homology level $f_i^{t_1,t_2} \colon \mathrm{H}_i(\mathcal{V}(\mathbb{X}, t_1)) \to \mathrm{H}_i(\mathcal{V}(\mathbb{X}, t_2))$ for any $t_1 \leq t_2$. The ith *persistent homology (PH)* of $\mathbb{X}$ with respect to the Vietoris-Rips construction is defined to be the collection of all these ith homology groups, together with the respective induced maps between them, and denoted by $\mathrm{PH}_i(\mathcal{V}(\mathbb{X}, \bullet))$. PH can therefore be viewed as a tool that keeps track of topological features such as holes and voids on multiple scales. For a more comprehensive introduction to PH in the context of machine learning, see Hensel et al. (2021). The so-called 'creation' and 'destruction' times of these features are summarised in a *persistence diagram* $\mathcal{D} \subset \mathbb{R} \times \mathbb{R} \cup \{\infty\}$, where any point $(b, d) \in \mathcal{D}$ corresponds to a homology class that arises at filtration step b, and lasts until filtration step d. The difference $|d - b|$ is referred to as the lifetime or eponymous *persistence* of this homology class. There are several distance measures for comparing persistence diagrams, one of them being the *bottleneck distance*, defined as $\mathrm{d_B}(\mathcal{D}, \mathcal{D}') := \inf_\gamma \sup_{x \in \mathcal{D}} \|x - \gamma(x)\|_\infty$, where γ ranges over all bijections between $\mathcal{D}$ and $\mathcal{D}'$.

Stratified spaces. Manifolds are widely studied and particularly well-behaved topological spaces as they locally resemble Euclidean space near any point. However, spaces that arise naturally often violate this local homogeneity condition (see Fig. 2 for an example). *Stratified spaces* generalise the concept of a manifold to address *singular spaces*. Being intrinsically capable of describing a wider class of spaces, we argue that stratified spaces are the right tool to analyse real-world data. In particular, the intrinsic dimension of points is allowed to vary in this setting, thus leading to high flexibility in comparison to manifolds. Subsequently, we define stratified spaces in the setting of simplicial complexes. A stratified simplicial complex of dimension 0 is a finite set of points with the discrete topology. A stratified simplicial complex of dimension n is an n-dimensional simplicial complex X, together with a filtration of closed subcomplexes $X = X_n \supset X_{n-1} \supset X_{n-2} \supset \cdots \supset X_{-1} = \emptyset$ such that $X_i \setminus X_{i-1}$ is an i-dimensional manifold for all i, and such that every point $x \in X$ possesses a distinguished local neighbourhood $U \cong \mathbb{R}^k \times c^\circ L$ in X, where L is a compact stratified simplicial complex of dimension $n-k-1$ and c° refers to the open cone construction (see Appendix A.1).

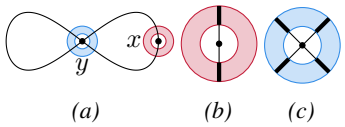


Figure 2: (a): Non-manifold space. (b): Annulus around a regular point x. (c): Annulus around a singular point. The neighbourhood around y is different from all others.

> If there exists a neighbourhood U of x that is homeomorphic to $\mathbb{R}^n$, we say that x is a *regular* point, otherwise we call x a *singularity*.

If X is a manifold, then independently of the point under consideration, L is given by a sphere since for a manifold, *any* point is regular by definition. By contrast, a small neighbourhood of the pinch point in a pinched torus can be described as the open cone on the disjoint union of two circles, and therefore the link L is given by $S^1 \sqcup S^1$ in this case, whose homology is different from homology of the link of any other point in the pinched torus (which is just given by a circle for every other point). The insets in Fig. 1 (left) depict the different neighbourhoods.

Local homology. We now formalise the idea of tracking the homology of a link of a point, following the previous description. Intuitively, the *local homology* of a point is obtained by the homology of an infinitesimal small punctured neighbourhood around the point (see Appendix A.3 for a more rigorous description). One can show that the local homology of a singularity usually differs from the local homology of a regular point. In particular, points that are of different intrinsic dimensions with respect to the stratified simplicial complex can be distinguished by local homology; this includes the disjoint union of manifolds of possibly varying dimensions. This observation motivates and justifies using local homology for detecting neighbourhoods of singular points, and serves as the primary motivation for our novel *Euclidicity* measure in Section 4.2.

[1] For readers familiar with persistent homology, we depart from the usual convention of using ϵ as the threshold parameter since we use it for the scale of our persistent local homology calculations.

3. Related Work

While manifold learning is concerned with the development of algorithms that extract geometric information under the assumption that the given data lie on a manifold, recent work starts to question this assumption. Brown et al. (2023), for instance, introduce the *union of manifolds hypothesis*, which augments the manifold hypothesis to spaces that can be modelled as (disjoint) unions of manifolds. Intrinsic dimension is thus allowed to vary across connected components of such a space. However, singularities are excluded under this assumption, whereas our method detects the correct intrinsic dimension for large classes of singular spaces. We assume a fundamental 'singularity-centric' perspective in this paper and argue that a multi-scale analysis of the local geometry and topology of data is necessary. In this context, methods from topological data analysis have started attracting attention in machine learning (Hensel et al., 2021). This is particularly due to *persistent homology*, which captures geometrical and topological properties of the underlying data set on different scales (Bubenik et al., 2020; Turkeš et al., 2022). The idea of tracking objects on multiple scales can at least be traced back to (Koenderink & van Doorn, 1986; Lindeberg, 1994), with scale space theory playing an eminent role in computer vision. However, the utility of persistent homology in the context of geometric singularities in data only came up more recently, since early work in persistent homology focuses predominantly on the simplification of functions on manifold domains (Edelsbrunner et al., 2002). While some research already discusses the utility of persistent homology for general unsupervised data analysis workflows (Chazal et al., 2013; Rieck & Leitte, 2017; 2016; 2015), it focuses more on global structures, whereas singularities are inherently local. We give a brief overview of methods in the emerging field of topology-driven singularity detection, outlining the differences to our approach below.

Topology-driven singularity detection. Several works assume a local perspective on homology to detect information about the intrinsic dimensionality of the data or the presence of certain singularities. Rieck et al. (2020) use pre-defined stratifications and *persistent intersection homology*, a technique developed by Bendich & Harer (2011), whereas Fasy & Wang (2016) and Bendich (2008) both develop persistent variants of local homology. By contrast, Stolz et al. (2020) approximate local homology as the absolute homology of a small annulus of a given neighbourhood, resulting in an algorithm for geometric anomaly detection (which requires knowing the intrinsic dimension of the data set). Moreover, Bendich et al. (2007) employ persistence vineyards, i.e. continuous families of persistence diagrams, to assess the local homology of a point in a stratified space, whereas Dey et al. (2014) use local homology to estimate the (global) intrinsic dimension of hidden, possibly noisy manifolds.

Key differences to existing approaches. While existing methods overall underscore the relevance of a local perspective, as well as the use of notions such as *stratified spaces*, our approach differs from them in essential components. In comparison to all aforementioned contributions, we capture additional local geometric information: we consider multiple scales of locality in a persistent framework for local homology. Concerning the overall construction, Stolz et al. (2020) is the closest to our method. However, the authors assume that the intrinsic dimension is known and the proposed algorithm uses one global scale, whereas our approach (i) operates in a multi-scale setting, (ii) provides local estimates of intrinsic dimensionality of the data space, and (iii) incorporates model spaces that serve as a comparison. We can thus measure the deviation from an idealised manifold, requiring fewer assumptions on the structure of the input data (see Appendix A.7 for a comparison).

4. Methods

Our framework TARDIS (Topological Algorithm for Robust DIscovery of Singularities) consists of two parts: (i) a method to calculate a local intrinsic dimension of the data, and (ii) *Euclidicity*, a measure for assessing the multi-scale deviation from a Euclidean space. TARDIS is based on the assumption that the intrinsic dimension of data may not be constant across the data set, and is thus best described by *local measurements*, i.e. measurements in a small neighbourhood of a given point. Since there is no canonical choice for the magnitude of such a neighbourhood, TARDIS analyses data on multiple scales. Our main idea involves constructing a collection of local (punctured) neighbourhoods for varying locality scales, and calculating their topological features. This procedure allows us to approximate local topological features (specifically, local homology) of a given point, which we use to measure the intrinsic dimensionality of a space. Moreover, by calculating the distance to Euclidean model spaces, we are capable of detecting singularities in a large range of input data sets. For the subsequent description of TARDIS, we only assume that data can be represented as a finite metric space (i.e. as a point cloud).

4.1. Persistent Intrinsic Dimension

For a finite metric space $(\mathbb{X}, \mathrm{d})$ and $x \in \mathbb{X}$, let $B_r^s(x) := \{y \in \mathbb{X} \mid r \leq \mathrm{d}(x,y) \leq s\}$ denote the *intrinsic annulus* of x in $\mathbb{X}$ with respect to radii r and s. Moreover, let $\mathcal{F}$ denote a procedure that takes as input a finite metric space and outputs an ascending filtration of topological spaces—such as a Vietoris–Rips filtration. By applying $\mathcal{F}$ to the intrinsic annulus of x, we obtain a tri-filtration $(\mathcal{F}(B_r^s(x), t))_{r,s,t}$, where t corresponds to the respective filtration step that is determined by $\mathcal{F}$. Note that this tri-filtration is covariant in s and t, but contravariant in r; we

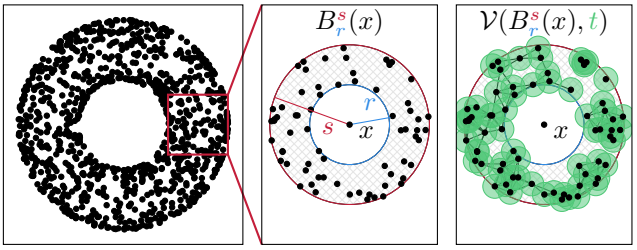


Figure 3: The intrinsic annulus $B_r^s(x)$ around a point x in a metric space $(\mathbb{X}, \mathrm{d})$, as well as one filtration step for some choice of t. By adjusting r and s, we obtain a tri-filtration.

denote it by $\mathcal{F}(B_\bullet^\bullet(x), \bullet)$. Applying ith homology to this filtration yields a tri-parameter persistent module that we call ith **persistent local homology (PLH)** of x, denoted by $\mathrm{PLH}_i(x; \mathcal{F}) := \mathrm{PH}_i(\mathcal{F}(B_\bullet^\bullet(x), \bullet))$. Fig. 3 illustrates how to obtain an annulus from a data set and depicts one step of the filtration process. To the best of our knowledge, this is the first time that PLH is considered as a multi-parameter persistence module. Since the Vietoris–Rips filtration is the pre-eminent filtration in TDA, we will use $\mathrm{PLH}_i(x) := \mathrm{PLH}_i(x; \mathcal{V})$ as an abbreviation. Before developing ways to detect singularities, we first show that our PLH formulation enjoys stability properties similar to the seminal stability theorem in persistent homology (Cohen-Steiner et al., 2007), making it robust to small parameter changes.

Theorem 1. *Given a finite metric space $\mathbb{X}$ and $x \in \mathbb{X}$, let $B_r^s(x)$ and $B_{r'}^{s'}(x)$ denote two intrinsic annuli with $|r - r'| \leq \epsilon_1$ and $|s - s'| \leq \epsilon_2$. Furthermore, let $\mathcal{D}, \mathcal{D}'$ denote the persistence diagrams corresponding to $\mathrm{PH}_i(\mathcal{V}(B_r^s(x), \bullet))$ and $\mathrm{PH}_i(\mathcal{V}(B_{r'}^{s'}(x), \bullet))$. Then $\frac{1}{2}\,\mathrm{d_B}(\mathcal{D}, \mathcal{D}') \leq \max\{\epsilon_1, \epsilon_2\}$.*

For a finite set of points $\mathbb{X} \subset \mathbb{R}^N$, we define the **persistent intrinsic dimension (PID)** of $x \in \mathbb{X}$ at scale ϵ as $i_x(\epsilon) := \max\{i \in \mathbb{N} \mid \exists\, r < s < \epsilon \text{ s.t. } \mathrm{PH}_{i-1}(\mathcal{F}(B_r^s(x), \bullet)) \neq 0\}$. This measure characterises the intrinsic dimension of data in a multi-scale fashion. We can also prove that we recover the correct dimension in case our data set constitutes a manifold sample.

Theorem 2. *Let $M \subset \mathbb{R}^N$ be an n-dimensional compact smooth manifold and let $\mathbb{X} := \{x_1, \ldots, x_S\}$ be a collection of uniform samples from M. For a sufficiently large S and $\mathcal{F} = \mathcal{V}$, there exist constants $\epsilon_1, \epsilon_2 > 0$ such that $i_x(\epsilon) = n$ for all $\epsilon_1 < \epsilon < \epsilon_2$ and any point $x \in \mathbb{X}$. Moreover, ϵ_1 can be chosen arbitrarily small by increasing S.*

Theorem 2 implies that $i_x(\epsilon)$ computes the correct intrinsic dimension of M in a certain range of values $\epsilon > 0$, provided the sample size is sufficiently large. Moreover, $i_x(\epsilon)$ persists in this range, which suggests considering a collection of $i_x(\epsilon)$ for varying ϵ to analyse the intrinsic dimension of x. We also have the following corollary, which specifically addresses stratified spaces such as the 'pinched torus,' implying that our method can correctly detect the intrinsic dimension of individual strata. PID is thus capable of handling large classes of 'non-manifold' data sets.

Corollary 1. *Let $X = X_n \supset X_{n-1} \supset X_{n-2} \supset \cdots \supset X_{-1} = \emptyset$ be an n-dimensional compact stratified simplicial complex, s.t. $X_i \setminus X_{i-1}$ is smooth for every i. For a fixed i, let $\mathbb{X}_i := \{x_1, \ldots, x_S\}$ be a collection of uniform samples from $X_i \setminus X_{i-1}$. For a sufficiently large S and $\mathcal{F} = \mathcal{V}$, there are constants $\epsilon_1, \epsilon_2 > 0$ such that $i_x(\epsilon) = i$ for all $\epsilon_1 < \epsilon < \epsilon_2$ and any point $x \in \mathbb{X}_i$. Moreover, ϵ_1 can be chosen arbitrarily small by increasing S.*

4.2. Euclidicity

Knowledge about the intrinsic dimension of a neighbourhood is crucial for measuring to what extent such a neighbourhood deviates from being Euclidean. We refer to this deviation as *Euclidicity*, with the understanding that low values indicate Euclidean neighbourhoods while high values indicate singular regions of a data set. *Euclidicity* can be calculated without stringent assumptions on manifoldness, requiring only an estimate of the intrinsic dimension n of x. The previously-described PID estimation procedure is applicable in this setting and may be used to obtain n, for example by calculating statistics on the set of $i_x(\epsilon)$ for varying locality parameters ϵ. *Euclidicity* can also use other dimension estimation procedures, which is advantageous when additional knowledge about the expected structures is available (see Camastra & Staiano (2016) for a survey).

The main idea of *Euclidicity* involves assessing how far a given neighbourhood of a point x is from being Euclidean.

To this end, we compare it to a Euclidean model space, measuring the deviation of their corresponding persistent local homology features. We first define the Euclidean annulus $\mathbb{E}\mathrm{B}_r^s(x)$ of x for parameters r and s to be a set of random uniform samples of $\{y \in \mathbb{R}^n \mid r \leq \mathrm{d}(x, y) \leq s\}$ such that $|\,\mathbb{E}\mathrm{B}_r^s(x)| = |B_r^s(x)|$. Here, r and s correspond to the inner and outer radius of the annulus, respectively. For $r' \leq r$ and $s \leq s'$ we extend $\mathbb{E}\mathrm{B}_r^s(x)$ by sampling additional points to obtain $\mathbb{E}\mathrm{B}_{r'}^{s'}(x)$ with $|\,\mathbb{E}\mathrm{B}_{r'}^{s'}(x)| = |B_{r'}^{s'}(x)|$. Iterating this procedure leads to a tri-filtration $(\mathcal{F}(\mathbb{E}\mathrm{B}_r^s(x), t))_{r,s,t}$ for any filtration $\mathcal{F}$, following our description in Section 4.1. We now define the persistent local homology of a Euclidean model space as

$$\mathrm{PLH}_i^{\mathbb{E}}(x; \mathcal{F}) := \mathrm{PH}_i(\mathcal{F}(\mathbb{E}\mathrm{B}_\bullet^\bullet(x), \bullet)). \tag{1}$$

Again, for a Vietoris–Rips filtration $\mathcal{V}$, we use a short-form notation, i.e. $\mathrm{PLH}_i^{\mathbb{E}}(x) := \mathrm{PLH}_i^{\mathbb{E}}(x; \mathcal{V})$. Notice that $\mathrm{PLH}_i^{\mathbb{E}}(x)$ implicitly depends on the choice of intrinsic dimension n, and on the samples that are generated randomly. To remove the dependency on the samples, we consider $\mathrm{PLH}_i^{\mathbb{E}}(x)$ to be a sample of a random variable $\mathbf{PLH}_i^{\mathbb{E}}(x)$.

Let $\mathrm{D}(\cdot,\cdot)$ be a distance measure for 3-parameter persistence modules, such as the *interleaving distance* (Lesnick, 2015). We then define the **Euclidicity** of x, denoted by $\mathfrak{E}(x)$, as the expected value of these distances, i.e.

$$\mathfrak{E}(x) := \mathrm{E}\Big[\mathrm{D}\Big(\mathrm{PLH}_{n-1}(x), \mathbf{PLH}^{\mathbb{E}}_{n-1}(x)\Big)\Big]. \quad (2)$$

This quantity essentially assesses how far x is from admitting a regular Euclidean neighbourhood. The choice of $\mathcal{F}$ and $\mathrm{D}(\cdot,\cdot)$ leaves us multiple ways of implementing Eq. (2) in practice. In the following, we describe one particular implementation with beneficial robustness properties.

Implementation. Calculating $\mathfrak{E}(x)$ requires different choices, namely (i) a range of locality scales, (ii) a filtration, and (iii) a distance metric between filtrations D. Using a grid Γ of possible radii (r,s) with $r<s$, we approximate Eq. (2) using the *mean of the bottleneck distances of fibred Vietoris–Rips barcodes*, i.e.

$$\mathfrak{E}(x) \approx \mathrm{D}\big(\mathrm{PLH}_i(x), \mathrm{PLH}^{\mathbb{E}}_i(x)\big) := \frac{1}{C}\sum_{(r,s)\in\Gamma} {\mathrm{d_B}}^{r,s} \quad (3)$$

where C is equal to the cardinality of the grid, and ${\mathrm{d_B}}^{r,s} := \mathrm{d_B}(\mathrm{PH}_i(\mathcal{V}(B^s_r(x),\bullet)), \mathrm{PH}_i(\mathcal{V}(\mathbb{E}\mathrm{B}^s_r(x),\bullet)))$, with $\mathrm{PLH}^{\mathbb{E}}_i(x)$ referring to a sample from a Euclidean annulus of the same size as the intrinsic annulus around x. Eq. (3) can be implemented using effective persistent homology calculation methods (Bauer, 2021), thus permitting an integration into existing TDA and machine learning frameworks (The GUDHI Project, 2015; Tauzin et al., 2020). Appendix A.4 provides pseudocode implementations, while Section 5 discusses how to pick these parameters in practice. We make our framework publicly available.[2]

As the following theorem shows, our approximation of Eq. (3) is justified in the sense that for smooth manifolds, $\mathfrak{E}(x)$ tends to be arbitrarily small in a large-sample regime.

Theorem 3. *Let $M \subset \mathbb{R}^N$ be a smooth n-dimensional manifold and let $\mathbb{X} \subset M$ be a finite sample of size $S := |\mathbb{X}|$. For a given $\epsilon > 0$, sufficiently large S and a point $x \in \mathbb{X}$, there exists $s_\epsilon > 0$ that only depends on ϵ and the curvature of x (w.r.t. M), such that the approximation of $\mathfrak{E}(x)$ via Eq. (3) is bounded above by ϵ, for any grid Γ with maximum outer radius s_ϵ.*

Properties. The main appeal of our formulation is that calculating both PID and Euclidicity does not require strong assumptions about the input data: we only assume that the intrinsic dimension n of the data is significantly lower than the ambient dimension N. Treating dimension as a local quantity that may vary across multiple scales makes Euclidicity broadly applicable. Moreover, as we showed in Section 4.1, our method is *guaranteed* to yield the right values for manifolds and stratified simplicial complexes. This increases both the practical applicability and expressivity, enabling our framework to handle unions of manifolds of varying dimensions, for instance. Euclidicity thus generalises to a larger class of spaces than existing approaches (Brown et al., 2023), permitting a more fine-grained structural assessment.

Limitations. Our implementation of Euclidicity makes use of the Vietoris–Rips complex, which is known to grow exponentially with increasing dimensionality. While all calculations of Eq. (2) can be performed *in parallel*—thus substantially improving scalability vis-à-vis persistent homology on the complete input data set, both in terms of dimensions and in terms of samples—the memory requirements for a full Vietoris–Rips complex construction may still prevent our method to be applicable for some high-dimensional data sets. This can be mitigated by using a different filtration (Anai et al., 2020; Sheehy, 2013). Our proofs do *not* assume a specific filtration, and we leave the derivation of filtration-specific theoretical properties for future work. Finally, we remark that the reliability of the Euclidicity score depends on the validity of the intrinsic dimension; otherwise, the comparison does not take place with respect to the appropriate model space.

5. Experiments

We demonstrate the expressivity of TARDIS in different settings, showing that it (i) calculates the correct intrinsic dimension, and (ii) detects singularities when analysing data sets with known singularities. We also conduct a brief comparison with one-parameter approaches, showcasing how our multi-scale approach results in more stable outcomes. Finally, we analyse Euclidicity scores of benchmark and real-world datasets, giving evidence that our technique can be used as a measure for the geometric complexity of data.

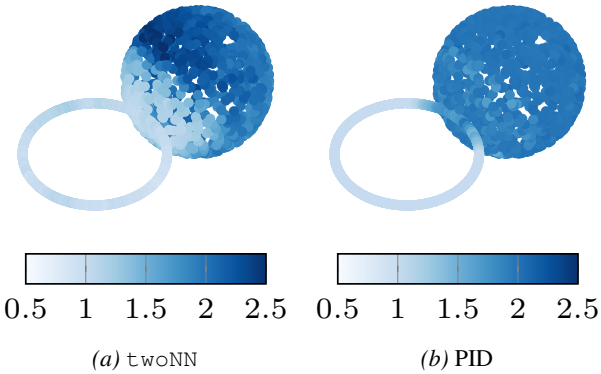
5.1. Parameter Selection

Since Eq. (2) intrinsically incorporates multiple scales of locality, we need to specify an upper bound for the radii $(r_{\min}, r_{\max}, s_{\min}, s_{\max})$ that define the respective annuli in practice. Given a point x, we found the following procedure to be useful in practice: we set $s_{\max}$, i.e. the maximum of the outer radius, to the distance to the kth nearest neighbour of a point, and $r_{\min}$, i.e the minimum inner radius, to the smallest non-zero distance to a neighbour of x. Finally, we set the minimum outer radius $s_{\min}$ and the maximum inner radius $r_{\max}$ to the distance to the $\lfloor \frac{k}{3} \rfloor$th nearest neighbour. While we find $k = 50$ to yield sufficient results, spaces with a high intrinsic dimension may require larger values. The advantage of using such a parameter selection procedure is that it works in a data-driven manner,

[2] See the supplementary materials for the code and experiments.

Table 1: Dimensionality estimates for the concatenation of S^1 and S^2, denoted by $S^1 \vee S^2$. Our PID measure is more capable of detecting the changes in dimensionality that arise from the concatenation.

	METHOD	MIN	$\mu \pm \sigma$	MAX
1D	lpca	1.00	1.42±0.78	3.00
	twoNN	0.83	1.00±0.07	1.20
	DANCo	1.00	1.00±0.01	1.16
	PID	1.00	1.12±0.24	1.97
2D	lpca	2.00	2.88±0.32	3.00
	twoNN	1.01	1.90±0.36	2.53
	DANCo	1.00	2.10±0.32	3.00
	PID	1.52	1.95±0.06	2.08



(a) twoNN *(b)* PID

Figure 4: Dimensionality estimates. PID is more nuanced in capturing changes in dimensionality, assigning ≈ 1 to almost all points of the circle, i.e. S^1, while highlighting that points closer to S^2 exhibit an increase in dimensionality.

accounting for differences in density. Since our approach is inherently *local*, we need to find a balance between sample sizes that are sufficiently large to contain topological information, while at the same time being sufficiently small to retain a local perspective. We found the given range to be an appropriate choice in practice. As for the number of steps, we discretise the parameter range using 20 steps by default. Higher numbers are advisable when there are large discrepancies between the radii, for instance when $s_{\max} \gg r_{\max}$.

5.2. Persistent Intrinsic Dimension is Expressive

We first analyse the behaviour of persistent intrinsic dimension (PID) on samples from a space obtained by concatenating S^1 (a circle) and S^2 (a sphere) at a gluing point. Table 1 shows a comparison of PID with state-of-the-art dimensionality estimators.[3] We find that PID outperforms all estimators in terms of mean and standard deviation for the 2D points, thus correctly indicating that the majority of all points admit non-singular 2D neighbourhoods. For the 1D points, the mean of the dimensionality estimate of PID is still close to the ground truth, while its standard deviation and maximum correctly capture the fact that some 1D points are situated closer to the gluing point. Fig. 4 exemplifies this behaviour and shows a comparison between one of the estimators and our PID measure. PID is more nuanced in capturing changes in local intrinsic dimension as one approaches the gluing point, correctly showing that points of the sphere admit 2D neighbourhoods. This behaviour is in line with our philosophy of considering dimensionality as an inherently local phenomenon. In case such behaviour is not desirable for a specific data set, Euclidicity calculations support *any* dimensionality estimator; since such estimators do not come with strong guarantees such as Theorem 2, their choice must be ultimately driven by the data set at hand. See Appendix A.6 for a more detailed analysis of these estimates.

In practice, the sample density may not be sufficiently high for Theorem 2 to apply. This means that there may appear artefact homological features in dimensions *higher* than the intrinsic dimension of a given space. We thus recommend to only consider features that exceed a certain persistence threshold in comparison to the persistence of features of lower dimension: for any data point x and the respective intrinsic annulus $B_r^s(x)$, we suggest to eliminate all topological features whose lifetimes are smaller than the maximum lifetime of features in one dimension below. This results in markedly stable estimates of intrinsic dimension, which are less prone to overestimations.

5.3. Euclidicity Captures Known Singularities

Before applying Euclidicity to high-dimensional spaces with unknown characteristics, we first analyse its behaviour on spaces with known singularities. Fig. 1 shows that Euclidicity is capable of detecting the singularity of the 'pinched torus.' Of particular relevance is the fact that Euclidicity also highlights that points in the vicinity of the singular point are *not* fully regular. This is an important property for practical applications since it implies that Euclidicity can detect such *isolated singularities* even in the presence of sampling errors.

Another prototypical example of singular spaces is given by $S^n \vee S^n$, the *wedge* of two n-dimensional spheres, which is obtained by gluing them together at one point. Denoting the gluing point by x_0, for a suitable triangulation of $X = S^n \vee S^n$, this space can be stratified by $X \supset \{x_0\}$. To assess the utility of TARDIS, we apply it to samples of

[3] Method names are taken from the scikit-dimension toolkit. See Appendix A.6 for more details.

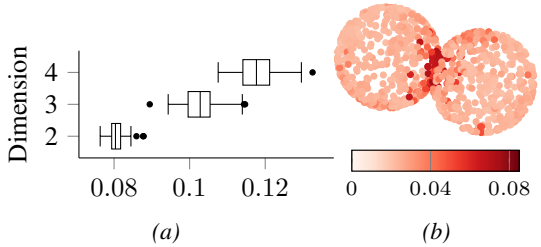


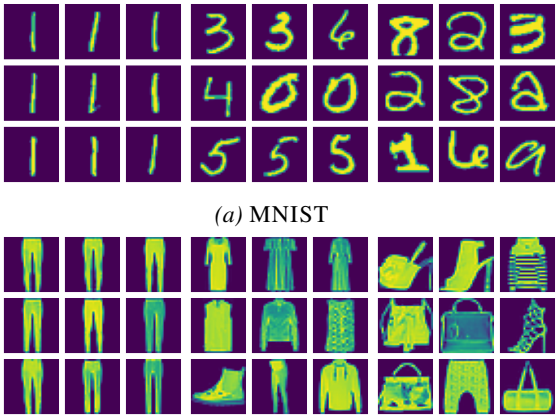
Figure 5: (a): Euclidicity scores of wedged spheres for different dimensions. High values indicate singular points/neighbourhoods. The Euclidicity of the singular point always constitutes a clear positive outlier. In 2D, *Euclidicity* (b) results in a clearly-delineated singular region.

such wedged spheres of dimensions $\{2, 3, 4\}$, calculating their respective Euclidicity scores. Since larger intrinsic dimensions require higher sample sizes to maintain the same density, we start with a sample size of 20000 in dimension 2 and increase it consecutively by a factor of 10. We then calculate Euclidicity of 50 random points in the respective data set, and additionally for the singular point x_0.

Fig. 5a shows the results of our experiments. We observe that the singular point exhibits a *significantly higher Euclidicity* score than the random samples. Moreover, we find that Euclidicity scores of non-singular points exhibit a high degree of variance across the data, which is caused by the fact that the sampled data does not perfectly fit the underlying space the points are being sampled from. This strengthens our main argument: assessing whether a specific point is Euclidean does not require a binary decision but a continuous measure such as Euclidicity.

Stability. Theorem 1 predicts that Euclidicity estimates are stable. Given the use of mini-batches in machine learning, we first assess whether Euclidicity is *robust towards sampling*: repeating the calculations for the 'pinched torus' over different batches results in highly similar distributions that are not distinguishable according to Tukey's range test (Tukey, 1949) at the $\alpha = 0.05$ confidence level. Moreover, choosing larger locality scales still enables us to detect singularities at higher computational costs and incorporating larger parts of the point cloud. Please refer to Appendix A.5 for a more detailed discussion of this aspect.

Comparison to single-parameter approach. We find that our Euclidicity measure also leads to significantly more stable results than a comparable single-parameter approach for geometry-based anomaly detection (Stolz et al., 2020). We analyse this behaviour in more detail in Appendix A.7, observing that a single-parameter approach, which examines data with a constant global scale, results in many 'false positives,' i.e. points with high anomaly scores that in fact



(a) MNIST

(b) FASHIONMNIST

Figure 6: Euclidicity captures local sample complexity in an unsupervised manner. From left to right: sample images exhibiting low, median, and high Euclidicity, respectively.

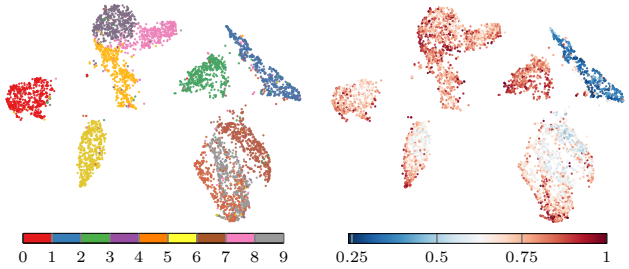
do admit a Euclidean neighbourhood. This behaviour will only be exacerbated in higher dimensions and sparser data, highlighting that the use of multiple locality scales by Euclidicity is advantageous.

The preceding experiments verified PID and Euclidicity in terms of *expressivity*, i.e. detecting singularities in spaces with a known ground truth, and *robustness*. Hence, our subsequent experiments will proceed with the analysis of high-dimensional spaces, exemplifying potential use cases of Euclidicity.

5.4. Euclidicity of High-Dimensional Spaces

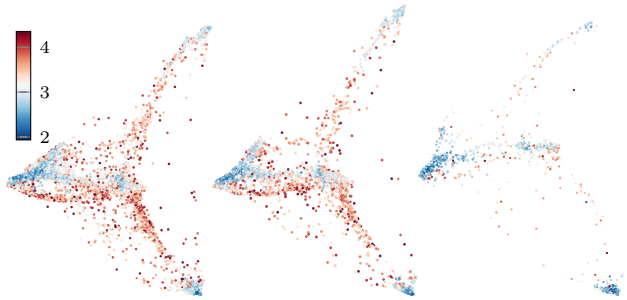
To test TARDIS in an unsupervised setting, we calculate Euclidicity scores for the MNIST and FASHIONMNIST data sets, selecting mini-batches of 1000 samples from a subsample of 10000 random images of these data sets. We 'flatten' all greyscale images, thus representing each image as a high-dimensional point, and a collection of images as a point cloud, for which we can calculate Euclidicity values. This procedure follows common practice in dimensionality reduction (McInnes et al., 2018; Moor et al., 2020). Following Pope et al. (2021), we assume an intrinsic dimension of 10; moreover, we use $k = 50$ neighbours for local scale estimation. To ensure that our results are representative, we repeat all calculations for five different subsamples. Euclidicity scores range from $[1.1, 5.3]$ for MNIST, and $[1.3, 5.6]$ for FASHIONMNIST. The scores of the two datasets appear to be following different distributions (see Appendix A.8 for a visualisation and a detailed depiction of the distributions).

Fig. 6 shows a selection of 9 images, corresponding to the lowest, median, and highest Euclidicity scores, respectively. We observe that high Euclidicity scores correspond



(a) Labels *(b)* Euclidicity scores

Figure 7: A UMAP (McInnes et al., 2018) embedding of the MNIST data, showing Euclidicity scores and class labels.



(a) 10,000 points *(b)* 5,000 points *(c)* 1,000 points

Figure 8: PHATE embeddings of a cytometry data set, with colours based on Euclidicity. Euclidicity highlights dense non-singular regions and remains stable under subsampling the iPSC data set. Minor variations in the point cloud shape are due to the PHATE embedding algorithm; Euclidicity was always calculated on the raw data.

to images with a high degree of non-linearity, whereas low Euclidicity scores correspond to images that exhibit less complex structures: for MNIST, these are digits of '1.' Interestingly, we observe the same phenomenon for FASHIONMNIST, where images with low Euclidicity ('pants') possess less geometric complexity in contrast to images with high Euclidicity. Since low Euclidicity can also be seen as an indicator of how close a neighbourhood is to being *locally linear*, this finding hints at the existence of simple substructures in the data, prompting the use of Euclidicity as an unsupervised measure of geometric complexity.

To analyse the relationship between geometric complexity and Euclidicity, we focus on MNIST. Fig. 7 shows an embedding of the data, with labels and *normalised Euclidicity scores* (by the maximum) highlighted (Euclidicity was only calculated on the raw data; the embedding is just used for visualisation purposes). We find that low Euclidicity scores are prevalent in clusters of '1's, whereas '5's are assigned both lower than average and higher than average Euclidicity scores (similar patterns hold for other classes). This lends credence to considering MNIST to consist of a *union of manifolds*, which are, however, not necessarily split along the different classes of digits (meaning that, as our scores indicate, even images having the same label may not 'live' on the same manifold).

To further assess this hypothesis, we analyse the connection between a classifier's ability to correctly classify a given sample from the test set, and its corresponding normalised Euclidicity score. To this end, we trained a simple neural network classifier for both MNIST and FASHIONMNIST, and subsequently compared Euclidicity scores of correctly and incorrectly classified samples. We note that the misclassified samples admit Euclidicity scores that are significantly higher than a random equally-sized sample of the correctly classified images. For MNIST, the average Euclidicity of correctly classified samples is 0.33, whereas misclassified samples admit an average score of 0.39. The respective averages for FASHIONMNIST are 0.31 and 0.35, respectively. Welch's t-test indicates significance of these results; please refer to Appendix A.8 for details on the model architecture and statistical significance. This insight encourages future uses of Euclidicity in models to improve or explain classification issues, either as an unsupervised preprocessing step, or by incorporating Euclidicity into the model architecture itself via appropriate loss terms.

5.5. Euclidicity of Cytometry Data

To highlight the utility of Euclidicity in unsupervised representation learning, we also calculate it on an induced pluripotent stem cell (iPSC) reprogramming data set (Zunder et al., 2015). The data set consists of 33 variables and around 220,000 samples; it depicts a progression of so-called *fibroblasts* diverging, and splitting into two different lineages. The data set is known to contain branching structures (Bhaskar et al., 2022) that can best be extracted using PHATE (Moon et al., 2019), a non-linear dimensionality reduction algorithm. We only employ this algorithm for visualisation purposes; all Euclidicity calculations are performed on the original data. We use `twoNN` to estimate dimensionality, obtaining a mean intrinsic dimension of 16, and select parameters as described in Section 5.4.

Fig. 8a shows an embedding obtained via PHATE and the Euclidicity scores of the original data. We find that high Euclidicity scores occur in regions that exhibit a lower density in the PHATE embedding; such points turn out to be situated in lower-dimensional subspaces.[4] We verify this fact using the `twoNN` dimensionality estimates depicted in Fig. 9. More specifically, we calculated the intrinsic dimension for the subsample, observing that the interquartile

[4] However, low-density regions in the PHATE visualisation need not necessarily correspond to low-density regions in the original dataset.

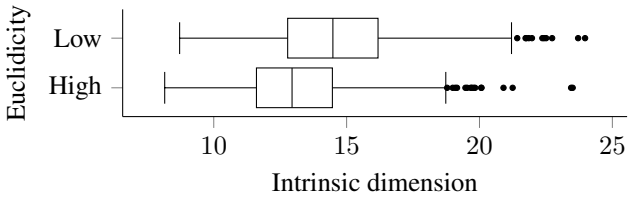


Figure 9: A comparison of intrinsic dimension estimates computed for points in the iPSC dataset that admit low (top) and high (bottom) Euclidicity scores.

range for the 1,000 points with *highest Euclidicity* is around 12–14, whereas the interquartile range of the 1,000 *lowest Euclidicity* points ranges between around 13–16. Again, we used the `twoNN` algorithm for intrinsic dimensionality estimates (using $k = 50$ nearest neighbours). Since lower-dimensional points in a space can be considered *singular* in the sense of stratified spaces, this is further evidence for Euclidicity to be a useful tool for detecting non-manifold regions in data. Finally, we remark that our Euclidicity estimates remain stable under *subsampling*. Fig. 8 depicts subsamples of different sizes for which we calculated Euclidicity (on the raw data, respectively, using PHATE to obtain embeddings). Euclidicity distributions remain stable and the same phenomena are highlighted for each subsample.

This analysis constitutes a proof of concept, showcasing how to use Euclidicity to 'probe' data sets and enrich general unsupervised data analysis workflows.

6. Discussion

We present TARDIS, a framework for locally estimating (i) the intrinsic dimension via PID, the *persistent intrinsic dimension*, and (ii) the 'manifoldness' via *Euclidicity*, a multi-scale measure of the deviation from Euclidean space of point clouds. Our method is based on a novel formulation of persistent local homology as a multi-parameter approach, and we provide theoretical guarantees for it in a dense sample setting. Our experiments showed significant improvements of stability compared to geometry-based anomaly detection methods with fixed locality scales, and we found that Euclidicity can detect singular regions in data sets with known singularities. Moreover, using high-dimensional benchmark and real-world data sets, we also find that Euclidicity can serve as an unsupervised measure of geometric complexity.

Future work. For future work, we envision two relevant research directions. First and foremost will be the inclusion of Euclidicity into machine learning models to make them 'singularity-aware.' In light of our experiments in Section 5.4 and Section 5.5, we believe that Euclidicity will be particularly useful in unsupervised scenarios, or provide an additional weight in classification settings (to ensure that singular examples are being given lower confidence scores). Moreover, Euclidicity could be used in the detection of adversarial samples—a task for which knowledge about the underlying topology of a space is known to be crucial (Jang et al., 2020). As a second direction, we want to further improve the properties of Euclidicity itself. To this end, we plan to investigate if incorporating custom distance measures for three-parameter persistence modules, i.e. different metrics for Eq. (3), lead to improved results in terms of stability, expressivity, and computational efficiency. Moreover, we hypothesise that replacing the Vietoris–Rips filtration will prove beneficial in reducing the number of samples for calculating Euclidicity (Anai et al., 2020; Sheehy, 2013). Along these lines, we also plan to derive theoretical results that relate specific filtrations and the expressivity of the corresponding Euclidicity measure. Another direction for future research concerns the approximation of a manifold from inherently singular data, i.e. finding the *best* manifold approximation to a given data set with singularities. This way, singularities could be resolved during the training phase of models, provided an appropriate loss function exists. Euclidicity may thus serve as a metric for assessing data sets, paving the way towards more trustworthy and faithful embeddings.

Acknowledgements

B.R. is supported by the Bavarian state government with funds from the *Hightech Agenda Bavaria*. The authors sincerely thank Francesco Conti for stimulating discussions. Moreover, the authors are grateful for the stimulating discussions with the anonymous reviewers. In particular the comments by reviewer `By5N` helped us in refining and positioning our method better in the context of other unsupervised machine learning methods.

Reproducibility

Our code is available under `https://github.com/aidos-lab/TARDIS`. All dependencies are listed in the respective `pyproject.toml` file, and the `README.md` discusses how to install our package and run our experiments.

References

Anai, H., Chazal, F., Glisse, M., Ike, Y., Inakoshi, H., Tinarrage, R., and Umeda, Y. DTM-based filtrations. In Baas, N. A., Carlsson, G. E., Quick, G., Szymik, M., and Thaule, M. (eds.), *Topological Data Analysis*, pp. 33–66, 2020.

Bauer, U. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

Bendich, P. *Analyzing stratified spaces using persistent versions of intersection and local homology*. PhD thesis, Duke University, 2008.

Bendich, P. and Harer, J. Persistent intersection homology. *Foundations of Computational Mathematics*, 11(3):305–336, 2011.

Bendich, P., Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Morozov, D. Inferring local homology from sampled stratified spaces. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 536–546, 2007.

Bhaskar, D., MacDonald, K., Fasina, O., Thomas, D. S., Rieck, B., Adelstein, I., and Krishnaswamy, S. Diffusion curvature for estimating local curvature in high dimensional data. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=AYII8AkvD1e.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. Verifying the union of manifolds hypothesis for image data. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Rvee9CAX4fi.

Bubenik, P., Hull, M., Patel, D., and Whittle, B. Persistent homology detects curvature. *Inverse Problems*, 36(2):025008, 2020.

Camastra, F. and Staiano, A. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, 2016.

Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

Chazal, F., Guibas, L. J., Oudot, S. Y., and Skraba, P. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 60(6), 2013.

Chazal, F., De Silva, V., and Oudot, S. Persistence stability for geometric complexes. *Geometriæ Dedicata*, 173(1):193–214, 2014.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

Dey, T. K., Fan, F., and Wang, Y. Dimension detection with local homology. In *Canadian Conference on Computational Geometry (CCCG)*, 2014.

Edelsbrunner, H. and Harer, J. *Computational topology: An introduction*. American Mathematical Society, Providence, RI, USA, 2010.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. J. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.

Fasy, B. T. and Wang, B. Exploring persistent local homology in topological data analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6430–6434, 2016.

Hensel, F., Moor, M., and Rieck, B. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4, 2021.

Jakubowski, A., Gašić, M., and Zibrowius, M. Topology of word embeddings: Singularities reflect polysemy. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pp. 103–113, 2020.

Jang, U., Jha, S., and Jha, S. On the need for topology-aware generative models for manifold-based defenses. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1lF_CEYwS.

Koenderink, J. J. and van Doorn, A. J. Dynamic shape. *Biological Cybernetics*, 53(6):383–396, 1986.

Lesnick, M. The theory of the interleaving distance on multidimensional persistence modules. *Foundations of Computational Mathematics*, 15(3):613–650, 2015.

Lindeberg, T. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of Applied Statistics*, 21(1-2):225–270, 1994.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. v. d., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., and Krishnaswamy, S. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019.

Moor, M., Horn, M., Borgwardt, K., and Rieck, B. Challenging Euclidean topological autoencoders. In *'Topological Data Analysis and Beyond' Workshop at NeurIPS*, 2020. URL https://openreview.net/forum?id=P3dZuOUnyEY.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact

on learning. In *International Conference on Learning Representations*, 2021.

Rieck, B. and Leitte, H. Persistent homology for the evaluation of dimensionality reduction schemes. *Computer Graphics Forum*, 34(3):431–440, 2015.

Rieck, B. and Leitte, H. Exploring and comparing clusterings of multivariate data sets using persistent homology. *Computer Graphics Forum*, 35(3):81–90, 2016.

Rieck, B. and Leitte, H. Agreement analysis of quality measures for dimensionality reduction. In Carr, H., Garth, C., and Weinkauf, T. (eds.), *Topological Methods in Data Analysis and Visualization IV*, pp. 103–117. 2017.

Rieck, B., Banagl, M., Sadlo, F., and Leitte, H. Persistent intersection homology for the analysis of discrete data. In Carr, H., Fujishiro, I., Sadlo, F., and Takahashi, S. (eds.), *Topological Methods in Data Analysis and Visualization V*, pp. 37–51. 2020.

Scoccola, L. and Perea, J. A. FibeRed: Fiberwise dimensionality reduction of topologically complex data with vector bundles. In Chambers, E. W. and Gudmundsson, J. (eds.), *International Symposium on Computational Geometry (SoCG)*, volume 258, pp. 56:1–56:18, 2023.

Sheehy, D. R. Linear-size approximations to the vietoris–rips filtration. *Discrete & Computational Geometry*, 49 (4):778–796, 2013.

Stolz, B. J., Tanner, J., Harrington, H. A., and Nanda, V. Geometric anomaly detection in data. *Proceedings of the National Academy of Sciences*, 117(33):19664–19669, 2020.

Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Medina-Mardones, A., Dassatti, A., and Hess, K. `giotto-tda`: A topological data analysis toolkit for machine learning and data exploration. *arXiv:2004.02551*, 2020.

The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 2015. URL http://gudhi.gforge.inria.fr/doc/latest/.

Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114, 1949.

Turkeš, R., Montúfar, G., and Otter, N. On the effectiveness of persistent homology. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=DRjUkfExCix.

Zunder, E., Lujan, E., Goltsev, Y., Wernig, M., and Nolan, G. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell Stem Cell*, 16(3):323–337, 2015.

A. Appendix

A.1. Notation

Symbol	Meaning
ϵ	local annulus scale parameter
$\mathbb{R}$	real numbers
H_i	ith (ordinary) homology functor (with $\mathbb{Z}/2\mathbb{Z}$ coefficients)
$\tilde{H}_i$	ith reduced homology functor (with $\mathbb{Z}/2\mathbb{Z}$ coefficients)
$\inf$	infimum
$\sup$	supremum
$\lvert\cdot\rvert_\infty$	uniform (infinity) norm
n	intrinsic dimension of the space under consideration
N	ambient dimension of the space under consideration
$\varinjlim$	(categorical) colimit
S^k	k-dimensional sphere
$c^\circ X := X \times (0,1]/X \times \{1\}$	open cone of a topological space X

A.2. Proofs of the Main Statements in the Paper

We restate the theorems from the main paper for the convenience of readers, along with their proofs, which were removed for space reasons. We first prove the stability theorem, first stated on p. 4 in the main text, which shows that our method enjoys stability properties with respect to radius changes of the intrinsic annuli.

Theorem 1. *Given a finite metric space $\mathbb{X}$ and $x \in \mathbb{X}$, let $B_r^s(x)$ and $B_{r'}^{s'}(x)$ denote two intrinsic annuli with $|r - r'| \leq \epsilon_1$ and $|s - s'| \leq \epsilon_2$. Furthermore, let $\mathcal{D}, \mathcal{D}'$ denote the persistence diagrams corresponding to $\mathrm{PH}_i(\mathcal{V}(B_r^s(x), \bullet))$ and $\mathrm{PH}_i(\mathcal{V}(B_{r'}^{s'}(x), \bullet))$. Then $\frac{1}{2}\,\mathrm{d_B}(\mathcal{D}, \mathcal{D}') \leq \max\{\epsilon_1, \epsilon_2\}$.*

Proof. The Hausdorff distance of two non-empty subsets $A, B \subset \mathbb{X}$ is $\mathrm{d_H}(A, B) := \inf\{\epsilon \geq 0 \mid A \subset B_\epsilon, B \subset A_\epsilon\}$, where $A_\epsilon = \cup_{a \in A}\{x \in \mathbb{X};\ \mathrm{d}(x, a) \leq \epsilon\}$ denotes the ϵ-thickening of A in X. Set $\epsilon := \max\{\epsilon_1, \epsilon_2\}$. By assumption, $B_r^s(x) \subset B_{r'}^{s'}(x)_\epsilon$ and $B_{r'}^{s'}(x) \subset B_r^s(x)_\epsilon$, i.e. $\mathrm{d_H}(B_r^s(x), B_{r'}^{s'}(x)) \leq \epsilon$. Using the geometric stability theorem of persistence diagrams (Chazal et al., 2014), we have $\frac{1}{2}\,\mathrm{d_B}(\mathcal{D}, \mathcal{D}') \leq \mathrm{d_H}(B_r^s(x), B_{r'}^{s'}(x))$, which proves the claim. □

Next, we prove that our *persistent intrinsic dimension* (PID) measure is capable of capturing the dimension of manifolds correctly, provided sufficiently many samples are present. This theorem was first stated on p. 4.

Theorem 2. *Let $M \subset \mathbb{R}^N$ be an n-dimensional compact smooth manifold and let $\mathbb{X} := \{x_1, \ldots, x_S\}$ be a collection of uniform samples from M. For a sufficiently large S and $\mathcal{F} = \mathcal{V}$, there exist constants $\epsilon_1, \epsilon_2 > 0$ such that $i_x(\epsilon) = n$ for all $\epsilon_1 < \epsilon < \epsilon_2$ and any point $x \in \mathbb{X}$. Moreover, ϵ_1 can be chosen arbitrarily small by increasing S.*

Proof. Let $x \in \mathbb{X}$ be an arbitrary point. Since M is a manifold, x admits a Euclidean neighbourhood U. Since M is smooth, we can assume U to be arbitrarily close to being flat by shrinking it. Thus, we can find $\epsilon_2 > 0$ with $B_r^s(x) \subset U$ for all $r, s < \epsilon_2$ such that $\mathrm{H}_i(\mathcal{V}(B_r^s(x), t)) = 0$ for all $i \geq n$, and all t. Hence, $\mathrm{PH}_i(\mathcal{V}(B_r^s(x), \bullet)) = 0$ for all $i \geq n$, and therefore $i_x(\epsilon_2) \leq n$. By contrast, for S sufficiently large, and r, s as before, there exists a parameter t such that $\mathcal{V}(B_r^s(x), t)$ is homotopy-equivalent to an $(n-1)$-sphere, and so $\mathrm{H}_{n-1}(\mathcal{V}(B_r^s(x), t))$ admits a generator, i.e. it is non-trivial. Consequently, $\mathrm{PH}_{n-1}(\mathcal{V}(B_r^s(x), \bullet)) \neq 0$, and $i_x(\epsilon_2) = n$. By further increasing S, we can ensure that the statement still holds when we decrease ϵ_2, which proves the two remaining claims. □

Finally, we prove that *Euclidicity*, an in particular its approximation in our implementation, i.e. Eq. (3), is consistent when dealing with manifolds in that it correctly assigns every point an infinitesimally small Euclidicity, provided a sufficient number of samples is available. This theorem was first stated on p. 5.

Theorem 3. *Let $M \subset \mathbb{R}^N$ be a smooth n-dimensional manifold and let $\mathbb{X} \subset M$ be a finite sample of size $S := |\mathbb{X}|$. For a given $\epsilon > 0$, sufficiently large S and a point $x \in \mathbb{X}$, there exists $s_\epsilon > 0$ that only depends on ϵ and the curvature of x (w.r.t. M), such that the approximation of $\mathfrak{E}(x)$ via Eq. (3) is bounded above by ϵ, for any grid Γ with maximum outer radius s_ϵ.*

Proof. Since M is a manifold, every point x admits a Euclidean neighbourhood U. Moreover, as M is smooth, we may assume U to be arbitrarily close to being flat (by potentially shrinking it). Thus, we can find an s_ϵ with the desired properties such that (i) $B_0^{s_\epsilon}(x) \subset U \cap \mathbb{X}$, and (ii) $\mathrm{d_H}(B_0^{s_\epsilon}(x), \mathbb{E}\mathrm{B}_0^{s_\epsilon}(x)) \leq \frac{\epsilon}{2}$, where the latter condition can be achieved by increasing S if necessary. Consequently, $\mathrm{d_H}(B_r^s(x), \mathbb{E}\mathrm{B}_r^s(x)) \leq \frac{\epsilon}{2}$ for *any* $0 \leq r \leq s \leq s_\epsilon$, and by making use of the geometric stability theorem (Chazal et al., 2014), the right-hand side of Eq. (3) reduces to

$$\frac{1}{C} \sum_{(r,s)\in\Gamma} \mathrm{d_B}(\mathrm{PH}_i(\mathcal{V}(B_r^s(x), \bullet)), \mathrm{PH}_i(\mathcal{V}(\mathbb{E}\mathrm{B}_r^s(x), \bullet))) \leq \frac{2}{C} \sum_{(r,s)\in\Gamma} \mathrm{d_H}(B_r^s(x), \mathbb{E}\mathrm{B}_r^s(x)) \leq \epsilon$$

for any grid Γ with the properties specified in the statement. This concludes the proof. □

A.3. Local Homology

Local homology quantifies topological properties of infinitesimal small neighbourhoods of a fixed point. For a topological space X and $x \in X$, its ith local homology group is defined as the categorical colimit $\mathrm{H}_i(X, X \setminus x) := \varinjlim_U \mathrm{H}_i(X, X \setminus U)$, where the direct system is given by the induced maps on homology that arise via the inclusion of (small) neighbourhoods of x. Heuristically, a local homology class can be thought of as a homology class of an infinitesimal small punctured neighbourhood of a point. When X is a simplicial complex, we may view x as a vertex in X, using subdivision if necessary. Its *star* $\mathrm{St}(x)$ is defined to be the union of simplices in X that have x as a face, whereas its *link* $\mathrm{Lk}(x)$ consists of all simplices in $\mathrm{St}(x)$ that do not have x as a face. By the excision axiom for homology, we have

$$\mathrm{H}_i(X, X \setminus x) \cong \mathrm{H}_i(\mathrm{St}(x), \mathrm{St}(x) \setminus x). \tag{4}$$

Since $\mathrm{St}(x)$ is *contractible*, the long exact reduced homology sequence of the pair $(\mathrm{St}(x), \mathrm{St}(x) \setminus x)$ records exactness of

$$0 = \tilde{\mathrm{H}}_i(\mathrm{St}(x)) \to \mathrm{H}_i(\mathrm{St}(x), \mathrm{St}(x) \setminus x) \to \tilde{\mathrm{H}}_{i-1}(\mathrm{St}(x) \setminus x) \to \tilde{\mathrm{H}}_{i-1}(\mathrm{St}(x)) = 0$$

for all i, and therefore $\mathrm{H}_i(\mathrm{St}(x), \mathrm{St}(x) \setminus x) \cong \tilde{\mathrm{H}}_{i-1}(\mathrm{St}(x) \setminus x)$. By noting that $\mathrm{St}(x) \setminus x$ deformation retracts to $\mathrm{Lk}(x)$, and by summarising the previous observations, we obtain

$$\mathrm{H}_i(X, X \setminus x) \cong \tilde{\mathrm{H}}_{i-1}(\mathrm{Lk}(x)). \tag{5}$$

The **key takeaway** here is that the homology of $\mathrm{Lk}(x)$ will usually differ from the homology of a sphere, once $\mathrm{Lk}(x)$ is not homotopy-equivalent to a sphere. For example, when x is an isolated singularity in a stratified simplicial complex X of dimension n, then its distinguished neighbourhood is given by $U \cong c^\circ L$. Thus, $\mathrm{Lk}(x) = L$ and $\mathrm{H}_i(X, X \setminus x) = \tilde{\mathrm{H}}_{i-1}(L)$ by Eq. (5), which is usually different from $\tilde{\mathrm{H}}_{i-1}(S^{n-1})$, when x does not admit a Euclidean neighbourhood in the appropriate dimension. In particular, points in a stratified simplicial complex that lie in strata of different dimensions can be distinguished in such a way, since homology of a sphere S^k records precisely one non-trivial generator in homology degree k. This observation motivates and justifies using local homology for detecting non-Euclidean neighbourhoods, and serves as the primary motivation for our *Euclidicity* measure in Section 4.2.

A.4. Pseudocode

We provide brief pseudocode implementations of the algorithms discussed in Section 4. In the following, we use $\#\,\mathrm{Bar}_i(\mathbb{X})$ to denote the number of i-dimensional persistent barcodes of $\mathbb{X}$ (w.r.t. the Vietoris–Rips filtration, but any other choice of filtration affords the same description). Algorithm 1 explains the calculation of *persistent intrinsic dimension* (see Section 4.1 in the main paper for details). For the subsequent algorithms, we assume that the estimated dimension of the intrinsic dimension of the data is n. We impose no additional requirements on this number; it can, in fact, be obtained by any choice of intrinsic dimension estimation method. As a short-hand notation, for $p_i = \mathrm{PH}_{n-1}(\mathcal{V}(\mathbb{E}\mathrm{B}_\bullet^\bullet(x), \bullet))$ w.r.t. some sample of $\{y \in \mathbb{R}^n \mid r \leq \mathrm{d}(x, y) \leq s\}$, we denote by $p_i^{r,s} = \mathrm{PH}_{n-1}(\mathcal{V}(\mathbb{E}\mathrm{B}_r^s(x), \bullet))$ the respective fibred persistent local homology barcode (calculated w.r.t. the same sample). Algorithm 2 then shows how to calculate the *Euclidicity* values, following Eq. (2) and one of its potential implementations, given in Eq. (3).

A.5. Stability of Euclidicity Estimates

Fig. 10 shows that Euclidicity is robust under sampling; repeating the calculations for smaller batches of the 'pinched torus' data set (500 points each) still lets us detect the singularity and its neighbours reliably. This robustness is an important

Algorithm 1 An algorithm for calculating the *persistent intrinsic dimension* (PID)

Require: $x \in \mathbb{X}$, $s_{\max}$, ℓ.
for $s \in \Gamma$ **do** {Iterate over the parameter grid}
  $i_x(s) \leftarrow 0$
  for $r < s \in \Gamma$ **do**
    for $i = 1, \ldots, N-1$ **do**
      Calculate $\# \operatorname{Bar}_i(B_r^s(x))$
      if $\# \operatorname{Bar}_i(B_r^s(x)) > 0$ **then**
        $i_x(s) \leftarrow i+1$
      end if
    end for
  end for
  return $i_x(s)$
end for

Algorithm 2 An algorithm for calculating the *Euclidicity values* δ_{jk}

Require: $x \in \mathbb{X}$, $s_{\max}$, ℓ, n, $\{p_1, \ldots, p_m\}$.
for $j = 1, \ldots, m$ **do**
  for $k = j+1, \ldots, m$ **do**
    for $s \in \Gamma$ **do**
      for $r \in \Gamma, r < s$ **do**
        Calculate $\mathrm{d_B}(p_j^{r,s}, p_k^{r,s})$ {Calculate bottleneck distance}
        return $\mathrm{d_B}(p_j^{r,s}, p_k^{r,s})$
      end for
    end for
    Calculate $\mathrm{D}(p_j, p_k)$ {Evaluate Eq. (3)}
    return $\mathrm{D}(p_j, p_k)$
  end for
end for

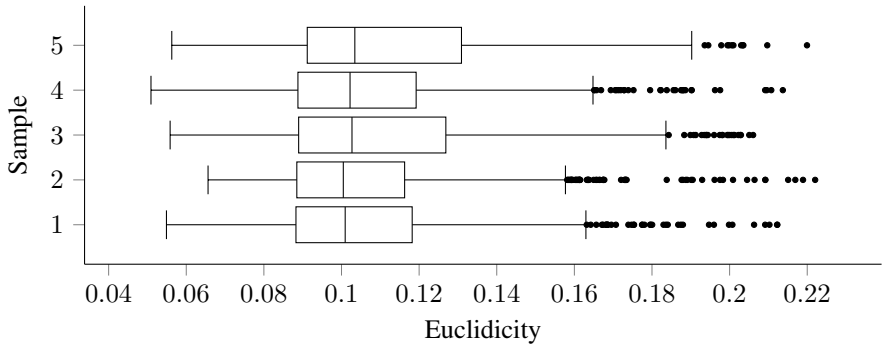
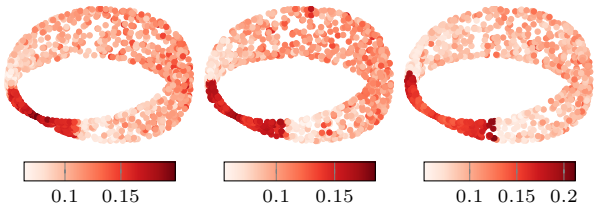


Figure 10: Boxplots of the Euclidicity values of different random samples of the 'pinched torus' data set. While each sample invariably exhibits some degree of geometric variation, we are able to reliably identify the singularity and its neighbourhood.



(a) $s_{\max} = 0.55$ *(b)* $s_{\max} = 0.65$ *(c)* $s_{\max} = 0.75$

Figure 11: Modifying the outer radius $s_{\max}$ still enables us to detect the singularity of the 'pinched torus.' Larger radii, however, progressively increase the field of influence of our method, thus starting to assign high Euclidicity values to larger regions of the point cloud.

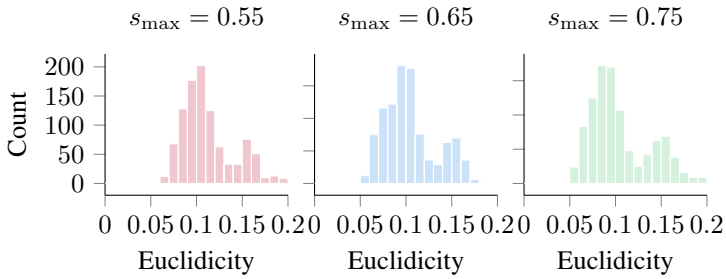


Figure 12: Histograms of the Euclidicity values for the point clouds shown in Fig. 11. Larger radii result in the distribution accumulating more probability mass at higher Euclidicity values, making the singularity detection procedure less local (but still succeeding in detecting the singularity and its environs).

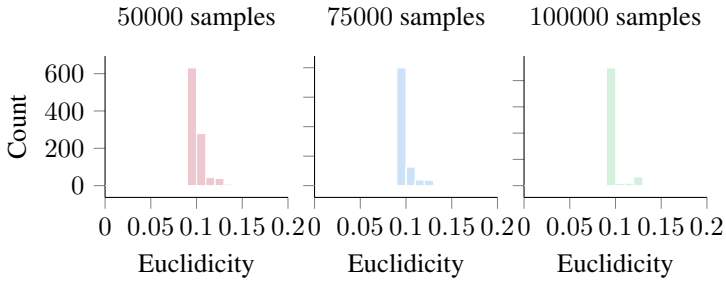
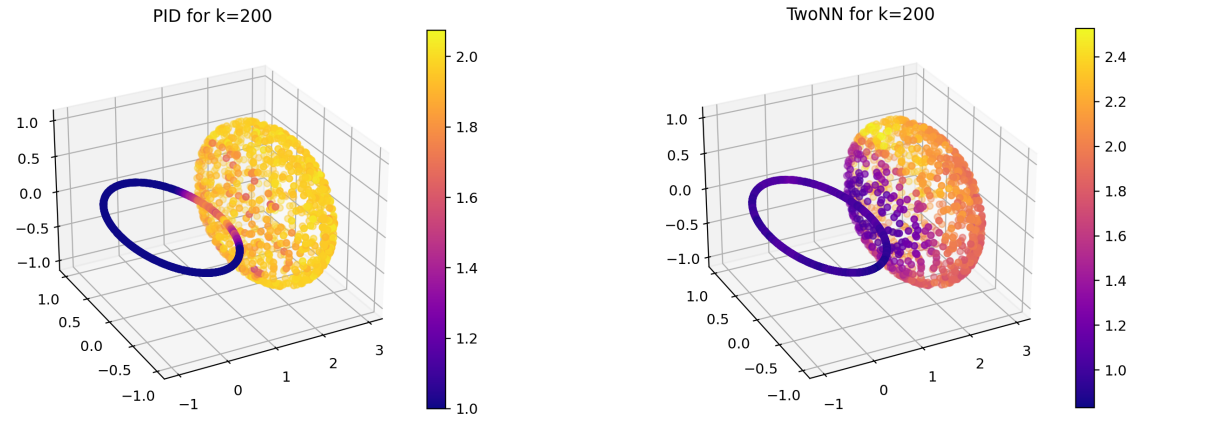


Figure 13: Histograms of the Euclidicity values for varying point sample densities. Each diagram represents Euclidicity values of 1000 random samples in the respective data space.



(a) PID *(b)* twoNN

Figure 14: Even for large values of k, PID still does not overestimate the local dimensionality of the data, exhibiting a clear distinction between the circle and the sphere, respectively.

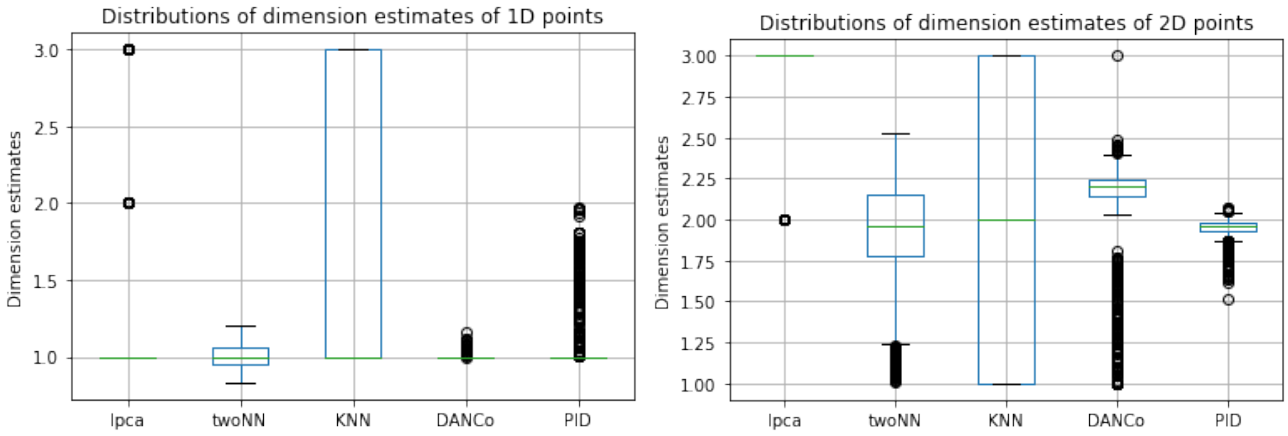
property in practice where we are dealing with samples from an unknown data set whose shape properties we want to capture. Euclidicity enables us to perform these calculations in a robust manner. Following the brief discussion in Section 5.1, we show the results of varying $s_{\max}$, the outer radius of the local annulus, for the 'pinched torus' data set. Fig. 11 depicts point clouds of 1000 samples; we observe that the singularity, i.e. the 'pinch point,' is always detected. For larger radii, however, this detection becomes progressively more *global*, incorporating larger parts of the point cloud. Fig. 12 depicts the corresponding histograms; we observe the same shift in probability mass towards the tail end of the distribution. For extremely large annuli, we estimate that we lose a clear distinction between singular values and non-singular values. Our data-driven parameter selection procedure is thus to be preferred in practice since it incorporates data density. However, even for markedly varying densities, we find that Euclidicity leads to reasonably stable results, as can be seen in Fig. 13. Here, we calculated Euclidicity scores for the pinched torus for varying data sample densities of $50000, 75000$ and 100000 samples. Each diagram shows the histogram of Euclidicity values for 1000 random samples in the respective data space. Although the histograms differ, the distributions exhibit the same overall characteristics.

A.6. Comparison of PID With Other Dimension Estimates

In order to assess the quality of PID, we decided to test its performance on a space that is both singular and has non-constant dimension. The data space we chose consists of 2000 samples of $S^1 \vee S^2$, i.e. a 1-sphere glued together with a 2-sphere at a certain concatenation point. We then applied the PID procedure for a maximum locality scale that was given by the k nearest neighbour distances, for $k \in \{25, 50, 75, 100, 125, 150, 175, 200\}$. We assigned to each point the average of the PID scores at the respective scales that are less than or equal to the k nearest neighbour bound. Subsequently, we compared the results with other local dimension estimates for the respective number of neighbours. The methods that were chosen for comparison include `lpca`, `twoNN`, `KNN`, and `DANCo`; we used the respective implementation from the `scikit-dimension` Python package.[5]

Fig. 14a shows the PID results for a maximum locality scale of 200 neighbours, with colours showing the estimated dimension values for each point. Overall, the correct intrinsic dimension is detected for most of the points. However, points that lie close to the singular point show a PID value between 1 and 2. Similarly to what we already discussed for Euclidicity, PID should therefore also be interpreted as a measure that incorporates the intrinsic dimension of a point on *multiple scales* of locality. For real-world data, the dimension will generally change when changing the locality scale. However, since there is no canonical choice of scale, we believe that any such scale provides valuable information about the intrinsic dimension that is worth being measured. We therefore argue that a multi-scale approach like ours is appropriate in practice, especially in a regime that is agnostic with respect to the underlying intrinsic dimension. By contrast, Fig. 14b shows the corresponding

[5] https://scikit-dimension.readthedocs.io/en/latest/



(a) 1-sphere dimension estimates

(b) 2-sphere dimension estimates

Figure 15: Estimates of the local intrinsic dimension for points that are close to the 1D-sphere, i.e. the circle, or the 2D-sphere, respectively.

dimension estimates for twoNN, where we observe less stable and reliable results across the dataset.

Fig. 15a shows boxplots of the distributions of the dimension estimates, for all points that lie on the 1D-sphere. We see that for PID, the mass is concentrated at a value of 1. Although there are outliers present, these correspond to points that are close to the singularity, as it was expected. We note that other methods like `KNN` and `lpca` might highly overestimate the dimension, and that the interquartile range is significantly higher for `twoNN` and `KNN`. Fig. 15b shows the same distributions for the points that lie on the 2D-sphere. Again, `lpca` highly overestimates the dimension since the median lies at a value of 3. Again, the interquartile range of PID is the tightest, and the estimates are closest to the ground truth. Moreover, the lower-value outliers again correspond to points that are close to the singular gluing point.

Fig. 16a and Fig. 16b show average dimension estimate scores of all investigated methods for varying values of k, both for points on the 1-sphere and the 2-sphere. We note that on average, only `twoNN` and `DANCo` lead to results which are comparable with the reliability of our method. However, as we already saw in Fig. 15a and Fig. 15b, the variance of the scores of our method is significantly lower, leading to more reliable outputs for each of the points.

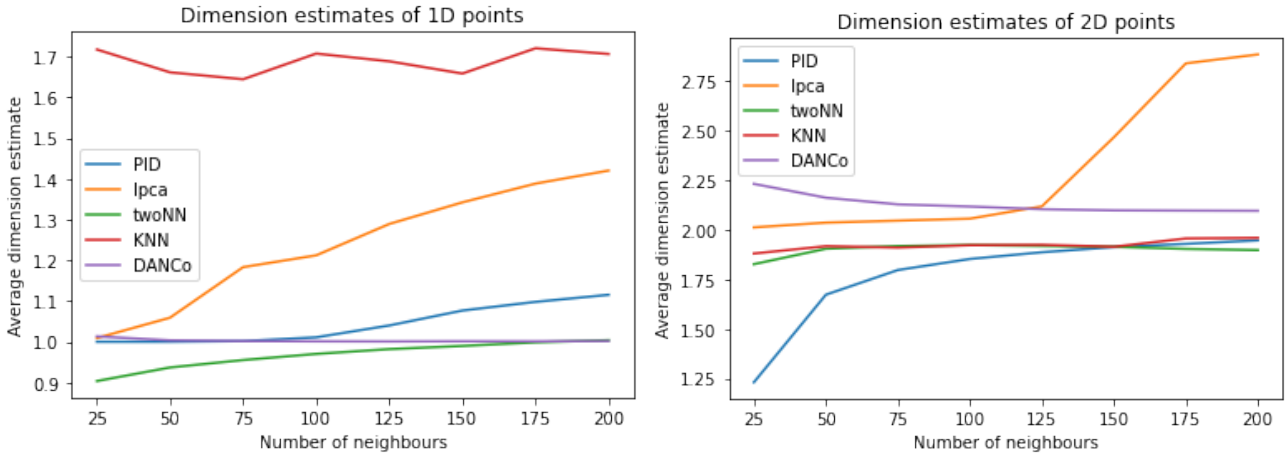
A.7. Euclidicity is More Expressive than Single-Parameter Approaches

Fig. 17 depicts a comparison between Euclidicity, an inherently multi-parametric approach, and a single-parameter approach, which assumes the *same* locality properties to hold for all points. This method is similar to the one described by Stolz et al. (2020). We observe that both methods are capable of detecting the isolated singularity of $S^2 \vee S^2$, with Euclidicity leading to a score that more clearly delineates the region around the singularity. Notice that this behaviour will become even more crucial in real-world data, which may exhibit strong sparsity patterns and varying density.

Further underscoring these findings, Fig. 18 shows the empirical distributions of Euclidicity scores for fixed locality parameters (left) and for our proposed multi-scale locality approach (right). We see that the variance is *significantly lower* in the multi-scale regime, indicating more stable and robust results. Moreover, the ratio of maximum and mean is higher in the multi-parameter setting, where high Euclidicity scores correspond to data points that lie close to the singularity, resulting in more reliable outcomes.

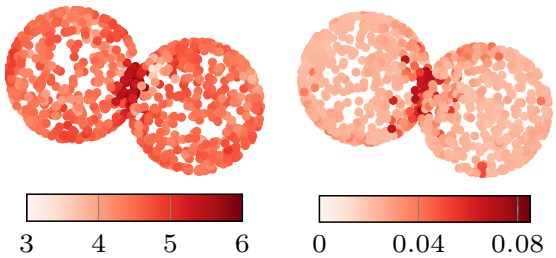
A.8. Euclidicity of MNIST and FASHIONMNIST

Fig. 19 and Fig. 20 show the Euclidicity results for the 4 additional runs on both the MNIST and FASHIONMNIST data sets. Again, we depicted the 9 images with lowest (left), medium (middle), and highest (right) Euclidicity scores for the two datasets. Moving from left to right, the images exhibit increases in the complexity of the local geometry, giving evidence for the reproducibility of the observation we remarked in Section 5.4.



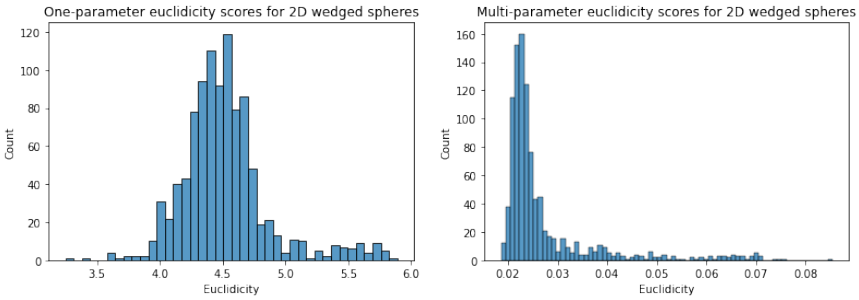
(a) *(b)*

Figure 16: Dimension estimates of the 1D-sphere and the 2D-sphere for different methods, plotted as a function of the number of neighours k.



(a) One-parameter approach *(b) Euclidicity*

Figure 17: In 2D, *Euclidicity* (b) results in a clearly-delineated singular region (also shown in the main paper), when compared to a single-parameter score (a).



(a) *(b)*

Figure 18: A comparison of Euclidicity values of a one-parameter approach (left) and our multi-parameter approach (right) demonstrates that multiple scales are necessary to adequately capture singularities.

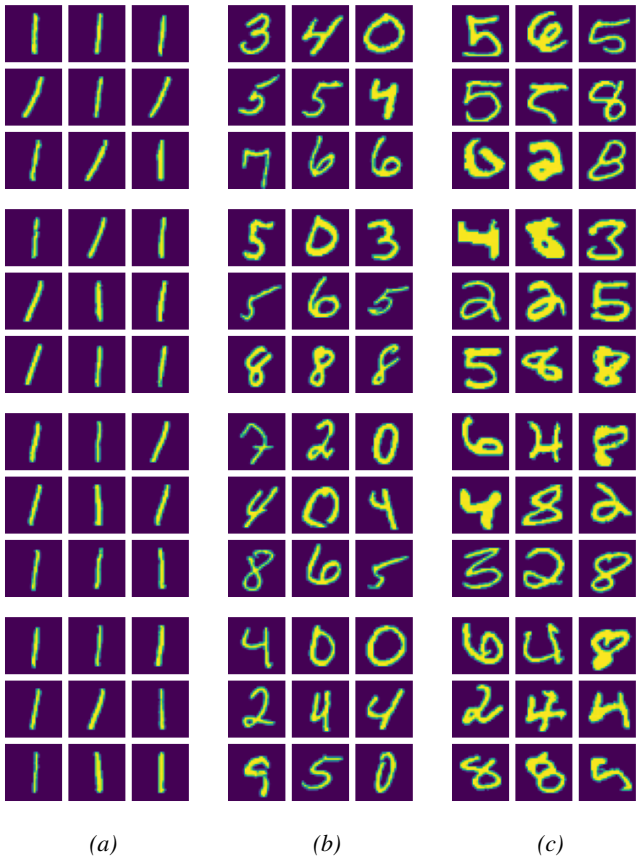
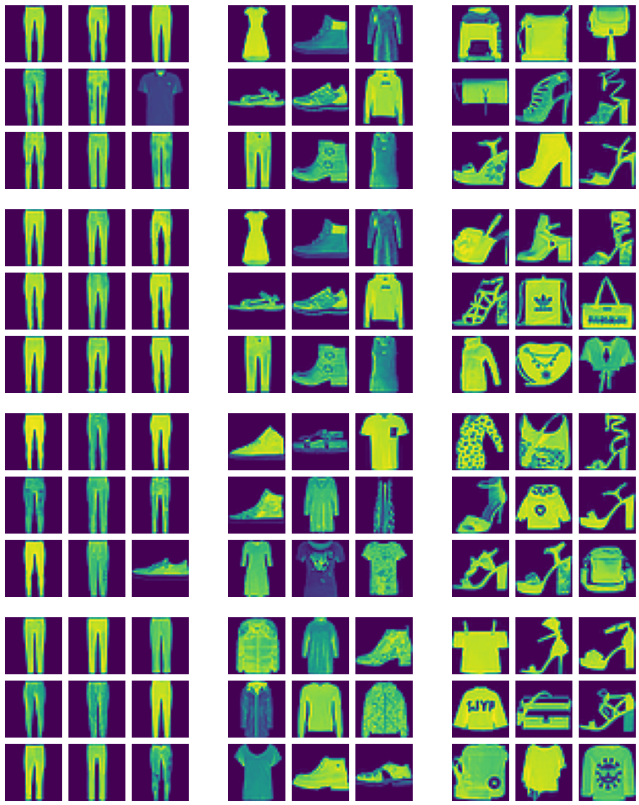
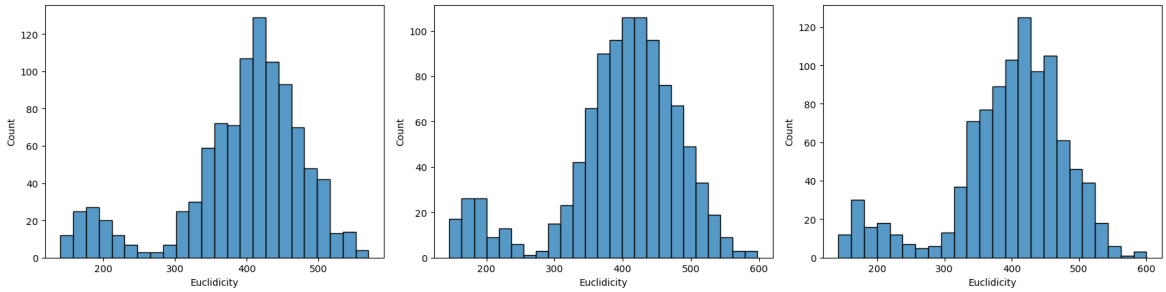


Figure 19: From left to right: more examples of low Euclidicity values, median Euclidicity values, and high Euclidicity values for the MNIST data set.



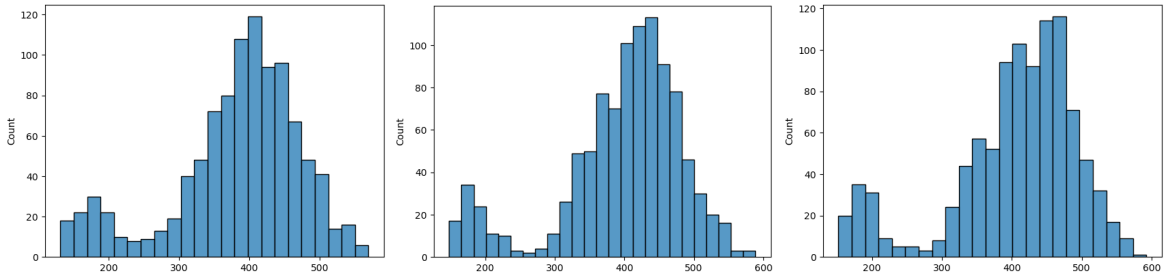
(a) *(b)* *(c)*

Figure 20: From left to right: more examples of low Euclidicity values, median Euclidicity values, and high Euclidicity values for the FASHIONMNIST data set.



(a) MNIST Euclidicity scores for different choices of the intrinsic dimension.

Figure 21: Left to right: distribution of MNIST Euclidicity scores for intrinsic dimension of 5, 10, and 15, respectively. As a consequence of the stability of our method, the distributions remain similar.



(a) MNIST Euclidicity scores for different choices of the nearest neighbour parameter k.

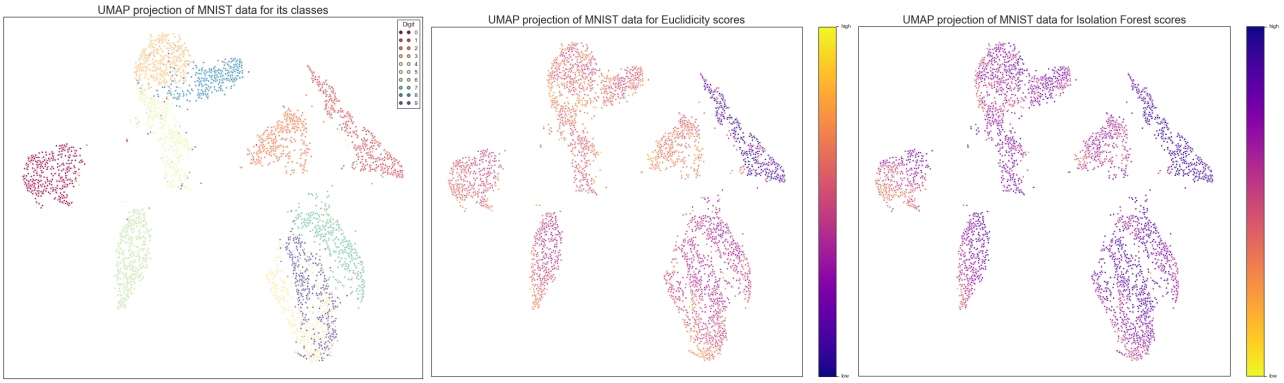
Figure 22: Left to right: distribution of MNIST Euclidicity scores for $k \in \{25, 50, 75\}$, respectively. As a consequence of the stability of our method, the distributions remain similar.

Moreover, as Fig. 25 shows, the empirical distributions of the calculated Euclidicity scores differ significantly for the MNIST and FASHIONMNIST data sets, with the distribution for MNIST exhibiting a bimodal behaviour, whereas the FASHIONMNIST Euclidicity value distribution is unimodal. We hypothesise that this corresponds to regions of simple complexity—and locally linear structures—in the MNIST data set, which are absent in the FASHIONMNIST data set. Finally, we observe that Euclidicity values for MNIST are surprisingly stable with respect to specific choices of the intrinsic dimension, as well as the nearest neighbour parameter k, which can be seen in Figs. 21 and 22. Here, we calculated Euclidicity scores for MNIST for intrinsic dimensions of $5, 10$ and 15, respectively, and $k \in \{25, 50, 75\}$. Although the scale of locality for which Euclidicity is computed varies significantly for different k, the respective distributions of Euclidicity values appear to be almost indistinguishable. We surmise that this is a consequence of the multi-scale nature of our approach.

Euclidicity and anomaly scores. Fig. 23 shows a comparison of Euclidicity scores and the scores of `IsolationForest`, an anomaly detection algorithm, for a UMAP projection of MNIST. Here, we used the implementation for `IsolationForest` from the `scikit-learn` Python package.[6] The clusters correspond to the different classes of MNIST, see Fig. 23a. Although the original purpose of Euclidicity is the detection of non-Euclidean neighbourhoods of points, we observe a correlation of low Euclidicity scores (i.e. points that are closer to admitting a Euclidean neighbourhood) and high `IsolationForest` scores (i.e. points that are less likely to be considered anomalies). This provides further evidence that Euclidicity could be expected to serve as a powerful anomaly detection method, on top of its ability to capture local geometric complexity. However, in our experiments we observe a (linear) correlation of only 0.5–0.7, indicating that Euclidicity detects *different* anomalous behaviour than just plain outliers. We plan on investigating this in future work.

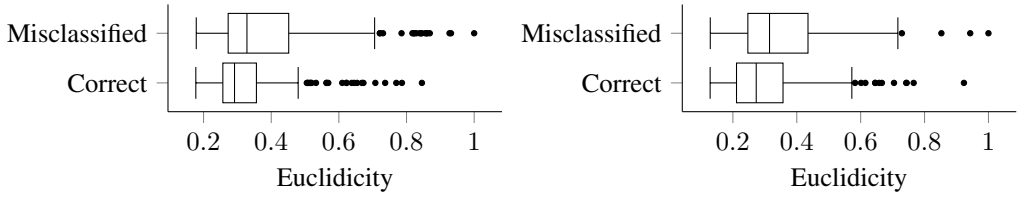
Details on neural network misclassification analysis. The neural network that we trained for the analysis in Section 5.4 consists of an input layer of 784 nodes, one dense hidden layer of 5 nodes and an output layer possessing 10 nodes,

[6] https://scikit-learn.org/stable/



(a) UMAP with class labels *(b)* UMAP Euclidicity scores *(c)* UMAP Isolation Forest scores

Figure 23: Euclidicity correlates with Isolation Forest scores for MNIST samples. Notice that we switched the colour maps because *low* Isolation Forest scores indicate anomalies.



(a) MNIST *(b)* FASHIONMNIST

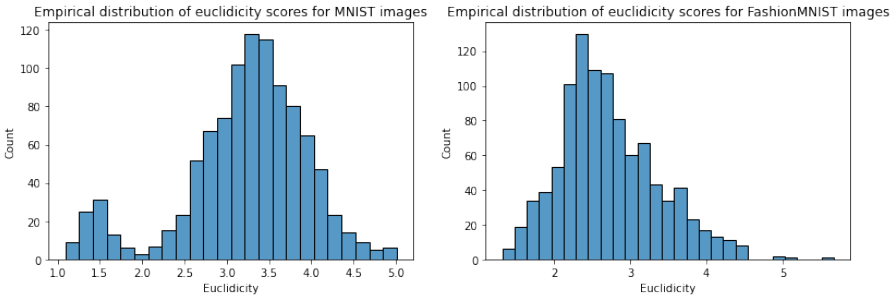
Figure 24: A comparison of Euclidicity scores for misclassified and correctly classified samples in two image data sets.

respectively. The same architecture was used for both MNIST and FASHIONMNIST, resulting in 3985 trainable parameters. Accuracy scores are 0.777 for MNIST, and 0.714 for FASHIONMNIST. Neither these scores nor the accuracy itself are supposed to be competitive; we merely use this network as an example to outline how Euclidicity could be potentially used to inform network training and highlight potentially problematic regions in data. Fig. 24 shows the respective results of Euclidicity scores for misclassified versus correctly-classified samples. To compare the statistical significance of the respective means for correctly and misclassified samples, we performed Welch's t-test, as the samples are approximately normally distributed. We obtained p-values of 2.8×10^{-5} in the case of MNIST, and 5.7×10^{-3} in the case of FASHIONMNIST. With respect to the common rejection thresholds in the literature, this leads to a clear rejection of the null hypothesis, which assumes the population means to be equal. However, we only claim a *correlation* between the likelihood for samples to be correctly classified and their corresponding Euclidicity scores. A statement in terms of *causality* is to be investigated in future work.

A.9. Euclidicity Captures Geometric Complexity in Histology Data

Finally, we applied Euclidicity to a benchmark histology data set.[7] The 96×96 pixel images were first scaled down to 60×60 pixels, to speed up the computation. Moreover, since the images are inherently coloured, we transformed them to greyscale images in order to naturally obtain a high-dimensional representation in Euclidean space, by flattening them. We then randomly picked 1000 of the training images and calculated Euclidicity scores of these samples (with respect to the whole space of training images). The intrinsic dimension was chosen to be 35, as suggested by `twoNN`. All the other hyperparameters were chosen as in the setting of Appendix A.8. The results of images admitting different scores are depicted in Fig. 26. Similarly to the observations in Appendix A.8, we see very different structural behaviour inside of the images for different Euclidicity values: while low scores correspond to images that admit large voids, medium-score images correspond to images with finer local geometric structure. Images of high Euclidicity scores, however, seem to admit *both* large voids *and* regions of fine-grained geometric structure. We therefore hypothesise the potential utility of Euclidicity for biological data, for instance to ensure data quality of (histology or biomedical) image datasets.

[7] https://github.com/basveeling/pcam



(a) MNIST *(b)* FASHIONMNIST

Figure 25: Both MNIST and FASHIONMNIST exhibit markedly different distributions in terms of Euclidicity: MNIST Euclidicity values are bimodal, whereas FASHIONMNIST Euclidicity values are unimodal.

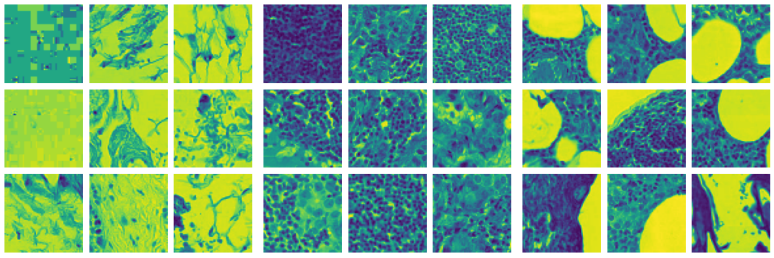


Figure 26: Left to right: sample histology images exhibiting low, median, and high Euclidicity, respectively. All samples are taken from the 'PatchCamelyon' benchmark data set.