CHALLENGING THE MANIFOLD HYPOTHESIS: A GEOMETRIC DIAGNOSTIC FOR DATA AND REPRESENTATION SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

The manifold hypothesis is widely invoked to motivate representation learning, dimensionality reduction, and geometric operations in high-dimensional data spaces. Yet it is often stated informally: data are said to lie near a low-dimensional "manifold" without specifying the geometric properties this object should satisfy or whether finite datasets exhibit them. We introduce the Geometric Manifoldness Index (GMI), a continuous diagnostic for two necessary signatures of smooth embedded manifold structure in point clouds: local intrinsic-dimension consistency and positive Federer reach. GMI is scalable to modern high-dimensional representations and separates dimension inconsistency from reach-related failures such as cusps and self-intersections. We validate GMI on synthetic manifolds, controlled singular examples, and image-transformation manifolds, then apply it to raw images, autoencoder latents, token embeddings, and single-cell transcriptomics data. Across these modalities, many data and representation spaces remain far from the smooth-manifold regime, even when low-dimensional structure is visible in projections or induced by compression. Learned embeddings become more manifold-like mainly when training objectives impose geometric regularity, and manifoldness shows no consistent positive correlation with reconstruction quality. Our results challenge common informal uses of the manifold hypothesis and motivate diagnosing manifoldness as an explicit geometric assumption rather than presuming it.

1 INTRODUCTION

High-dimensional data in machine learning is often assumed to have much lower-dimensional structure than its ambient representation suggests. For example, a fixed-size grayscale image can be represented as a vector in a high-dimensional pixel space, with one coordinate per pixel. Most vectors in this space look like random noise rather than meaningful images, which suggests that natural images occupy a much smaller subset of the full ambient space (Figure 1). The manifold hypothesis attempts to formalise this intuition by positing that this subset lies near a low-dimensional geometric structure. In the idealised version, this subset is taken to be a smooth embedded manifold, so that valid data can be described by few intrinsic degrees of freedom.

This intuition has been influential across representation learning (Bengio et al., 2014; Brahma et al., 2015), generative modelling, dimensionality reduction, and geometric analyses of embeddings. It helps explain why high-dimensional observations can sometimes be encoded by low-dimensional latent variables, and why traversals in learned latent spaces can produce smooth semantic changes. However, the term "manifold" is used loosely in the community. A low-dimensional or semantically meaningful representation need not be a smooth manifold in the *mathematical sense*: it may have singularities, self-intersections, boundaries, varying local dimension, disconnected components, or stratified structure, to name a few.

This distinction matters because many manifold-based methods implicitly rely on geometric properties that are stronger than low effective dimension alone. A smooth embedded manifold should have locally consistent intrinsic dimension and should avoid reach-zero pathologies such as cusps, intersections, and bottlenecks. Existing empirical support for the manifold hypothesis often demon-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

strates compression, visualisation structure, or low intrinsic dimension, but these observations do not by themselves establish manifoldness. This motivates a diagnostic question: *to what extent do real data and learned representations satisfy the geometric conditions expected of a smooth embedded manifold?*

In this work, we propose the Geometric Manifoldness Index (GMI), a continuous diagnostic for smooth embedded manifoldness in finite point clouds (Figure 2). We use it to evaluate raw data spaces and learned representation spaces across images, token embeddings, and single-cell transcriptomics. Our results suggest that many commonly studied data and embedding spaces do not behave like smooth low-dimensional manifolds under this operational definition, even when they exhibit low-dimensional or visually coherent structure. Rather than using the manifold hypothesis as an implicit justification, we argue that manifoldness should be made explicit, measured, and validated before manifold-based methods are applied.

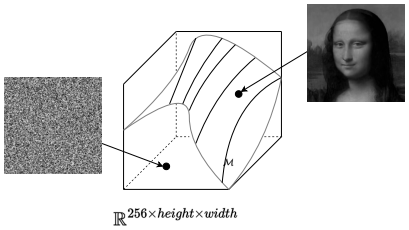


Figure 1: An illustration of the manifold hypothesis: natural images are often depicted as lying near a low-dimensional smooth manifold $\mathcal{M} \subset \mathbb{R}^{256\times h\times w}$, with off-manifold points interpreted as noise. This intuition need not hold mathematically. Our work tests whether idealised geometry is supported by real data and learned representations.

Our main contributions are: (I) We operationalise a geometric diagnostic for the manifold hypothesis by separating low-dimensional structure from two necessary signatures of smooth embedded manifoldness: local intrinsic-dimension consistency and positive Federer reach. (II) We implement GMI: a continuous, scalable diagnostic that combines a local intrinsic-dimension consistency score, s_{const}, with a positive-reach score, s_{reach}. We validate GMI on synthetic manifolds, controlled singular examples, and image-transformation manifolds. (III) We conduct empirical asymptotic studies across raw images, autoencoder latents, token embeddings, and single-cell transcriptomic data. Across these modalities, we find persistent low manifoldness in many raw data and learned representation spaces, challenging the manifold hypothesis under Definition 1. (IV) We show that manifoldness is not a reliable proxy for representation quality: on MNIST autoencoder families, GMI has no consistent positive relationship with test reconstruction quality and is sometimes anticorrelated with it.

Importantly, we would like to stress that GMI is an operational diagnostic rather than a certificate of manifoldness. High GMI values indicate finite-sample evidence for two necessary signatures of smooth embedded manifold structure but do not prove that the underlying support is a smooth manifold. Low GMI values should likewise be interpreted as failures of these criteria at the sampled scales, rather than as a formal impossibility result for all manifold models.

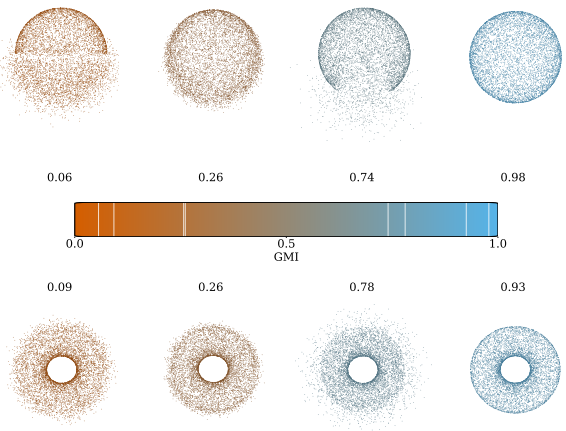


Figure 2: GMI on point clouds sampled from spheres (top row) and tori (bottom row) in $\mathbb{R}^3$ with increasingly strong additive Gaussian noise (left). GMI degrades continuously which is the behaviour required of a diagnostic for data that are only assumed to lie near a manifold.

2 RELATED WORK

The manifold hypothesis (M.H.) is central to representation learning, manifold learning, dimensionality reduction, and geometric operations on embeddings such as UMAP (Jeon et al., 2025; Robinson et al., 2025; Tenenbaum et al., 2000). Many methods implicitly prefer smooth low-dimensional geometry, including autoencoders (Kramer, 1991) and diffusion models (Song et al., 2020). Despite this reliance, the hypothesis is not well formulated: the literature mixes data-space and representation-space claims, imposes different smoothness requirements, and adopts conflicting structural interpretations (Brown et al., 2022; Fefferman et al., 2013; Li & Sarwate, 2025; Robinson et al., 2024; Wang & Wang, 2024). The M.H. also matters because learning rates and sample complexity can depend on intrinsic dimension (Block et al., 2022; Fefferman et al., 2013; Narayanan & Mitter, 2010) and it is sometimes quoted to justify techniques for manifold learning, as well as operations such as inner products, cosine similarity, and UMAP (McInnes et al., 2018; Jeon et al., 2025).

The M.H. is not well formulated in the existing literature. Some work treats raw data in $\mathbb{R}^D$ as concentrated near an embedded submanifold; others treat representation spaces $\mathcal{Z}$ as the manifold (Fefferman et al., 2013; Robinson et al., 2025). Regularity assumptions vary (C^1, C^2, C^∞, smooth, Riemannian), and alternative hypotheses include unions of manifolds (Brown et al., 2022), stratified spaces (Li & Sarwate, 2025), CW complexes (Wang & Wang, 2024), and fibre bundles (Robinson et al., 2025). In a footnote, Bengio et al. (2014) mention that any low-dimensional structure may consist of several (disconnected) manifolds of varying intrinsic dimension. This caveat has been forgotten in the literature but we emphasise it empirically. Extensive work supports low effective dimension in images and token embeddings (Pope et al., 2021; Ethayarajh, 2019; Whiteley et al., 2022), yet prior challenges remain fragmented, theoretical, and without a continuous cross-modality diagnostic.

Prior tests target only single necessary conditions and return binary decisions without a continuous measure of how close a finite point cloud is to a smooth embedded manifold (Lim et al., 2025; Robinson et al., 2025). HADES (Lim et al., 2025) aggregates per-point singularity p-values, and Robinson et al. (2025) apply a sliding-window intrinsic-dimension test. On synthetic shapes, these tests can miss reach-zero singularities and topological intersections that GMI flags through s_{reach} and s_{const}. The concept of data lying in the *vicinity* of a manifold necessitates a continuous diagnostic rather than traditional hypothesis testing in order to probe data spaces and embedding spaces where downstream models operate.

3 PRELIMINARIES

For this work, we consider $\mathcal{M}_d$ as a smooth embedded submanifold with positive reach, which aligns with existing canonical references (Bengio et al., 2014; Fefferman et al., 2013). We use the term *manifoldness* to refer to the degree to which a finite point cloud behaves like samples from a regular embedded manifold. This section briefly introduces key preliminaries. Further definitions are available in Appendix B.

Definition 1 (Manifold hypothesis (Bengio et al., 2014; Fefferman et al., 2013))**.** The M.H. states that real-world data in high-dimensional spaces $\mathbb{R}^D$ reside in the *vicinity* of a lower-dimensional *smooth* (at least $C^{1,1}$) manifold $\mathcal{M}_d$ embedded in $\mathbb{R}^D$, with $d \ll D$.

Definition 2 (Tangent space)**.** The tangent space $T_p\mathcal{M}$ at a point p on a smooth manifold is a vector space containing all possible smooth curves (tangent vectors) that pass through p.

Definition 3 (Reach (Federer, 1959))**.** Let $\mathcal{M} \subset \mathbb{R}^D$. The reach $\tau(\mathcal{M})$ is the largest $\tau \geq 0$ such that every point $x \in \mathbb{R}^D$ with $\operatorname{dist}(x, \mathcal{M}) < \tau$ has a unique nearest point on $\mathcal{M}$. Via tangent planes, Federer's characterisation is

$$\tau(\mathcal{M}) = \inf_{\substack{a,b \in \mathcal{M} \\ a \neq b}} \frac{\|b-a\|^2}{2\left\|(b-a) - T_a T_a^\top (b-a)\right\|}, \tag{1}$$

where T_a is an orthonormal basis matrix for $T_a\mathcal{M}$. Positive reach is a property of a $C^{1,1}$ smooth manifold (Fefferman et al., 2013; Aamari et al., 2019). Equation 1 is the basis of the s_{reach} diagnostic.

4 METHOD

We score point clouds in raw data space $\mathcal{X}$ or in learned embedding space $\mathcal{Z}$ (see Figure 10 for a summary) with GMI, a continuous scalar $\mathrm{GMI}(\mathbf{X}) \in [0, 1]$ that provides a diagnostic for how closely a point cloud $\mathbf{X} \in \mathbb{R}^{N \times D}$ satisfies two necessary conditions for smooth embedded submanifolds: (i) a consistent local intrinsic dimension d throughout the dataset, and (ii) positive reach $\tau > 0$ (i.e., note $\tau = 0$ typically indicates a singularity). Figure 3 illustrates the computation pipeline.

From the initial point cloud, we retain only active columns with variance above $\varepsilon_{\mathrm{col}} = 10^{-8} \cdot \max_k \mathrm{Var}(X_{:,k})$. A global Two-NN (Facco et al., 2017) query on X_{active} supplies $\varepsilon_{\mathrm{density}}$, $\varepsilon_{\mathrm{norm}}$ (median first-neighbour distance), and coarse dimension d_{coarse}. The median distance sets both the farthest-point sampling (FPS) covering scale ($\varepsilon_{\mathrm{density}}$) and reach normalisation ($\varepsilon_{\mathrm{norm}}$), which links patch density to sampling resolution.

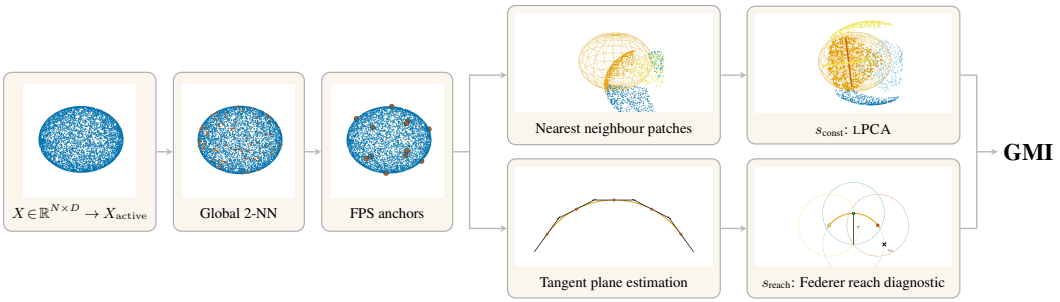


Figure 3: The GMI pipeline. We begin with point cloud X, and run Two-NN to obtain $\varepsilon_{\mathrm{density}}$, $\varepsilon_{\mathrm{norm}}$, and d_{coarse}. The constants are used to build an FPS anchor set of patch centres. Overlapping k-NN patches are used in LPCA calculation and tangent planes on FPS anchors provide the reach score.

4.1 LOCAL DIMENSIONAL CONSISTENCY

A d-dimensional manifold has the same intrinsic dimension at every point. The score s_{const} tests whether local dimension estimates agree across spatially overlapping patches of points. We use Fukunaga–Olsen (FO) local principal component analysis (PCA) as the default per-patch estimator (Fukunaga & Olsen, 1971). FPS builds patch centres, each patch uses adaptive k_{patch} clipped to minimum and maximum values (Hein & Audibert, 2005; Eldar et al., 1997). FO LPCA counts the number of significant singular values per patch. The equation for the score is:

$$s_{\mathrm{const}} = \exp\left(-\frac{\mathrm{std}(\boldsymbol{d})}{\log\big(1 + \mathrm{mean}(\boldsymbol{d})\big)}\right) \in (0, 1], \qquad \boldsymbol{d} = (d_{\mathrm{LPCA},1}, \ldots, d_{\mathrm{LPCA},n_{\mathrm{centres}}}) \tag{2}$$

and the common patch size is clipped by

$$k_{\mathrm{lo}} = \max\big(k_{\mathrm{patch,min}},\, 2D_{\mathrm{active}}\big), \qquad k_{\mathrm{hi}} = \min\big(k_{\mathrm{patch,max}},\, \lfloor N/4 \rfloor\big), \tag{3}$$

This means $s_{\mathrm{const}} = 1$ exactly when all patches agree (equation 2). Since s_{const} depends on the spread of the per-patch counts, it measures *consistency* of the intrinsic-dimension estimate across the cloud, not correctness. An estimator with a bias that is uniform over patches still gives $s_{\mathrm{const}} \approx 1$. FO LPCA, for example, returns $\hat{d} = d + 1$ on spheres with zero variance (Appendix Table 7), so $s_{\mathrm{const}} = 1$ despite the overestimate. Therefore, a high GMI suggests dimensional homogeneity, not an accurate d. Hyperparameter ablations are in the supplementary material.

4.2 POSITIVE REACH

Equation 1 is a population infimum over all pairs of $\mathcal{M}$ evaluated with exact tangent spaces, and a point cloud supplies neither component. Tangents are only available as plug-in estimates $\hat{T}_a$ from LPCA, and the infimum is attained at a single pair, so a literal plug-in version is decided by the

worst tangent estimate in the cloud and is driven towards zero by sampling noise. Aamari et al. (2019) show that plug-in reach is stable only on sparsified samples, and existing reach estimators are analysed on low-dimensional, densely sampled manifolds. Our target regime is the opposite, $N \geq 10^4$ points in $D_{\text{active}} \geq 10^3$ dimensions, where those guarantees do not apply. We therefore compute a *proxy* for reach and accept a known bias in exchange for a quantity that is computable for modern high-dimensional data.

Two relaxations make the computation tractable. First, we restrict the pairs: anchors are the FPS net of Section 4, following the recommendation of Aamari et al. (2019) to sparsify before plug-in reach estimation, and each anchor is paired only with its k_{reach} nearest points. Second, we aggregate mean-then-min: the ratios at an anchor are separated only at the k-NN scale, so we treat them as noisy observations of one local curvature scale and average them into $\hat{\tau}_i$, then take $\hat{\tau} = \min_i \hat{\tau}_i$, which retains the Federer-like step over the worst region. The score s_{reach} maps the normalised proxy monotonically into $[0, 1]$ (equation 7).

Both relaxations act in the direction of our claims and can only raise $\hat{\tau}$ relative to the infimum over all pairs. Local pairs cannot see bottlenecks between sheets that are far apart along $\mathcal{M}$ but close in $\mathbb{R}^D$, and the inner mean discards the worst pair at each anchor, so $\hat{\tau}$ sits above the quantity in equation 1 and s_{reach} is a proxy. We keep this bias explicit: a low s_{reach} suggests reach 0 singularities, but a high s_{reach} implies that no reach failure was found at the sampled scale. Furthermore, s_{reach} requires ambient dimension exceeding intrinsic dimension (Section 6).

The final index is the geometric mean of two components:

$$\text{GMI} = \left(s_{\text{const}} \cdot s_{\text{reach}}\right)^{1/2} \in [0, 1]. \tag{4}$$

Both components must be large for high GMI; if either is zero, $\text{GMI} = 0$. The construction parallels Fefferman's class $\mathcal{G}(d, V, \tau)$ of compact d-manifolds with volume $\leq V$ and reach $\geq \tau > 0$ (Fefferman et al., 2013; Narayanan & Mitter, 2010), but as a top-down computationally tractable method on the observed cloud rather than an exhaustive bottom-up search over (d, V, τ).

5 Results

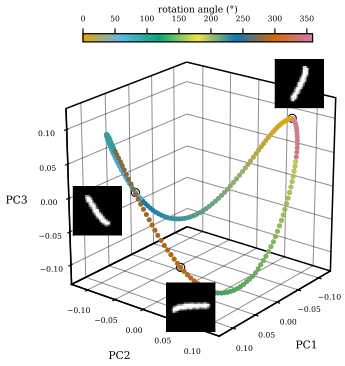
We report GMI across controlled synthetic examples, raw data spaces, and learned embeddings[1]. Complete dataset specifications, hyperparameters, and ablation studies are in Appendix E. We use a fixed $\alpha_{\text{FO}} = 0.175$ across all experiments, determined on synthetic shapes and yielding MNIST dimension estimates consistent with prior work (Brown et al., 2022; Pope et al., 2021).

Validating the geometric manifoldness index On synthetic shapes in $\mathbb{R}^{128}$, true manifolds score high ($\text{GMI} > 0.83$) while nearly-smooth and non-manifold controls score lower (Table 1; Figure 13). The score separates clean embeddings from forced self-intersections (Klein bottle in $\mathbb{R}^4$ versus $\mathbb{R}^3$), identifies reach-zero points on a lollipop shape, and penalises dimension inconsistency under noise. Validation extends to high intrinsic and ambient dimension (Figure 12; Table 2). A synthetic 1-D rotation manifold of an example MNIST image recovers high GMI. We rotate a randomly selected digit 1 image through $N = 360$. The equally spaced point cloud takes constant-angle views of this orbit ($n = 72$ is $5°$); and random subsets draw n of the 360 views uniformly without replacement. Equally spaced clouds stay high at every subsample ($\text{GMI} \geq 0.73$), while random subsets score lower at small n, where uneven angular coverage degrades both components, and increase with n towards the full-orbit score (Figure 4).

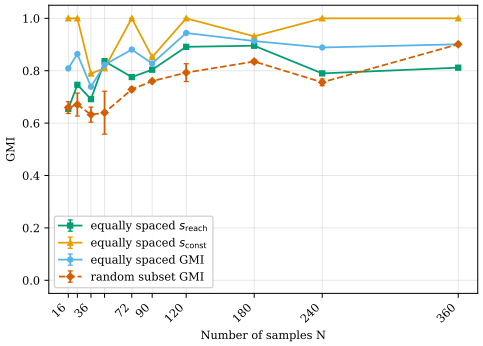
5.1 Manifoldness in data and embedding spaces

Image and text token point clouds provide evidence challenging the M.H. We tested four image datasets in raw pixel space: MNIST, Fashion-MNIST, CIFAR-10, and SVHN. All-class GMI declines with N, driven by collapsing s_{const} while s_{reach} stays comparatively flat (Figure 5). In the case of a smooth embedded manifold we would expect both components to stabilise at non-zero values as N grows; instead dimensional inconsistency dominates. Per-class scores exceed all-class scores substantially, consistent with the union of manifolds theory (Brown et al., 2022), which suggests class boundaries introduce non-manifold structure absent within individual classes. For digit 0,

[1]Unless otherwise stated, error bars indicate ± 1 standard deviation over five seeds.

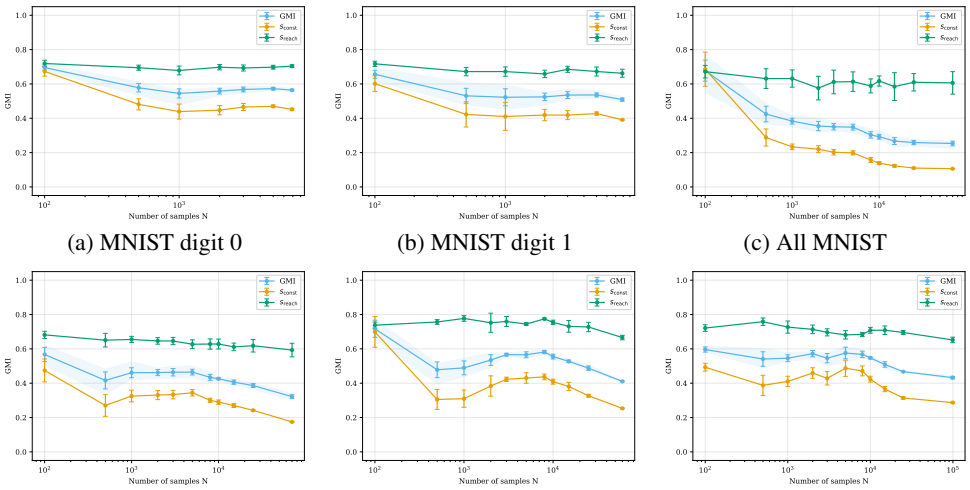


(a) PCA-3D of the $N = 360$ in-plane rotation orbit.



(b) Equally spaced s_{reach}, s_{const}, and GMI, and random subset GMI (mean $\pm$ 1 s.d. over 3 seeds). At $n = 360$, GMI coincides.

Figure 4: Results of the synthetic rotation experiment for one randomly drawn MNIST digit 1.



(a) MNIST digit 0 (b) MNIST digit 1 (c) All MNIST

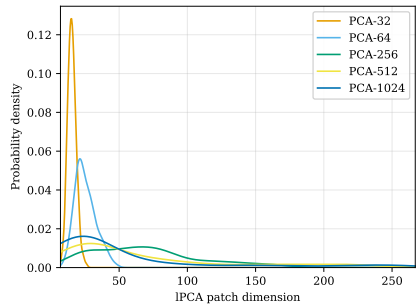
(d) All FMNIST (e) All CIFAR-10 (f) All SVHN

Figure 5: Raw-pixel GMI vs. N. Top: MNIST digit 0, digit 1, and the all-class pool. Bottom: all-class Fashion-MNIST, CIFAR-10, and SVHN. Per-class panels for Fashion-MNIST, CIFAR-10, SVHN, and additional MNIST classes are in Figure 27.

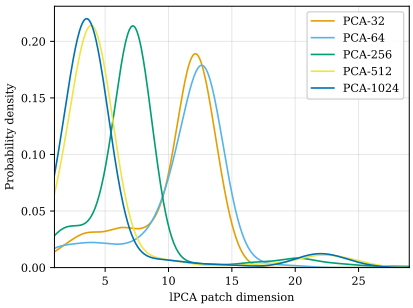
higher GMI may reflect stroke-shape variation along PC1 (Figure 28). Nevertheless, per-class GMI also remains below MNIST synthetic positive controls. This provides a challenge to claims for a smooth (disconnected) manifold of constant intrinsic dimension. One interesting note is that the per-class scores bound manifoldness within a semantic partition but do not exhaust unions of manifolds or stratified supports that need not align with class labels. Asymptotic studies of token embeddings provide supporting evidence for the same conclusions (Figure 35).

Two-stage pipelines confound scRNA-seq manifoldness. scRNA-seq data offer a setting where $D \gtrsim 17{,}000$ and N is often only thousands of cells (Table 3). Many scRNA-seq processing methods linearly reduce D before clustering and visualisation (Huang et al., 2022). In the Puram dataset, GMI collapses near zero but PCA projection can inflate scores by compressing LPCA rather than through manifold-like geometry (Figure 6; full asymptotic GMI results are in Appendix J). In contrast, Baron2 stays comparatively stable across PCA caps.

Higher manifoldness tracks projection stability. Stable manifoldness in Baron2 across PCA projections mirrors qualitatively consistent UMAP and t-SNE visualisations. For example, all 3 UMAP visuals for Baron2 indicate a strand of points protruding left from a central cluster (Fig-



(a) Puram LPCA density



(c) Baron2 LPCA density



(b) Puram has less qualitatively stable UMAP and t-SNE visualisations



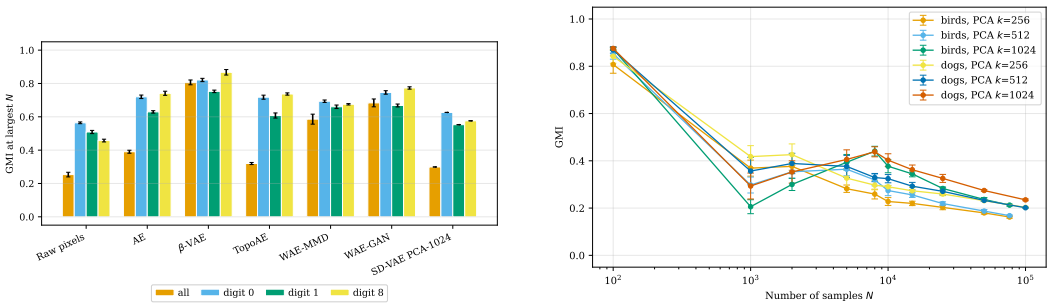
(d) Baron2 has qualitatively stabler UMAP and t-SNE visualisations

Figure 6: Puram and Baron2 scRNA-seq PCA sweeps and visualisations. Accompanying asymptotic GMI vs. N charts are in Appendix J. $N = 5{,}000$ for Puram and Baron2 and visualisations present for PCA-D' representations ($D' \in \{32, 256, 1024\}$), coloured by per-point d_{LPCA}. Baron2 has higher and more consistent manifoldness across PCA dimensions and qualitatively more stable low-dimensional visualisations than Puram.

ure 6). Puram's unstable projections may be due to dimensional inconsistency. Even in the $N \ll D$ regime, GMI can flag when dimensionality-reduction comparisons are confounded by projection choice (Huang et al., 2022); and any supporting evidence on linear projections alone does not establish that raw gene space satisfies the manifold hypothesis.

Autoencoders learn more manifold-like per-digit latents but not at dataset scale. We evaluated six encoder families on MNIST: a standard autoencoder (AE), variational autoencoder (VAE), topological autoencoder (TopoAE), Wasserstein autoencoder with a maximum mean discrepancy based penalty (WAE-MMD), Wasserstein autoencoder with a generative adversarial network based penalty (WAE-GAN), and stable diffusion variational autoencoder (SD-VAE). Further details are in Appendix E.1. Per-digit AE latents score higher than all-digit representation spaces (Figure 7a) and both s_{const} and s_{reach} diverge with N, which is similar to the raw-pixel space results. VAE all-digit scores are elevated by KL regularisation toward a spherical prior which suggests that manifold-like geometry is imposed by construction rather than discovered in the data. We also evaluated SD-VAE latents on ImageNet birds and dogs. While the other encoders are ordered roughly by distributional regularisation strength, SD-VAE latents show low s_{const} from dimensional inconsistency even within classes (Figures 7b and 30).

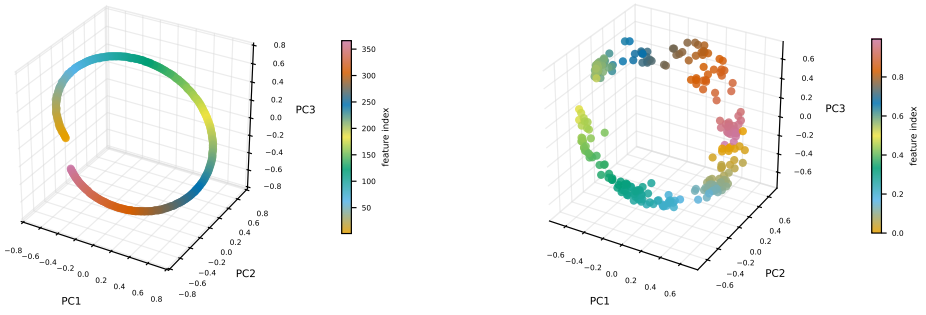
The GMI is capable of analysing token feature manifolds. We tested seven token-embedding tables: GPT-2, BERT, T5, Mistral, Pythia, Llemma, and OpenAI 3. Full token datasets exhibit



(a) MNIST GMI at largest N by encoder family and digit subset.

(b) ImageNet SD-VAE bird/dog supercategories: GMI vs. N for PCA-k.

Figure 7: Manifoldness of learned image latents.



(a) Dates (GMI $= 0.556$, $n = 366$).

(b) Colours (GMI $= 0.025$, $n = 265$).

Figure 8: 3D PCA projections of token subsets on OpenAI 3.

very low GMI (Figure 35) while curated subsets (dates, suffixes) can score higher. One reason is that PCA could be inflating manifoldness through the same projection-cap artefact as scRNA-seq. However, a token subset such as calendar dates still exhibits locally smooth structure consistent with Modell et al. (2025). Figure 8 demonstrates that English colour names trace a coherent hue loop in a 3D PCA projection yet have near zero GMI. This suggests structure visible in low-dimensional projections is not sufficient for evidence of feature manifolds.

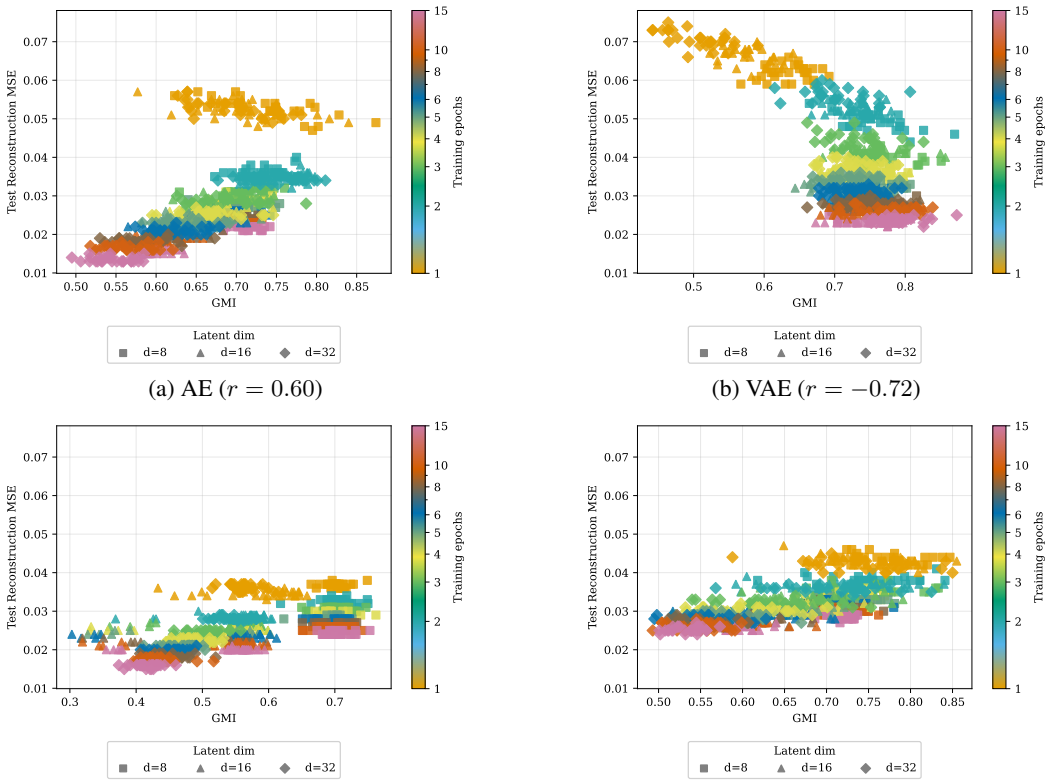
5.2 Manifoldness and Representation Quality

We trained five MNIST encoder families across latent dimensions $d \in \{8, 16, 32\}$ with 30 random seeds each, measuring test reconstruction MSE. Figure 9 plots GMI against MSE.

Manifoldness does not reliably predict representation quality. For AE, TopoAE, and WAE-MMD, higher manifoldness correlates with *worse* reconstruction: perhaps since higher latent dimensions likely contain richer non-manifold structure with lower MSE. VAE and WAE-GAN reverse the sign under distributional regularisation. Our results indicate that direction of the manifoldness–quality relationship is determined by encoder inductive bias, not by any universal property of manifold-like geometry in raw data and embedding spaces. A manifold-like latent is also not a prerequisite for strong reconstruction.

6 Discussion and Future Work

The GMI is computable on high-dimensional point clouds and is empirically validated on synthetic controls. Across modalities, we find evidence counter-intuitive to the M.H. under Definition 1 as GMI remains low even where PCA projections or prior work suggest low-dimensional manifold structure. Meanwhile, encoders raise manifoldness primarily when objectives impose it. A contin-



(a) AE ($r = 0.60$)

(b) VAE ($r = -0.72$)

(c) TopoAE ($r = 0.62$)

(d) WAE-MMD ($r = 0.70$)

Figure 9: GMI vs. test MSE on MNIST latents (30 seeds, $d \in \{8, 16, 32\}$; WAE-GAN: Figure 31).

uous *diagnostic* expands how data and representation spaces are analysed and we hope it motivates using manifoldness to gauge the validity of techniques that rely on manifold assumptions. Below we note caveats and directions for future work; Appendix A.2 extends the discussion.

Limitations of the reach score. When the entire embedding space is to be considered a manifold, d_{reach} saturates D_{active} and s_{reach} is undefined since a normal tangent plane for the point cloud cannot be estimated. This implies s_{reach} is also not reparametrisation-invariant, as the reach calculation uses the ambient Euclidean metric of $\mathcal{Z}$ (Arvanitidis et al., 2018; Hauberg, 2022; Shao et al., 2017). Our reach diagnostic is also sensitive to sampling near singular loci (see the synthetic cone shape).

Lack of manifoldness to motivate rebuilding new techniques from mathematical foundations. Where manifoldness is absent, we should not force a manifold model but rebuild tools on (non-Euclidean) foundations that drop the manifold assumption. Disjoint and union-of-manifold learning (Khayatkhoei et al., 2018; Brown et al., 2022; Elhamifar & Vidal, 2013; Vidal, 2011), stratified, CW-complex, and fibre-bundle hypotheses (Li & Sarwate, 2025; Wang & Wang, 2024; Robinson et al., 2025), and complex generative supports via normalising flows (Yu et al., 2025; Papamakarios et al., 2021; Kingma & Welling, 2013) are possible approaches. Persistent homology, stratified Morse theory, fractal geometry, and optimal transport (Edelsbrunner & Harer, 2008; Goresky & MacPherson, 1988; de Ocáriz Borde & Kratsios, 2024; Falconer, 2003; Peyré & Cuturi, 2019) describe singular geometry yet remain under-explored in machine learning. The Mona Lisa illustrates the challenge: this image plausibly lies on several semantic manifolds at once: the manifold of women, of Renaissance art, of portraits, of tourism and of France. This would likely represent a self-intersection where local structure is no longer that of a smooth manifold (Whitney, 1944; Jakubowski et al., 2020).

REPRODUCIBILITY STATEMENT

Anonymised source code is provided in the supplementary material. Dataset specifications and all hyperparameters are in Appendix E (Tables 3–5); further implementation details are described in Appendix G and Appendix F. Ablation studies are in Appendix I.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, implement experiments, clean and reformat datasets, and provide feedback on research methodology. We have not used generative AI tools to propose or refine hypotheses, or to interpret results. The rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artefacts produced with the aid of generative AI.

REFERENCES

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359–1399, 2019. URL https://arxiv.org/abs/1705.04565.

Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 29–38, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783405. URL https://doi.org/10.1145/2783258.2783405.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: On the curvature of deep generative models. In *International Conference on Learning Representations*, 2018.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics, 2023. URL https://arxiv.org/abs/2310.10631.

Maayan Baron, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K. Wagner, Shai S. Shen-Orr, Allon M. Klein, Douglas A. Melton, and Itai Yanai. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems*, 3(4):346–360.e4, 2016. doi: 10.1016/j.cels.2016.08.011.

Aimée Bastidas-Ponce, Sophie Tritschler, Leander Dony, Katrin Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian J. Theis, and Heiko Lickert. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146(12), 2019. doi: 10.1242/dev.173849. URL https://doi.org/10.1242/dev.173849.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014. URL https://arxiv.org/abs/1206.5538.

Zelong Bi and Pierre Lafaye de Micheaux. A survey and comparative evaluation of intrinsic dimension estimators under the manifold hypothesis. *arXiv preprint arXiv:2509.15517*, 2025. URL https://arxiv.org/abs/2509.15517.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2304.01373.

Adam Block, Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. Intrinsic dimension estimation using Wasserstein distance. *Journal of Machine Learning Research*, 23(313):1–37, 2022. URL https://jmlr.org/papers/v23/21-0815.html.

Pratik Prabhanjan Brahma, Dapeng Wu, and Yiyuan She. Why deep learning works: A manifold disentanglement perspective. *IEEE transactions on neural networks and learning systems*, 27 (10):1997–2008, 2015.

Bradley CA Brown, Anthony L Caterini, Brendan Leigh Ross, Jesse C Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. *arXiv preprint arXiv:2207.02862*, 2022. URL https://arxiv.org/abs/2207.02862.

Spyros Darmanis, Steven A. Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M. Shuer, Melanie G. Hayden Gephart, Ben A. Barres, and Stephen R. Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015. doi: 10.1073/pnas.1507125112.

Haitz Sáez de Ocáriz Borde and Anastasis Kratsios. Neural snowflakes: Universal latent graph inference via trainable latent geometries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.15003.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. URL https://doi.org/10.1109/CVPR.2009.5206848.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. URL http://yann.lecun.com/exdb/mnist/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019. URL https://arxiv.org/abs/1810.04805.

Herbert Edelsbrunner and John Harer. Persistent homology - a survey. In *Surveys on Discrete and Computational Geometry: Twenty Years Later*, volume 453 of *Contemporary Mathematics*, pp. 257–282. American Mathematical Society, Providence, RI, 2008. URL http://dx.doi.org/10.1090/conm/453/12.

Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y. Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997. doi: 10.1109/83.623193.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013. URL https://arxiv.org/abs/1203.1005.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, Sep 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-11873-y. URL https://doi.org/10.1038/s41598-017-11873-y.

Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, Chichester, 2 edition, 2003. URL https://doi.org/10.1002/0470013850.

Tyler Farghly, Peter Potaptchik, Samuel Howard, George Deligiannidis, and Jakiw Pidstrigach. Diffusion models and the manifold hypothesis: Log-domain smoothing is geometry adaptive, 2025. URL https://arxiv.org/abs/2510.02305.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *arXiv preprint arXiv:1310.0425*, 2013. URL https://arxiv.org/abs/1310.0425.

K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. Comput.*, 20(2):176–183, February 1971. ISSN 0018-9340. doi: 10.1109/T-C.1971.223208. URL https://doi.org/10.1109/T-C.1971.223208.

Mevlana C. Gemici, Danilo Jimenez Rezende, and Shakir Mohamed. Normalizing flows on riemannian manifolds, 2016. URL https://arxiv.org/abs/1611.02304.

Anna C. Gilbert and Kevin O'Neill. Ca-pca: Manifold dimension estimation, adapted for curvature. *arXiv preprint arXiv:2309.13478*, 2024. URL https://arxiv.org/abs/2309.13478.

Mark Goresky and Robert MacPherson. *Stratified Morse Theory*, volume 14 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, Berlin, 1988. URL https://doi.org/10.1007/978-3-642-71714-7.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Søren Hauberg. Differential geometry for generative modeling. ACML 2021 tutorial notes, 2022. URL https://www2.compute.dtu.dk/~sohau/weekendwithbernie/.

Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *Proceedings of the 22nd international conference on Machine learning*, pp. 289–296, 2005. URL https://doi.org/10.1145/1102351.1102388.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Christian Horvat and Jean-Pascal Pfister. Intrinsic dimensionality estimation using normalizing flows. *Advances in Neural Information Processing Systems*, 35:12225–12236, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4f3878b6639d6e4b8523c9135064a37c-Paper-Conference.pdf.

Christian Horvat and Jean-Pascal Pfister. On gauge freedom, conservativity and intrinsic dimensionality estimation in diffusion models. *arXiv preprint arXiv:2402.03845*, 2024.

Michael E. Houle. Local intrinsic dimensionality i: An extreme-value-theoretic foundation for similarity applications. In Christian Beecks, Felix Borutta, Peer Kröger, and Thomas Seidl (eds.), *Similarity Search and Applications*, pp. 64–79, Cham, 2017. Springer International Publishing. ISBN 978-3-319-68474-1. URL https://doi.org/10.1007/978-3-319-68474-1_6.

Haiyang Huang, Yingfan Wang, Cynthia Rudin, and Edward P. Browne. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Communications Biology*, 5(1):719, 2022. doi: 10.1038/s42003-022-03628-x.

Alexander Jakubowski, Milica Gasic, and Marcus Zibrowius. Topology of word embeddings: Singularities reflect polysemy. In Iryna Gurevych, Marianna Apidianaki, and Manaal Faruqui (eds.), *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pp. 103–113, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.starsem-1.11/.

Noémie Jaquier, Leonel Rozo, Miguel González-Duque, Viacheslav Borovitskiy, and Tamim Asfour. Bringing motion taxonomies to continuous domains via GPLVM on hyperbolic manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2210.01672.

Hyeon Jeon, Jeongin Park, Sungbok Shin, and Jinwook Seo. Stop misusing t-sne and umap for visual analytics, 2025. URL https://arxiv.org/abs/2506.08725.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Hamidreza Kamkari, Brendan L Ross, Rasa Hosseinzadeh, Jesse C Cresswell, and Gabriel Loaiza-Ganem. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. *Advances in Neural Information Processing Systems*, 37:38307–38354, 2024. URL https://arxiv.org/abs/2406.03537.

Mahyar Khayatkhoei, Maneesh K. Singh, and Ahmed Elgammal. Disconnected manifold learning for generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1806.00880.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kin Kwan Leung, Rasa Hosseinzadeh, and Gabriel Loaiza-Ganem. On convolutions, intrinsic dimension, and diffusion models, 2025. URL https://arxiv.org/abs/2506.20705.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. In L. Saul, Y. Weiss, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/74934548253bcab8490ebd74afed7031-Paper.pdf.

Xin Li and Anand Sarwate. Unraveling the localized latents: Learning stratified manifold structures in llm embedding space with sparse mixture-of-experts. *arXiv preprint arXiv:2502.13577*, 2025. URL https://arxiv.org/abs/2502.13577.

Uzu Lim, Harald Oberhauser, and Vidit Nanda. Hades: Fast singularity detection with local measure comparison. *SIAM Journal on Mathematics of Data Science*, 7(4):1882–1903, 2025. doi: 10.1137/24M1665763. URL https://doi.org/10.1137/24M1665763.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018. URL https://arxiv.org/abs/1801.02613.

David J. C. MacKay and Zoubin Ghahramani. Comment on "maximum likelihood estimation of intrinsic dimension" by e. levina and p. bickel, 2005. URL http://www.inference.org.uk/mackay/dimension/.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models, 2025. URL https://arxiv.org/abs/2505.18235.

Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019. doi: 10.1038/s41587-019-0336-3. URL https://doi.org/10.1038/s41587-019-0336-3.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/1906.00722. arXiv:1906.00722.

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. *Advances in neural information processing systems*, 23, 2010. URL https://proceedings.neurips.cc/paper/2010/file/49182f81e6a13cf5eaa496d51fea6406-Paper.pdf.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011. URL https://papers.nips.cc/paper/2011/hash/ebea851d20058b56f2f4581f190e290f-Abstract.html.

OpenAI. New embedding models and api updates, 2024. URL https://openai.com/blog/new-embedding-models-and-api-updates.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. URL https://arxiv.org/abs/1803.00567.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning, 2021. URL https://arxiv.org/abs/2104.08894.

Sidharth V. Puram, Itay Tirosh, Anuraag S. Parikh, Anoop P. Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L. Luo, Edmund A. Mroz, Kevin S. Emerick, Daniel G. Deschler, Mark A. Varvares, Rajan Mylvaganam, Orit Rozenblatt-Rosen, James W. Rocco, William C. Faquin, Daniel T. Lin, Aviv Regev, and Bradley E. Bernstein. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624.e24, 2017. doi: 10.1016/j.cell.2017.10.044.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Michael Robinson, Sourya Dey, and Shauna Sweet. The structure of the token space for large language models, 2024. URL https://arxiv.org/abs/2410.08993.

Michael Robinson, Sourya Dey, and Tony Chiang. Token embeddings violate the manifold hypothesis, 2025. URL https://arxiv.org/abs/2504.01002.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Haitz Sáez de Ocáriz Borde and Michael Bronstein. Mathematical foundations of geometric deep learning, 2025. URL https://arxiv.org/abs/2508.02723.

Hang Shao, Abhishek Kumar, and P. Fletcher. The riemannian geometry of deep generative models. 11 2017. doi: 10.48550/arXiv.1711.08014.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. URL https://arxiv.org/abs/2011.13456.

Jan Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Your diffusion model secretly knows the dimension of the data manifold. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31326–31338, 2022. URL https://arxiv.org/abs/2212.12611.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319. URL https://www.science.org/doi/10.1126/science.290.5500.2319.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1711.01558. arXiv:1711.01558.

René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011. URL https://doi.org/10.1109/MSP.2010.939739.

Julius Von Rohrscheidt and Bastian Rieck. Topological singularity detection at multiple scales. In *International Conference on Machine Learning*, pp. 35175–35197. PMLR, 2023. URL https://arxiv.org/abs/2210.00069.

Yi Wang and Zhiren Wang. CW complex hypothesis for image data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51596–51613. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24bs.html.

Nick Whiteley, Annie Gray, and Patrick Rubin-Delanchy. Statistical exploration of the manifold hypothesis. *arXiv preprint arXiv:2208.11665*, 2022. URL https://arxiv.org/abs/2208.11665.

Hassler Whitney. The self-intersections of a smooth n-manifold in $2n$-space. *Annals of Mathematics*, 45(2):220–246, 1944.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Hanlin Yu, Søren Hauberg, Marcelo Hartmann, Arto Klami, and Georgios Arvanitidis. Learning geometry and topology via multi-chart flows. *arXiv preprint arXiv:2505.24665*, 2025.

A Extended related work and discussion

A.1 Extended related work

Extrinsic and intrinsic formulations. Fefferman et al. (2013) distinguish an *extrinsic* from an *intrinsic* formulation of the M.H. The extrinsic formulation is the standard one: the data are assumed to concentrate near a d-dimensional submanifold embedded in a high-dimensional ambient space $\mathbb{R}^D$ with $d \ll D$, so manifoldness is judged relative to the ambient coordinates. The intrinsic formulation discards the ambient embedding and treats the data as a finite metric space, asking whether its pairwise distances can be realised, up to low distortion, by a d-dimensional manifold. See Figure 10 for further detail.

Topological data analysis. A topological alternative to GMI is TARDIS (Von Rohrscheidt & Rieck, 2023), which computes persistent local homology and produces a Euclidicity value. Euclidicity assesses locally varying intrinsic dimension, which is similar to s_{const}. We did not adopt it, or related persistent homology diagnostics, because they do not meet the scalability requirements of the diagnostic we set out to build. Persistent homology remains complementary where sample size and ambient dimension permit.

A.2 Extended discussion

Manifoldness to gauge the validity of using existing techniques in data and embedding spaces. Most of our experiments suggest data and embedding spaces may not satisfy the M.H., but some subsets score high enough that a noisy manifold with bounded distortion is plausible. For these spaces, GMI can act as an indicator or precondition for downstream methods. Further comparisons can be run against generative-quality metrics such as FID (Heusel et al., 2017) or quality of UMAP (McInnes et al., 2018; Jeon et al., 2025) outputs. Representation space manifoldness can flag neighbourhoods with low τ and high d_{LPCA} for targeted repair; and probing Riemannian and hyperbolic latent spaces (Gemici et al., 2016; Jaquier et al., 2024; Farghly et al., 2025) remain open.

B MANIFOLD DEFINITIONS

Definition 1 treats $\mathcal{M}_d$ as a smooth embedded submanifold of $\mathbb{R}^D$ with positive reach. We record the standard differential-topology definitions underlying that claim (Sáez de Ocáriz Borde & Bronstein, 2025; Federer, 1959; Fefferman et al., 2013). GMI does not construct charts; it tests necessary geometric signatures of this object on finite samples (Section 4).

Definition 4 (Homeomorphism)**.** A map $\varphi : X \to Y$ between topological spaces is a homeomorphism if it is bijective, continuous, and has a continuous inverse.

Definition 5 (Topological manifold)**.** An n-dimensional topological manifold is a Hausdorff, second-countable topological space $\mathcal{M}$ such that for every $p \in \mathcal{M}$ there exist an open neighbourhood $U \subseteq \mathcal{M}$ and a homeomorphism $\varphi : U \to \varphi(U) \subseteq \mathbb{R}^n$.

Definition 6 (Coordinate chart and atlas)**.** A coordinate chart on an n-dimensional topological manifold $\mathcal{M}$ is a pair (U, φ) with $U \subseteq \mathcal{M}$ open and $\varphi : U \to \hat{U} \subseteq \mathbb{R}^n$ a homeomorphism onto an open set. An atlas is a cover of $\mathcal{M}$ by charts $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in A}$ such that whenever $W = U_\alpha \cap U_\beta \neq \emptyset$, the transition map $\varphi_\beta \circ \varphi_\alpha^{-1} : \varphi_\alpha(W) \to \varphi_\beta(W)$ is a homeomorphism between open subsets of $\mathbb{R}^n$.

Definition 7 (Smooth manifold)**.** A smooth (or C^∞) manifold is a topological manifold equipped with an atlas whose transition maps are smooth (infinitely differentiable) as maps between open subsets of Euclidean space. A C^k manifold is defined analogously with C^k transitions; $C^{1,1}$ requires continuously differentiable transitions with Lipschitz derivatives.

Definition 8 (Embedded submanifold of $\mathbb{R}^D$)**.** A subset $\mathcal{M} \subset \mathbb{R}^D$ is a d-dimensional C^r embedded submanifold if for every $p \in \mathcal{M}$ there exist an open neighbourhood $U \subset \mathbb{R}^D$ of p and a C^r diffeomorphism $\phi : U \to \phi(U) \subseteq \mathbb{R}^D$ such that

$$\phi(U \cap \mathcal{M}) = \phi(U) \cap \left(\mathbb{R}^d \times \{0\}^{D-d}\right). \tag{5}$$

Under Definition 1 we require at least $C^{1,1}$ regularity. Positive Federer reach (Definition 3) is compatible with this class and forbids cusps and self-intersections (Federer, 1959; Fefferman et al., 2013).

The tangent space $T_p\mathcal{M}$ of Section 3 is then the d-dimensional subspace of $\mathbb{R}^D$ obtained by differentiating curves on $\mathcal{M}$ through p, equivalently the preimage under $D\phi_p$ of $\mathbb{R}^d \times \{0\}$. Reach and local PCA operate in this ambient embedding.

Definition 9 (Ambient dimension)**.** The ambient dimension D is the dimension of the Euclidean space containing the observed vectors. For raw RGB images of size 256×256, $D = 256 \cdot 256 \cdot 3$. For an encoder representation, D is the size of the embedding vector.

Definition 10 (Intrinsic dimension)**.** Let $\mathcal{M} \subset \mathbb{R}^D$ be a d-dimensional manifold. The intrinsic dimension of $\mathcal{M}$ is d, the number of local coordinates required to describe a neighbourhood of any point in $\mathcal{M}$. Equivalently, every point $p \in \mathcal{M}$ has a chart to an open subset of $\mathbb{R}^d$.

Definition 11 (Local intrinsic dimension)**.** The local intrinsic dimension at a point p is the dimension inferred from a neighbourhood of p. On a single smooth d-manifold it equals d everywhere, whereas unions, branches, and stratified supports can vary spatially (Houle, 2017; Hein & Audibert, 2005).

C ROUTING GMI TO DATA AND REPRESENTATION SPACES

Figure 10 summarises the routes described in Section 4. Throughout, we use *embedding space*, *feature space*, *latent space*, and *representation space* interchangeably for the learned vector input representations of a deep learning model.

D SYNTHETIC VALIDATION

Figure 11 shows representative synthetic shapes used throughout validation; component scores are in Table 1. Figures 12–14 report high-dimensional asymptotics and per-centre diagnostics. Appendix D.1 extends the MNIST image-transformation control of the main text to CIFAR-10 and Fashion-MNIST.

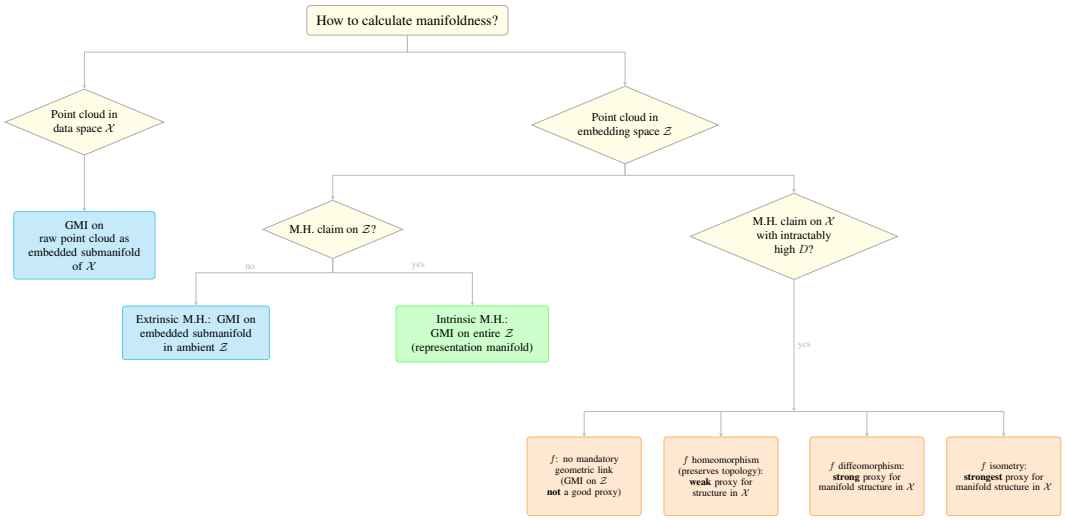


Figure 10: Routes to GMI. Yellow: decision points. Blue: GMI for embedded submanifolds in data space $\mathcal{X}$ or in ambient $\mathcal{Z}$. Green: intrinsic claim on the full support in $\mathcal{Z}$. Orange: proxy evidence for structure in high-dimensional $\mathcal{X}$ via $\mathcal{Z}$, weighted by encoder regularity (homeomorphism $\prec$ diffeomorphism $\prec$ isometry).

Table 1: Component scores for synthetic shapes in $\mathbb{R}^{128}$, $N = 10{,}000$, 3 seeds. Rows grouped: clean manifolds, manifolds with boundary, nearly manifold, non-manifold.

Shape	s_{const}	s_{reach}	GMI
Manifold			
Circle	1.000 ± 0.000	0.992 ± 0.002	0.996 ± 0.001
Sphere S2	1.000 ± 0.000	0.977 ± 0.001	0.989 ± 0.001
Sphere S16	0.937 ± 0.006	0.784 ± 0.004	0.857 ± 0.004
Torus 16	0.927 ± 0.006	0.752 ± 0.001	0.835 ± 0.003
Torus 32	0.940 ± 0.002	0.800 ± 0.001	0.867 ± 0.001
Disjoint S2	1.000 ± 0.000	0.965 ± 0.003	0.982 ± 0.001
Klein Bottle R4	1.000 ± 0.000	0.902 ± 0.011	0.950 ± 0.006
Manifold w/ Boundary			
Swiss Roll	1.000 ± 0.000	0.939 ± 0.016	0.969 ± 0.008
Hollow Cube C16	0.840 ± 0.003	0.771 ± 0.000	0.805 ± 0.001
Hollow Cube C32	0.861 ± 0.002	0.785 ± 0.001	0.822 ± 0.001
Fibre Bundle	0.906 ± 0.006	0.643 ± 0.058	0.763 ± 0.037
Nearly Manifold			
S2+noise(0.02)	0.664 ± 0.003	0.667 ± 0.024	0.665 ± 0.012
S2+noise(0.15)	0.485 ± 0.003	0.633 ± 0.011	0.554 ± 0.006
Cone (Singular)	0.987 ± 0.019	0.657 ± 0.184	0.797 ± 0.127
Lollipop	0.804 ± 0.008	0.000 ± 0.000	0.000 ± 0.000
Pinch Torus 2	0.892 ± 0.005	0.278 ± 0.060	0.495 ± 0.057
Intersecting S2	0.693 ± 0.001	0.725 ± 0.037	0.709 ± 0.018
Klein Bottle R3	0.737 ± 0.006	0.676 ± 0.058	0.705 ± 0.032
Not Manifold			
Spheres (TopoAE)	0.136 ± 0.063	0.544 ± 0.002	0.266 ± 0.060
Pure Noise	0.062 ± 0.002	0.639 ± 0.012	0.200 ± 0.005

D.1 Image-transformation manifolds for CIFAR-10 and Fashion-MNIST

We repeat the main-text MNIST rotation control (Figure 4) on a CIFAR-10 airplane ($D = 3{,}072$) and a Fashion-MNIST trouser ($D = 784$) under an identical protocol: one randomly drawn image is rotated in-plane through $N = 360$ views at 1° steps with a circular mask and cubic interpolation, the displayed base is the median-GMI image of a pool of 32 random bases from the class, equally spaced clouds take constant-angle views of that orbit, and random subsets draw n of the 360 views uniformly without replacement over 3 seeds. Both datasets reproduce the MNIST behaviour.

Table 2: Component scores for $S^d \subset \mathbb{R}^D$, $d \in \{256, 512, 768\}$, $D = 4096$, 5 seeds. At $N = 10{,}000$ all three spheres converge to GMI $\in [0.89, 0.91]$.

Shape	N	s_{const}	s_{reach}	GMI
S^{256}	1,000	0.833 ± 0.013	0.840 ± 0.003	0.836 ± 0.008
S^{512}	1,000	0.966 ± 0.009	0.861 ± 0.004	0.912 ± 0.004
S^{768}	1,000	1.000 ± 0.000	0.873 ± 0.002	0.934 ± 0.001
S^{256}	10,000	0.930 ± 0.002	0.856 ± 0.001	0.892 ± 0.001
S^{512}	10,000	0.923 ± 0.002	0.879 ± 0.001	0.901 ± 0.001
S^{768}	10,000	0.920 ± 0.002	0.889 ± 0.001	0.904 ± 0.001

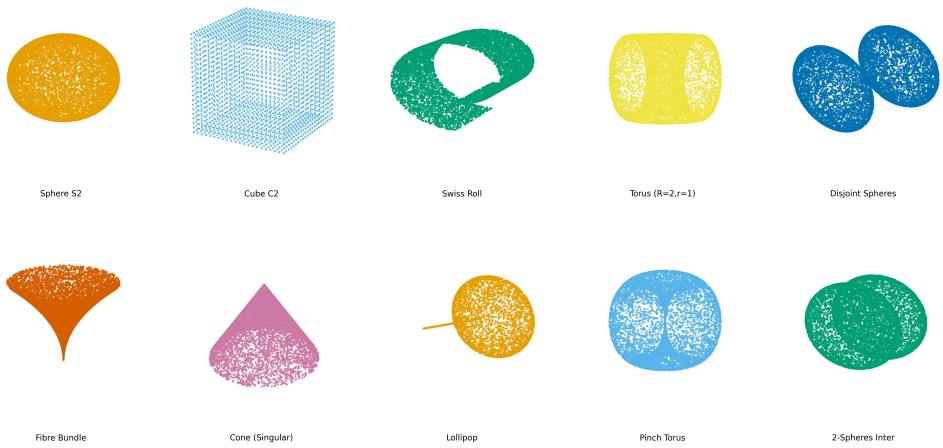
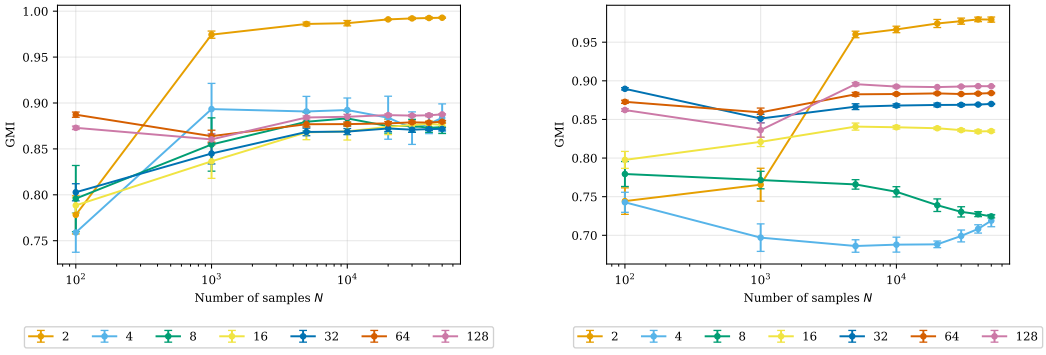


Figure 11: Representative synthetic shapes: clean manifolds (top), manifolds with boundary, nearly-manifold singularities, and non-manifold controls (bottom). Full catalogue in Section D.



(a) Sphere S^d

(b) Torus T^d

Figure 12: GMI vs. N for $d \in \{2, 4, 8, 16, 32, 64, 128\}$ in $\mathbb{R}^{512}$ (10 seeds).

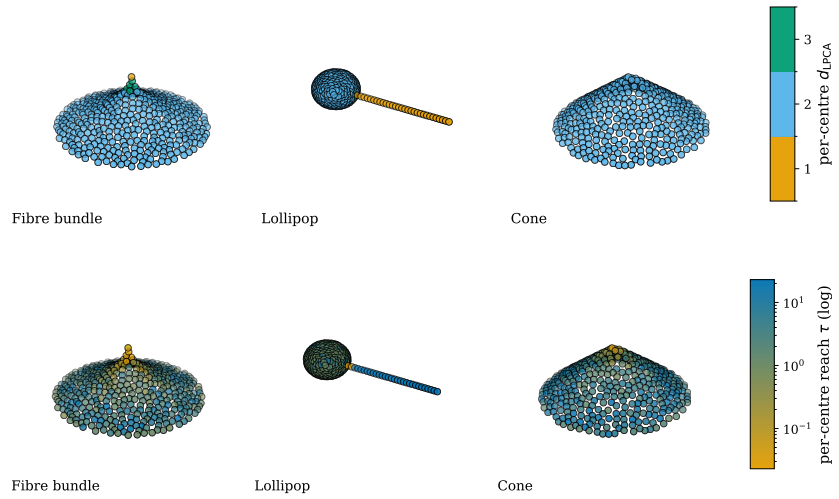


Figure 13: Per-centre d_{LPCA} (top) and reach τ (bottom) on fibre bundle, lollipop, and cone ($N = 10{,}000$, $D = 128$). Figure 15 shows that low-reach FPS centres near the cone apex support the high s_{reach} variance reported above; noisy-sphere diagnostics are in Figure 14.

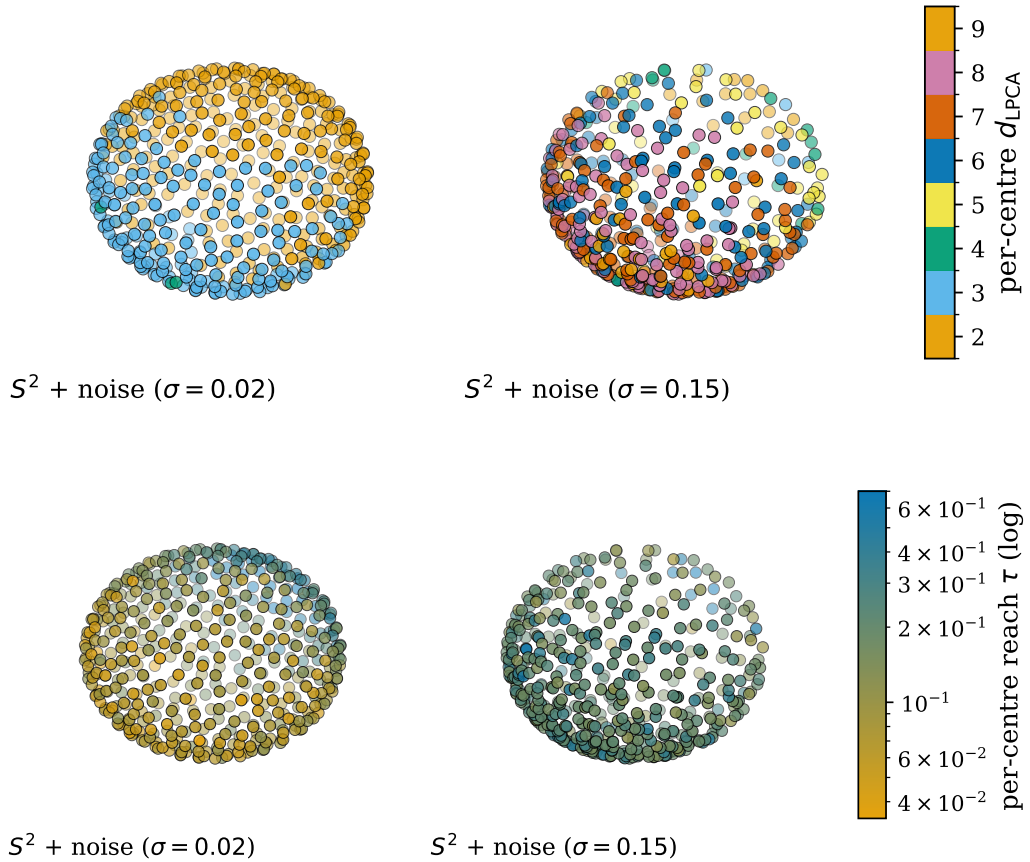


Figure 14: Per-centre d_{LPCA} (top) and reach τ (bottom) on noisy spheres ($N = 10{,}000$, $D = 128$).

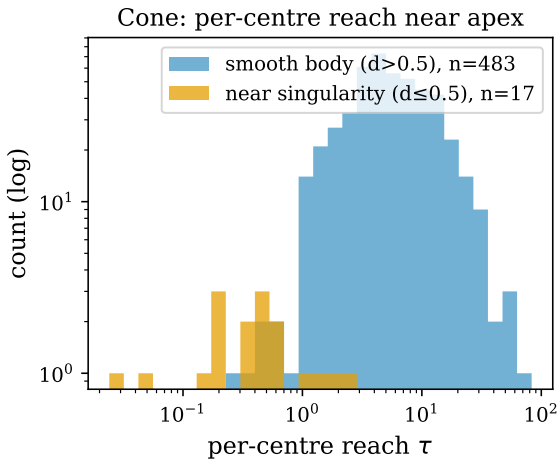
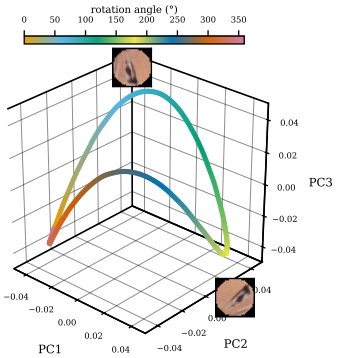
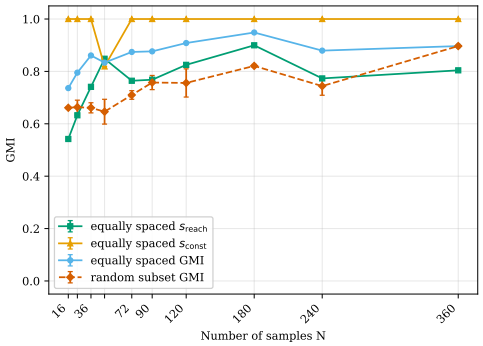


Figure 15: Distribution of per-centre reach τ on FPS centres near the cone apex (orange) versus the smooth body (blue), $N = 10{,}000$, $D = 128$, seed 0.

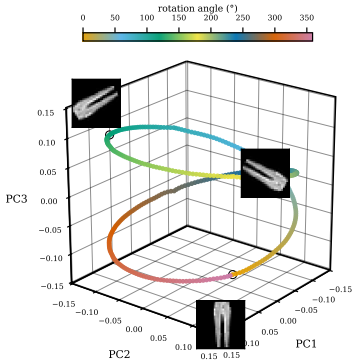


(a) PCA-3D of the $N = 360$ in-plane rotation orbit.

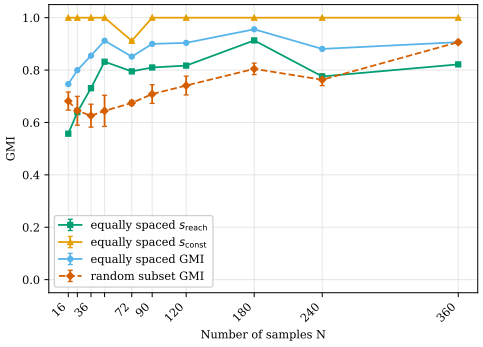


(b) Equally spaced s_{reach}, s_{const}, and GMI, and random subset GMI (mean $\pm$ 1 s.d. over 3 seeds). At $n = 360$, GMI coincides.

Figure 16: Results of the synthetic rotation experiment for one randomly drawn CIFAR-10 airplane.



(a) PCA-3D of the $N = 360$ in-plane rotation orbit.



(b) Equally spaced s_{reach}, s_{const}, and GMI, and random subset GMI (mean $\pm$ 1 s.d. over 3 seeds). At $n = 360$, GMI coincides.

Figure 17: Results of the synthetic rotation experiment for one randomly drawn Fashion-MNIST trouser.

Table 3: Real-world point clouds: sample size N, ambient dimension D, and published intrinsic-dimension estimates where available.

Dataset / source	N	D	Published ID estimate
Images			
MNIST (Deng, 2012)	70,000	784	7–11 (Pope et al., 2021)
Fashion-MNIST (Xiao et al., 2017)	70,000	784	13–21 (Pope et al., 2021)
CIFAR-10 (Krizhevsky et al., 2009)	60,000	3,072	13–21 (Pope et al., 2021)
SVHN (Netzer et al., 2011)	99,289	3,072	MLE 9–14 (Pope et al., 2021)
ImageNet birds/dogs (SD-VAE) (Deng et al., 2009; Rombach et al., 2022)	76k–100k	16,384	26–43 (Pope et al., 2021)
Token embeddings			
GPT-2 / BERT / T5 (Radford et al., 2019; Devlin et al., 2019; Raffel et al., 2020)	30k–50k	768	
Mistral / Pythia / Llemma (Jiang et al., 2023; Biderman et al., 2023; Azerbayev et al., 2023)	32k–50k	4096	
OpenAI 3 (OpenAI, 2024)	50,257	3,072	
scRNA-seq			
Embryoid body (Moon et al., 2019)	31,161	33,694	
Pancreatic endocrinogenesis (Bastidas-Ponce et al., 2019)	36,351	17,327	
Baron2 (Baron et al., 2016)	8,569	20,125	
Puram HNSCC (Puram et al., 2017)	5,902	23,686	
Darmanis (brain) (Darmanis et al., 2015)	466	22,088	

E Datasets and hyperparameters

Table 3 lists the real-world point clouds; Table 4 lists constructor defaults and Table 5 the fixed internal constants.

E.1 Encoder families

All MNIST encoders share the same convolutional encoder/decoder backbone and differ only in the training objective; SD-VAE is used pretrained. The families are:

- **AE**: deterministic autoencoder trained with reconstruction loss alone, with no constraint on the aggregate latent distribution (Kramer, 1991).
- **VAE**: adds a KL divergence between the per-point posterior and an isotropic Gaussian prior (Kingma & Welling, 2013).
- **TopoAE**: adds a persistent-homology loss that matches the input and latent Vietoris–Rips filtrations, explicitly penalising topological distortion (Moor et al., 2020).
- **WAE-MMD**: penalises the *aggregate* posterior against the prior with a maximum mean discrepancy term rather than a per-point KL (Tolstikhin et al., 2018).
- **WAE-GAN**: same Wasserstein objective with the divergence estimated by an adversarial discriminator in latent space (Tolstikhin et al., 2018).
- **SD-VAE**: the pretrained Stable Diffusion KL-regularised autoencoder, applied without fine-tuning (Rombach et al., 2022).

TopoAE is the only objective that regularises geometry directly; VAE, WAE-MMD, and WAE-GAN regularise the latent *distribution*, which is what orders their manifoldness scores in Section 5.1.

Table 4: Constructor hyperparameters (`ManifoldDetector`).

Parameter	Default	Role
`n_fps_max`	2000	Cap on FPS centres
`k_patch_min/max`	10 / 2000	Patch size bounds
`k_tangent_min`	5	Tangent neighbourhood floor
`k_reach_nn`	10	Reach averaging neighbours
`id_method`	`lpca`	Per-patch ID estimator

Table 5: Fixed internal constants.

Constant	Value
`alphaFO`	0.175
`col_var_eps`	10^{-8}
`eps_norm_pct`	50
`adaptive_max_divisor`	20
`k_patch_divisor`	4
`n_small`	200 (use all points as net)
`flat_thresh`	10^{-4}
`svd_batch`	256
`tangent_d_max`	256
`lowrank_d_thresh` / `lowrank_q`	256 / 32

F Determinism

The GMI calculation is mostly deterministic: every pipeline stage is an exact function of the point cloud and fixed hyperparameters except randomised truncated SVD for tangent estimation when $D_{\text{active}} > 256$ and $k_{\tan} > 32$. That stochastic step affects only s_{reach}; s_{const} is fully deterministic. Principal-angle error from `svd_lowrank` (Halko et al., 2011) is negligible relative to reported accuracy; variance across seeds therefore comes primarily from random subsamples of X.

G Implementation

Implementation follows `manifold_detector.py` (`ManifoldDetector._fit`). After active-column filtering, a single k-NN query on X_{active} supplies $\varepsilon_{\text{density}}$, median $\varepsilon_{\text{norm}}$, and global TwoNN d_{coarse}. The same FPS net defines patch centres for s_{const} and reach anchors for s_{reach}.

KNN backend. Exact k-NN under squared Euclidean distance on float32 X_{active}: GPU chunked torch matmul when CUDA is available, else FAISS `IndexFlatL2`.

Patch size. All patches share one size. The implementation clips a size proposal to a lower bound $k_{\text{lo}} = \max(k_{\text{patch,min}}, 2D_{\text{active}})$, which keeps the patch SVD overdetermined, and an upper bound $k_{\text{hi}} = \min(k_{\text{patch,max}}, \lfloor N/\nu_{\text{patch}} \rfloor)$, which keeps a patch a bounded fraction of the cloud; when $k_{\text{lo}} > k_{\text{hi}}$ the lower bound is dropped and $k_{\text{patch}} = k_{\text{hi}}$. In every high-dimensional setting we report, k_{patch} is determined by D_{active} and N alone: for example $k_{\text{patch}} = 2D_{\text{active}} = 512$ on PCA-256 clouds with $N = 10^4$, and $k_{\text{patch}} = k_{\text{hi}}$ whenever $2D_{\text{active}} > k_{\text{hi}}$. Patches in that regime are a fixed fraction of the cloud rather than asymptotically local, and FO counts are capped at $\min(k_{\text{patch}}, D_{\text{active}}) - 1$. Table 10 reports sensitivity to k_{patch}.

Alternative `id_method`. Reach always uses $d_{\text{reach}} = \min\big(d_{2\text{NN}}, \text{round}(d_{\text{ref}}^{\text{LPCA}})\big)$, independent of `id_method`; optional per-patch estimators are:

`id_method`	Per-patch $d_{\text{local},i}$	Notes
`lpca` (default)	FO count on patch singular values	production
`twonn`	TwoNN on patch members' 1-/2-NN	fallback $d_{2\text{NN}}$
`mle`	Levina–Bickel MLE from centre to neighbours	fallback $d_{2\text{NN}}$

Tangent SVD. For $D_{\text{active}} > 256$: Gram eigendecomposition when $k_{\tan} \leq 32$, else truncated SVD. Patch FO LPCA uses batched `torch.linalg.svdvals` with batch size `svd_batch`$=256$.

Table 6: Wall-clock seconds for one GMI fit on ImageNet SD-VAE PCA-D_{active} clouds (NVIDIA L4), mean ± s.d. over 2 supercategories × 5 seeds; the $N = 10^5$ row is dogs only (5 seeds).

N	$D_{\text{active}} = 256$	512	1024
1,000	$1.3_{\pm 0.7}$	$1.4_{\pm 0.7}$	$1.6_{\pm 0.7}$
2,000	$2.4_{\pm 1.3}$	$4.8_{\pm 2.0}$	$5.9_{\pm 2.2}$
5,000	$5.9_{\pm 3.2}$	$11.3_{\pm 4.6}$	$33.5_{\pm 9.9}$
10,000	$12.4_{\pm 6.5}$	$22.2_{\pm 9.5}$	$58.6_{\pm 20.2}$
25,000	$30.7_{\pm 16.4}$	$53.7_{\pm 22.0}$	$150.8_{\pm 41.0}$
50,000	$49.1_{\pm 26.8}$	$90.7_{\pm 32.9}$	$262.1_{\pm 74.7}$
100,000	$66.1_{\pm 24.0}$	$105.7_{\pm 26.7}$	$283.4_{\pm 62.5}$

Complexity. Write $M = n_{\text{centres}} \le n_{\text{fps,max}}$ for the number of FPS centres/anchors. The global exact k-NN query on all of X_{active} (for $\varepsilon_{\text{density}}$, $\varepsilon_{\text{norm}}$, and d_{coarse}) costs $O(N^2 \cdot D_{\text{active}})$; FPS and the per-centre k-NN queries cost $O(N \cdot M \cdot D_{\text{active}})$; patch SVD $O(M \cdot k_{\text{patch}}^2 \cdot D_{\text{active}})$; tangent SVD $O(M \cdot k_{\text{tan}}^2 \cdot D_{\text{active}})$; the reach projection $O(M \cdot k_{\text{reach}} \cdot d_{\text{reach}} \cdot D_{\text{active}})$. Intrinsic dimension enters only through $k_{\text{tan}} = \min(\max(\hat{d}, k_{\text{tan,min}}), d_{\max})$ and d_{reach}, both capped by `tangent_d_max`$= 256$; $M \le n_{\text{fps,max}} = 2000$ and $k_{\text{patch}} \le k_{\text{patch,max}} = 2000$. Once these caps bind, N appears only in the k-NN and FPS terms, so a fit is quadratic in N at fixed D_{active}, dominated by the global k-NN query; all other stages are linear in N.

Hardware. Experiments used an NVIDIA L4 GPU on Python 3.12.3 (dual-socket Intel Xeon Silver 4514Y, 64 logical CPUs). Wall-clock per ID estimator call is in Table 9.

Runtime. Table 6 reports end-to-end wall-clock for a single `ManifoldDetector.fit` call on the ImageNet SD-VAE PCA clouds of Figure 7b, excluding data loading and PCA. The largest cloud we score, $N = 10^5$ at $D_{\text{active}} = 1024$, takes 283 ± 62 s on one L4. On the scRNA PCA clouds of Figure 25 ($N = 2000$), a fit costs 0.06 s at $D' = 32$ and 9.9 s at $D' = 1024$, against 14–16 s for the UMAP embedding of the same cloud.

H Properties of the Geometric Manifoldness Index

We record elementary properties of the GMI map of equation 4, whose components are defined by equation 2 above and equation 7 below. These are statements about the diagnostic, not statistical consistency of the reach diagnostic (Section 4.2). Per-patch FO LPCA counts significant singular values via

$$d_{\text{local},i} = \#\{j : s_{i,j} > \alpha_{\text{FO}}\, s_{i,1}\}, \qquad d_{\text{local},i} \ge 1. \tag{6}$$

Consistency of these counts is scored by equation 2, whose denominator uses $d_{\text{ref}} = \text{mean}(\boldsymbol{d})$. The denominator converts the absolute spread of the counts into a spread relative to the dimension level: FO counts are integers, so a one-count disagreement between patches is a large relative inconsistency at $d_{\text{ref}} = 2$ and a small one at $d_{\text{ref}} = 16$. This means an unnormalised $\text{std}(\boldsymbol{d})$ would penalise high-dimensional manifolds for quantisation noise alone. Section I.2 recomputes s_{const} from the same per-patch counts under the normalisers 1, $\log(1 + d_{\text{ref}})$, $\sqrt{d_{\text{ref}}}$, and d_{ref} (Figure 19): linear d_{ref} over-normalises and reorders shapes that the others agree on, no normalisation shrinks the margin between a clean S^8 and the singular lollipop, and we keep $\log(1 + d_{\text{ref}})$ as the more conservative of the two sub-linear choices that preserve the ordering.

For reach, displacements $v_{ij} = x_j - s_i$ use perpendicular components $v_{ij}^{\perp} = v_{ij} - \hat{T}_i^{\top}(\hat{T}_i v_{ij})$ and Federer ratios $\rho_{ij} = \|v_{ij}\|^2/(2\|v_{ij}^{\perp}\|)$ when $\|v_{ij}^{\perp}\| \ge$ `flat_thresh` $\|v_{ij}\|$; otherwise the pair is treated as flat. Per anchor, $\hat{\tau}_i$ is the mean of the finite ratios at s_i, or $+\infty$ if all pairs are flat; $\hat{\tau} = \min_i \hat{\tau}_i$ over finite $\hat{\tau}_i$. With $d_{\text{scale}} = \sqrt{d_{\text{reach}}}$ and $\varepsilon_{\text{norm}}$ the median first-neighbour distance,

$$s_{\text{reach}} = \begin{cases} 1 & \text{if } \hat{\tau} = +\infty \\ \dfrac{\hat{\tau}\, d_{\text{scale}}}{\hat{\tau}\, d_{\text{scale}} + \varepsilon_{\text{norm}}} & \text{otherwise} \end{cases} \tag{7}$$

The case $d_{\text{reach}} \ge D_{\text{active}}$ is outside the domain of equation 7: Federer ratios require a nonzero perpendicular component, which vanishes when the estimated tangent span equals the active ambient

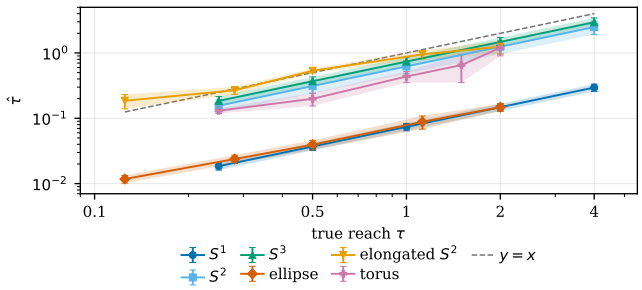


Figure 18: Calibration of the mean-then-min proxy $\hat{\tau}$ against analytic Federer reach on families with closed-form τ: S^1, S^2, and S^3 of radius $\tau \in \{0.25, 0.5, 1, 2, 4\}$; a planar ellipse with semi-axes $(2, b)$ and an elongated S^2 with semi-axes $(2, b, b)$, both of reach $\tau = b^2/2$ for $b \in \{0.5, 0.75, 1, 1.5, 2\}$; and a torus with $R = 5$ and tube radius $\tau = r \in \{0.25, 0.5, 1, 1.5, 2\}$. $N = 1{,}500$ points per cloud, 3 seeds; markers are seed means, error bars and bands are ± 1 s.d., both axes are logarithmic, and the dashed line is $y = x$.

space. That case is undefined. The Federer ratio is computed with the ambient Euclidean metric of the supplied coordinates (Arvanitidis et al., 2018; Hauberg, 2022), so s_{reach} is extrinsic and two encoders that realise the same manifold in different coordinates need not receive the same score. On clouds where the diagnostic is defined, $s_{\text{reach}} = 1$ if and only if $\hat{\tau} = +\infty$.

Proposition 1 (Range of the score)**.** On any finite point cloud where the reach diagnostic is defined, $s_{\text{const}} \in (0, 1]$, $s_{\text{reach}} \in [0, 1]$, and hence $\text{GMI} = (s_{\text{const}}\, s_{\text{reach}})^{1/2} \in [0, 1]$. Moreover $s_{\text{const}} = 1 \iff \text{std}(\boldsymbol{d}) = 0$, $s_{\text{reach}} = 1 \iff \hat{\tau} = +\infty$, and $s_{\text{reach}} = 0 \iff \hat{\tau} = 0$.

Proof. The dimension vector $\boldsymbol{d}$ is finite, so $\text{std}(\boldsymbol{d}) \in [0, \infty)$ and the denominator in equation 2 is positive. Hence $s_{\text{const}} = \exp(-(\cdot)) \in (0, 1]$, with equality at 1 iff $\text{std}(\boldsymbol{d}) = 0$. Restrict to clouds on which the reach diagnostic is defined. For s_{reach}, the map $x \mapsto x/(x + \varepsilon_{\text{norm}})$ with $\varepsilon_{\text{norm}} > 0$ takes values in $[0, 1)$, attaining 0 iff $\hat{\tau} = 0$; the $\hat{\tau} = +\infty$ branch contributes 1, and only that branch attains 1. □

Proposition 2 (Invariance under similarity transformations after active-column filtering)**.** Let $X_{\text{active}} \subset \mathbb{R}^{D_{\text{active}}}$ be a finite point cloud after the active-column filter of Section 4, and let $\Phi(x) = cRx + t$ with $c > 0$, $R \in O(D_{\text{active}})$, and $t \in \mathbb{R}^{D_{\text{active}}}$. If the reach diagnostic is defined on X_{active}, then $\text{GMI}(\Phi(X_{\text{active}})) = \text{GMI}(X_{\text{active}})$.

Proof. The map equation 4 is applied to X_{active}. Pairwise distances and $\varepsilon_{\text{density}}$ both scale by c under Φ, so the FPS net, d_{coarse}, and k_{patch}, which depend only on their ratios, are invariant. FO counts (equation 6) depend only on singular-value ratios of centred patches, likewise invariant. Tangent subspaces rotate with R; Federer ratios and $\varepsilon_{\text{norm}}$ both scale by c and cancel in equation 7. Thus s_{const}, s_{reach}, and GMI are unchanged. □

I Ablation Studies

We ablate per-patch ID estimators (Tables 7–8), FO significance threshold (Figure 21), reach neighbourhood size (Figure 22), patch size (Table 10), and the Federer-ratio aggregator (Figure 20).

Figure 18 compares our method of mean-then-min $\hat{\tau}$ with analytic Federer reach on shapes where τ varies and is known in closed form.

I.1 Aggregating Federer ratios

Figure 20 compares the mean-then-min aggregator of Section 4.2 against the literal pairwise minimum of equation 1. The pairwise minimum is decided by a single anchor–neighbour pair: it collapses $\hat{\tau}$ considerably on shapes with positive reach, and its seed variance swamps the gap between the clean torus and the singular cone, which mean-then-min preserves.

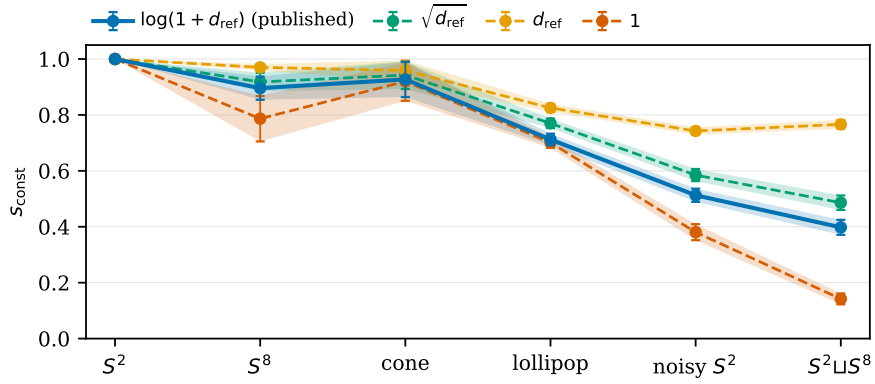


Figure 19: Values of s_{const} for different normalisation functions. All four are recomputed from the same per-patch FO counts ($N = 1{,}200$, $D = 16$, mean $\pm$ 1 s.d. over 3 seeds).

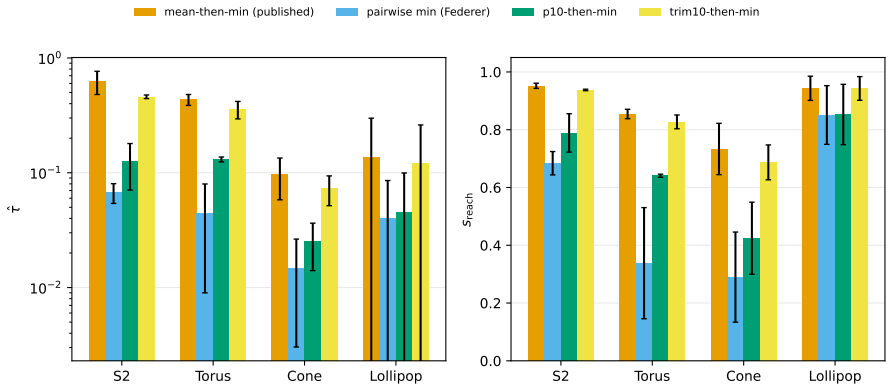


Figure 20: Aggregator ablation for $\hat{\tau}$ (log scale) and s_{reach} ($N = 1{,}500$, $D = 16$, mean $\pm$ 1 s.d. over 3 seeds). All four aggregators are applied to the same finite Federer-ratio matrix and share the normalisation constants of the mean-then-min fit, so only the aggregation step differs. S^2 and the torus have $\tau = 1$; the cone and lollipop have $\tau = 0$.

I.2 Normalisation inside s_{const}

Figure 19 recomputes s_{const} from a single fit per (shape, seed) under four normalisation functions of $\text{std}(\boldsymbol{d})$. The shapes span clean manifolds at low and high d (S^2, S^8), a singular cone, a lollipop, a noisy S^2, and a mixed-dimension union $S^2 \sqcup S^8$; the last two are the cases s_{const} must penalise.

All four normalisers separate clean manifolds from the noisy and mixed-dimension clouds, so our conclusions do not depend highly on this choice. The sub-linear normalisers $\log(1 + d_{\text{ref}})$ and $\sqrt{d_{\text{ref}}}$ preserve the same ordering of shapes. Linear d_{ref} over-normalises: it compresses the range, lifts $S^2 \sqcup S^8$ to within 0.2 of the clean shapes, and reorders pairs that the other three agree on (it ranks $S^2 \sqcup S^8$ above noisy S^2, and S^8 above the cone). No normalisation (1) under-normalises: it roughly halves the margin between clean S^8 and the singular lollipop, confounding high intrinsic dimension with inconsistency. We keep $\log(1 + d_{\text{ref}})$, the more conservative of the two remaining choices.

I.3 Intrinsic dimension estimators

We summarise the per-patch ID estimators considered for s_{const} (Section 4.1).

Maximum likelihood estimators. Prior work on the M.H. often uses the Levina–Bickel MLE, which models local neighbour distances under a Poisson approximation (Levina & Bickel, 2004):

for k neighbours at distances $r_1, \ldots, r_k$, $\hat{d}_k(x) = \left(\frac{1}{k-1}\sum_{i=1}^{k-1}\log(r_k/r_i)\right)^{-1}$. The MacKay–Ghahramani correction is also commonly used in image-manifold studies (Brown et al., 2022; MacKay & Ghahramani, 2005; Pope et al., 2021). These provide high accuracy in low-curvature regions but suffer significant bias near manifold boundaries or when the local distribution is non-uniform (Amsaleg et al., 2015; Hein & Audibert, 2005; Ma et al., 2018), and remain extremely sensitive to k.

Geometric estimators. TwoNN uses only the ratio $\mu = r_2/r_1$ between the first and second nearest neighbours, with d following from $-\log(1 - P(\mu)) = d \log \mu$ (Facco et al., 2017). This reduces dependence on a manually chosen k but increases pointwise variance. FO LPCA is a geometric estimator rather than an MLE: it estimates the rank of a local tangent patch through singular values (equation 6) rather than fitting a distance distribution (Fukunaga & Olsen, 1971). caPCA calibrates the same singular values against a quadratic embedding instead of a flat ball, accounting for curvature (Gilbert & O'Neill, 2024).

Flow-based methods. Recent estimators use deep generative models to handle high-dimensional spaces like images. Stanczuk et al. (2022) and Horvat & Pfister (2024) use diffusion models, relating the score function to the intrinsic dimension; NF-ID uses normalising flows (Horvat & Pfister, 2022); and FLIPD gives $O(1)$ inference per point once trained (Kamkari et al., 2024), with correctness proven by Leung et al. (2025). These involve a high training cost but fast inference, and their accuracy depends heavily on the underlying density model. FLIPD is designed for pixel-space inference where $D = 262{,}144$, whereas GMI must score an arbitrary point cloud at lower D without training a generative model per dataset.

Our choice. A recent empirical benchmark of ID estimators ranks TwoNN and caPCA as the strongest contenders, with local PCA also reliable at high intrinsic dimension (Bi & Lafaye de Micheaux, 2025). We avoid caPCA because of its computational overhead at high ambient dimension (Table 9) and instead use FO LPCA, which has competitive empirical accuracy across low- and high-dimensional synthetic benchmarks (Table 7). The distance-based estimators can underestimate dimension in high intrinsic dimension, high curvature, or finite-sample regimes, which appears in our estimator ablation (Table 8).

Table 7 compares MLE, TwoNN, FO LPCA, and caPCA on spheres. Table 8 swaps only the per-patch ID estimator in s_{const} (reach always uses FO LPCA $\hat{d}$). On clean manifolds, FO LPCA matches or exceeds TwoNN/MLE discrimination; on S^{16}, T^{16}, and hollow cubes, distance-based estimators collapse s_{const} because of high-variance local $\hat{d}$. On pure noise, MLE reports $s_{\text{const}} = 0.469$ versus 0.062 for FO LPCA. TwoNN per-patch flags the lollipop junction ($s_{\text{const}} = 0.003$) but misses the singularity on the intersecting S^2 and the Klein bottle in $\mathbb{R}^3$, where FO LPCA scores lower.

Table 9 reports wall-clock time per estimator call in data space (seconds, mean $\pm$ std over 3 trials). FO LPCA remains tractable at $D = 1000$; caPCA ($K = 50$) exceeds 500 s per call at medium–high D, motivating our default choice.

J scRNA-SEQ APPENDIX RESULTS

Figure 6 in the main text focuses on Puram and Baron2 PCA LPCA densities and low-dimensional visualisations. This section reports asymptotic GMI sweeps for all five scRNA-seq datasets in our study: Puram and Baron2 in Figure 23, and Darmanis, pancreatic endocrinogenesis, and embryoid body in Figure 24. Figure 25 quantifies the analysis on projection-stability.

Table 7: Intrinsic dimension estimates for spheres S^d in $\mathbb{R}^D$, $N = 2000$, mean $\pm$ std over 3 trials. caPCA uses $K = 50$ neighbours. Latent columns use an AE with latent dimension d_ℓ. Dashes: estimator failure in 1-D latent space.

	Data space					Latent space			
d	MLE	TwoNN	lPCA	caPCA	d_ℓ	MLE	TwoNN	lPCA	caPCA
Low ambient ($D = 10$)									
3	$3.00_{\pm.02}$	$2.92_{\pm.03}$	$4.00_{\pm.00}$	$3.00_{\pm.00}$	3	$2.96_{\pm.01}$	$2.92_{\pm.04}$	$3.00_{\pm.00}$	$2.99_{\pm.00}$
3	$3.00_{\pm.02}$	$2.92_{\pm.03}$	$4.00_{\pm.00}$	$3.00_{\pm.00}$	1	–	–	–	$1.00_{\pm.00}$
5	$4.88_{\pm.02}$	$4.86_{\pm.06}$	$6.00_{\pm.00}$	$5.00_{\pm.00}$	5	$4.82_{\pm.03}$	$4.91_{\pm.09}$	$5.00_{\pm.00}$	$5.00_{\pm.00}$
5	$4.88_{\pm.02}$	$4.86_{\pm.06}$	$6.00_{\pm.00}$	$5.00_{\pm.00}$	3	$3.02_{\pm.02}$	$2.98_{\pm.08}$	$3.00_{\pm.00}$	$3.00_{\pm.00}$
Medium ambient ($D = 100$)									
3	$3.00_{\pm.02}$	$2.92_{\pm.03}$	$4.00_{\pm.00}$	$3.00_{\pm.00}$	3	$2.87_{\pm.09}$	$2.85_{\pm.04}$	$3.00_{\pm.00}$	$2.96_{\pm.05}$
10	$8.90_{\pm.07}$	$9.35_{\pm.08}$	$11.00_{\pm.00}$	$10.00_{\pm.00}$	10	$8.10_{\pm.00}$	$8.34_{\pm.11}$	$10.00_{\pm.00}$	$9.81_{\pm.05}$
50	$29.51_{\pm.29}$	$32.51_{\pm.59}$	$51.00_{\pm.00}$	$43.95_{\pm.01}$	50	$19.06_{\pm.49}$	$21.66_{\pm.89}$	$30.33_{\pm.47}$	$30.60_{\pm.88}$
50	$29.51_{\pm.29}$	$32.51_{\pm.59}$	$51.00_{\pm.00}$	$43.95_{\pm.01}$	10	$8.83_{\pm.03}$	$9.43_{\pm.28}$	$10.00_{\pm.00}$	$10.00_{\pm.00}$
High ambient ($D = 1000$)									
3	$3.00_{\pm.02}$	$2.92_{\pm.03}$	$4.00_{\pm.00}$	$3.00_{\pm.00}$	3	$2.67_{\pm.03}$	$2.86_{\pm.07}$	$3.00_{\pm.00}$	$2.75_{\pm.01}$
10	$8.90_{\pm.07}$	$9.35_{\pm.08}$	$11.00_{\pm.00}$	$10.00_{\pm.00}$	10	$7.79_{\pm.12}$	$8.48_{\pm.06}$	$9.33_{\pm.47}$	$9.85_{\pm.14}$
20	$15.45_{\pm.09}$	$16.79_{\pm.34}$	$21.00_{\pm.00}$	$20.25_{\pm.01}$	20	$11.52_{\pm.11}$	$12.56_{\pm.20}$	$14.00_{\pm.82}$	$16.39_{\pm.40}$
50	$29.51_{\pm.29}$	$32.51_{\pm.59}$	$51.00_{\pm.00}$	$43.95_{\pm.01}$	20	$12.75_{\pm.08}$	$13.98_{\pm.24}$	$16.33_{\pm.94}$	$18.71_{\pm.44}$

Table 8: ID method ablation: s_{const} (mean $\pm$ std, 3 seeds) for FO lPCA, TwoNN, and MLE per FPS patch, on the same synthetic shapes as Table 1 ($\mathbb{R}^{128}$, $N = 10{,}000$). s_{reach} is fixed (always from FO lPCA $\hat{d}$).

Shape	$s_{\text{const}}^{\text{lPCA}}$	$s_{\text{const}}^{\text{TwoNN}}$	$s_{\text{const}}^{\text{MLE}}$
Manifold			
Circle	1.000 ± 0.000	0.891 ± 0.008	1.000 ± 0.000
Sphere S2	1.000 ± 0.000	0.817 ± 0.003	0.848 ± 0.004
Sphere S16	0.937 ± 0.006	0.365 ± 0.003	0.409 ± 0.005
Torus 16	0.927 ± 0.006	0.410 ± 0.007	0.421 ± 0.009
Torus 32	0.940 ± 0.002	0.391 ± 0.008	0.444 ± 0.001
Disjoint S2	1.000 ± 0.000	0.810 ± 0.002	0.862 ± 0.022
Klein Bottle R4	1.000 ± 0.000	0.817 ± 0.015	0.825 ± 0.006
Manifold w/ Boundary			
Swiss Roll	1.000 ± 0.000	0.814 ± 0.009	0.863 ± 0.007
Hollow Cube C16	0.840 ± 0.003	0.345 ± 0.004	0.431 ± 0.006
Hollow Cube C32	0.861 ± 0.002	0.390 ± 0.011	0.422 ± 0.015
Fibre Bundle	0.906 ± 0.006	0.807 ± 0.021	0.856 ± 0.016
Nearly Manifold			
S2+noise(0.02)	0.664 ± 0.003	0.510 ± 0.009	0.656 ± 0.003
S2+noise(0.15)	0.485 ± 0.003	0.371 ± 0.008	0.400 ± 0.002
Cone (Singular)	0.987 ± 0.019	0.808 ± 0.009	0.842 ± 0.006
Lollipop	0.804 ± 0.008	0.003 ± 0.002	0.629 ± 0.020
Pinch Torus 2	0.892 ± 0.005	0.839 ± 0.004	0.827 ± 0.006
Intersecting S2	0.693 ± 0.001	0.813 ± 0.010	0.718 ± 0.003
Klein Bottle R3	0.737 ± 0.006	0.822 ± 0.002	0.661 ± 0.004
Not Manifold			
Spheres (TopoAE)	0.136 ± 0.063	0.431 ± 0.011	0.017 ± 0.002
Pure Noise	0.062 ± 0.002	0.252 ± 0.001	0.469 ± 0.008

Table 9: Wall-clock time (seconds) per ID estimator call in data space for spheres S^d in $\mathbb{R}^D$, $N = 2000$, mean $\pm$ std over 3 trials. caPCA uses $K = 50$ neighbours.

d	MLE	TwoNN	lPCA	caPCA
Low ambient ($D = 10$)				
3	$0.05_{\pm.00}$	$0.01_{\pm.00}$	$0.00_{\pm.00}$	$1.21_{\pm.05}$
5	$0.07_{\pm.00}$	$0.03_{\pm.00}$	$0.00_{\pm.00}$	$1.28_{\pm.06}$
Medium ambient ($D = 100$)				
3	$1.68_{\pm.69}$	$0.40_{\pm.08}$	$0.69_{\pm.17}$	$905.43_{\pm529.64}$
10	$2.32_{\pm1.35}$	$0.54_{\pm.17}$	$1.05_{\pm.57}$	$1001.60_{\pm310.79}$
50	$0.93_{\pm.49}$	$0.21_{\pm.11}$	$0.52_{\pm.37}$	$555.03_{\pm95.88}$
High ambient ($D = 1000$)				
3	$0.84_{\pm.27}$	$1.31_{\pm.80}$	$62.76_{\pm34.36}$	$569.97_{\pm28.68}$
10	$1.07_{\pm.09}$	$1.98_{\pm.20}$	$89.64_{\pm4.19}$	$579.85_{\pm23.18}$
20	$1.17_{\pm.24}$	$1.09_{\pm.66}$	$88.25_{\pm12.14}$	$590.31_{\pm66.04}$
50	$1.41_{\pm.15}$	$2.05_{\pm.58}$	$92.21_{\pm4.03}$	$621.19_{\pm66.11}$

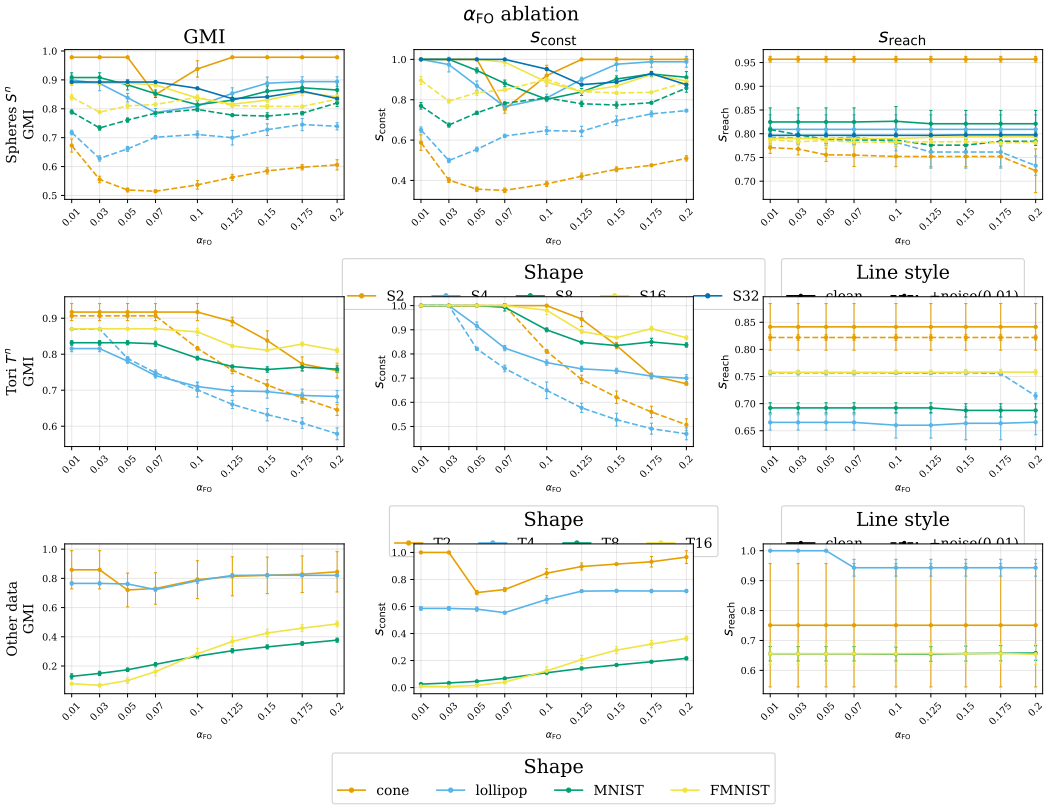


Figure 21: GMI, s_{const}, and s_{reach} vs. $\alpha_{\text{FO}} \in [0.01, 0.20]$ on synthetic shapes and MNIST / Fashion-MNIST ($N = 10{,}000$, 5 seeds).

Table 10: Component scores with k_{patch} forced to a fixed value on S^2 and S^{16}, $N = 10{,}000$, 3 seeds. Both bounds of Appendix G are overridden so that the reported k_{patch} is used directly.

Shape	k_{patch}	$\hat{d}$	s_{const}	s_{reach}	GMI
S^2	11	2.00	1.000	0.974	0.987
S^{16}	20	13.89	0.790	0.793	0.791

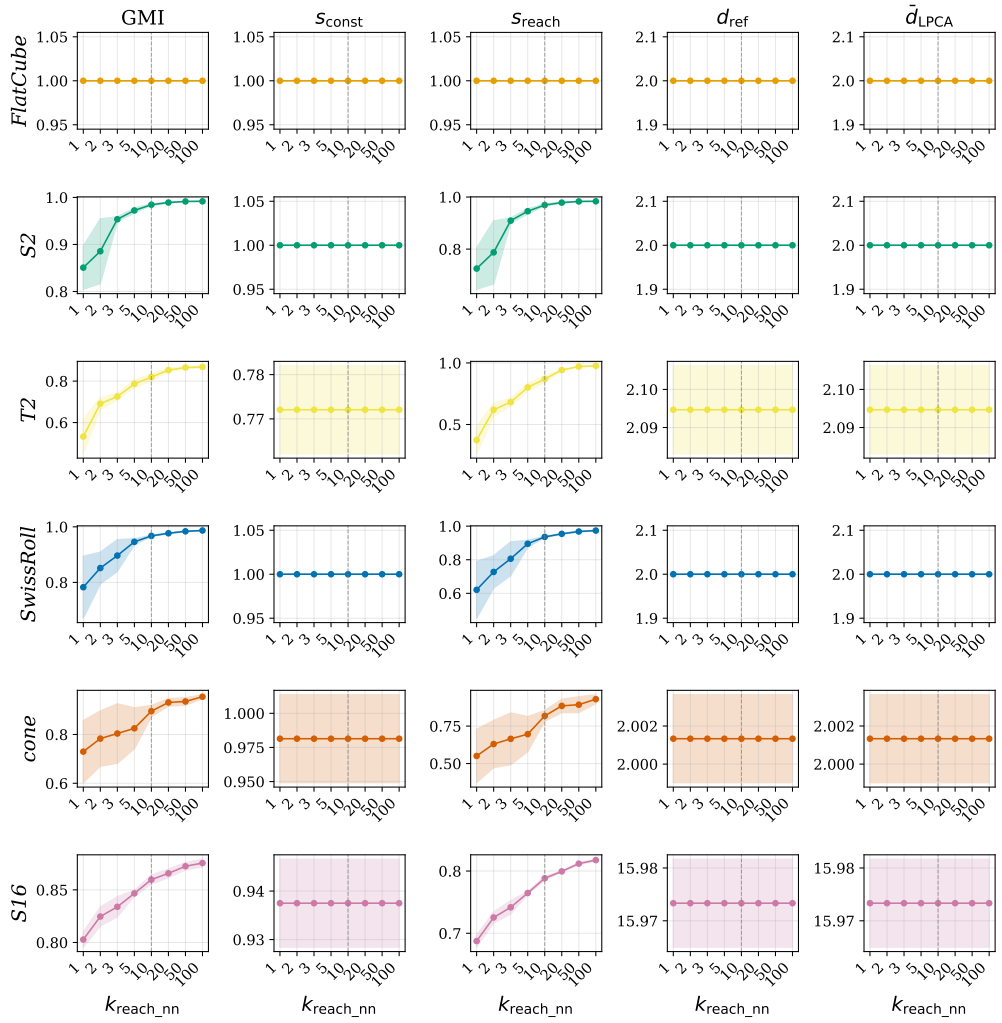
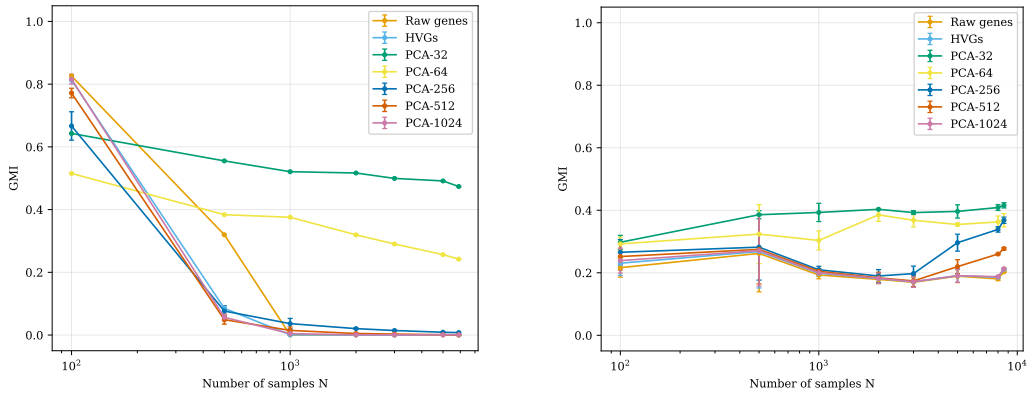


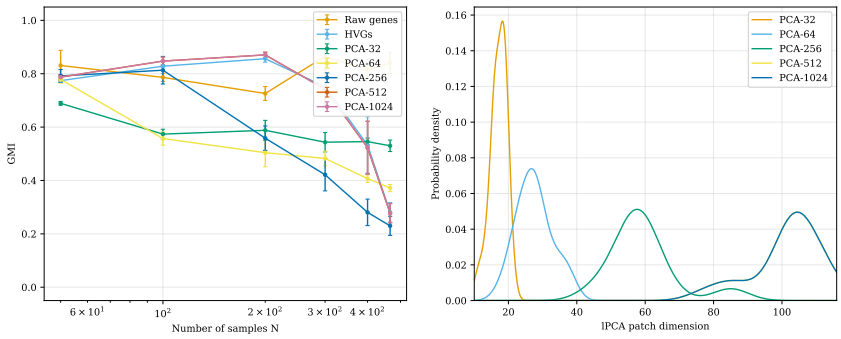
Figure 22: Sensitivity of s_{reach} to $k_{\text{reach_nn}}$ ($N = 5000$). Default $k_{\text{reach_nn}} = 10$.



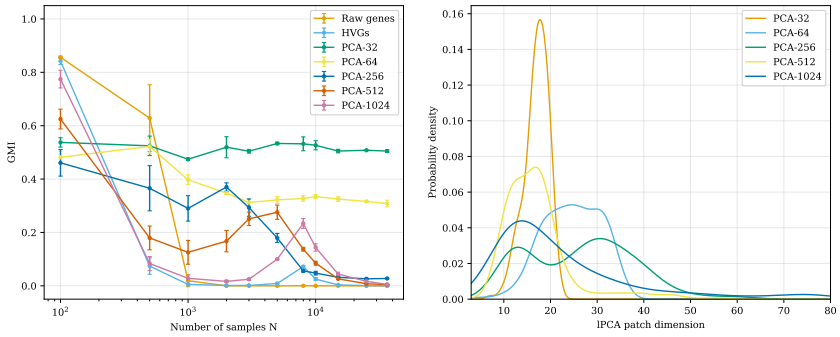
(a) Puram ($N = 5{,}902$, $D = 23{,}686$)

(b) Baron2 ($N = 8{,}569$, $D = 20{,}125$)

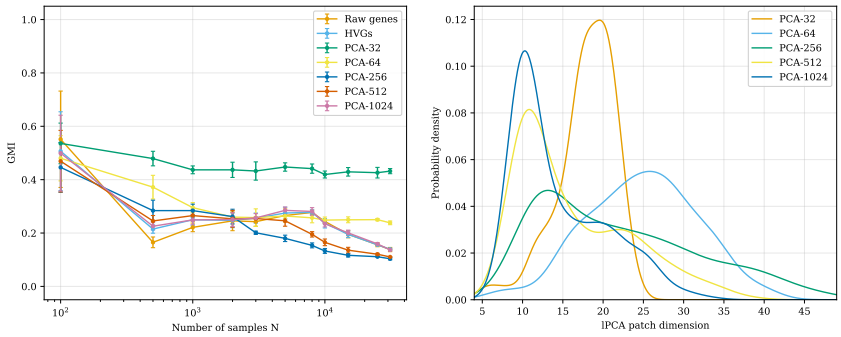
Figure 23: GMI vs. N for raw genes, HVGs, and PCA-D' ($D' \in \{32, 64, 256, 512, 1024\}$) on Puram and Baron2 scRNA-seq (3 seeds).



(a) Darmanis ($N = 466$, $D = 22{,}088$)



(b) Pancreatic endocrinogenesis ($N = 36{,}351$, $D = 17{,}327$)



(c) Embryoid body ($N = 31{,}161$, $D = 33{,}694$)

Figure 24: GMI vs. N for raw genes, HVGs, and PCA-D' ($D' \in \{32, 64, 256, 512, 1024\}$) (left of each panel) and per-patch LPCA dimension densities (Gaussian KDE) for the PCA representations (right of each panel), 3 seeds. Darmanis (a) shows elevated raw-gene GMI in the $N \ll D$ regime where sample-complexity guarantees fail (Narayanan & Mitter, 2010); we treat this as a finite-sample artefact rather than evidence for manifold structure in gene space.

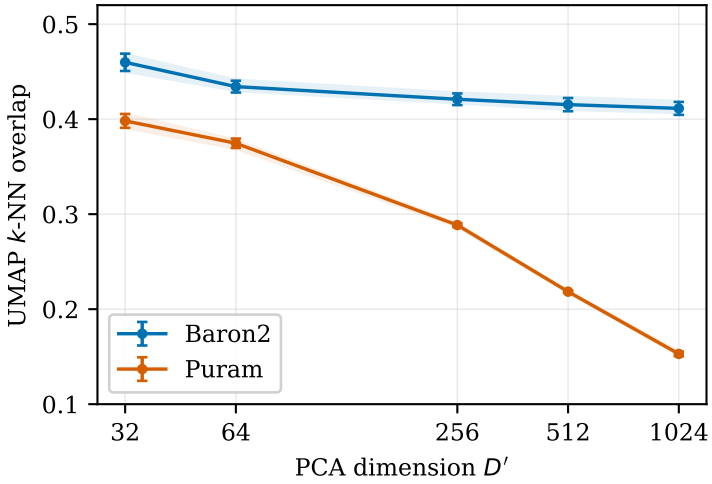


Figure 25: UMAP k-NN overlap (k=15) against PCA dimension D', for the same pair of datasets as Figure 6. $N = 2{,}000$ cells, 3 seeds; error bars are ± 1 standard deviation and shaded bands span the min–max across seeds, both small relative to the trend.

Quantifying projection stability. For each PCA-D' cloud, we take each cell's $k = 15$ nearest neighbours in the cloud and in its 2D UMAP and report the mean Jaccard overlap of the two neighbour sets (Figure 25). Baron2's overlap trends down slowly while Puram's collapses faster. This tracks the collapse of Puram's GMI under the same PCA projections. Puram's 2D map therefore stops representing the cloud it was computed from once that cloud is not aggressively truncated: its apparent low-dimensional structure is a property of the chosen PCA cap, not of gene space.

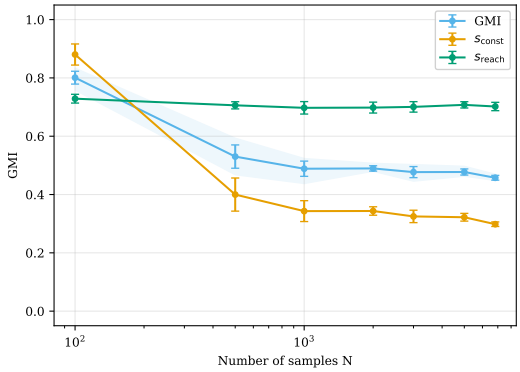
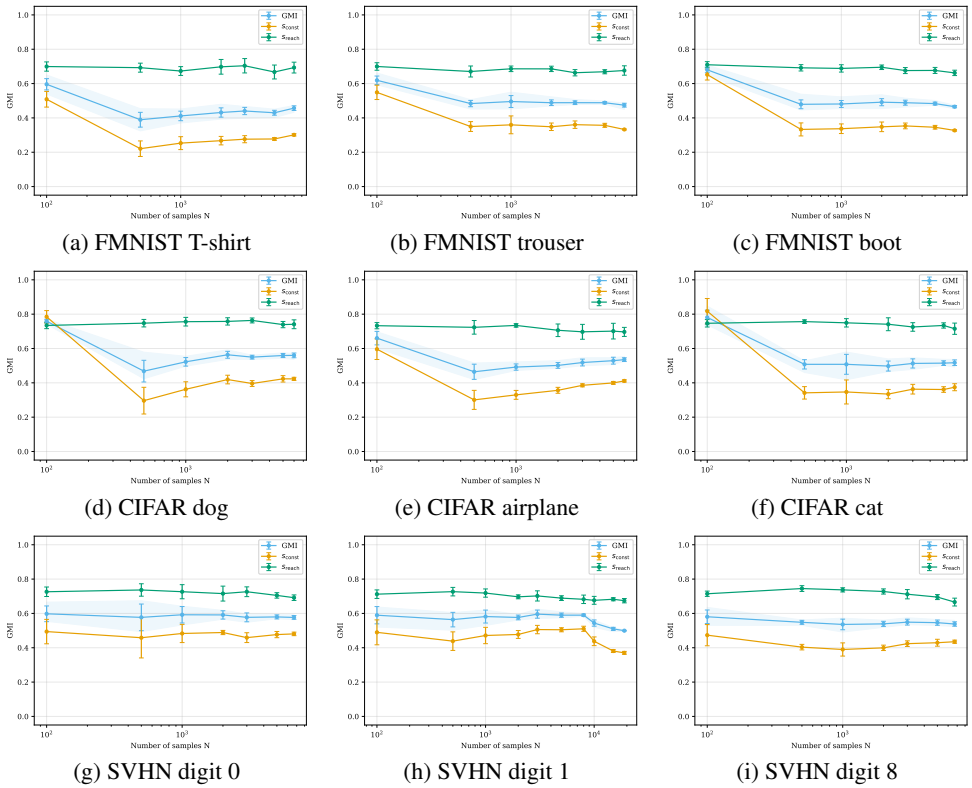


Figure 26: Raw-pixel GMI vs. N for the MNIST digit 8 class subset, omitted from Figure 5.



(a) FMNIST T-shirt (b) FMNIST trouser (c) FMNIST boot

(d) CIFAR dog (e) CIFAR airplane (f) CIFAR cat

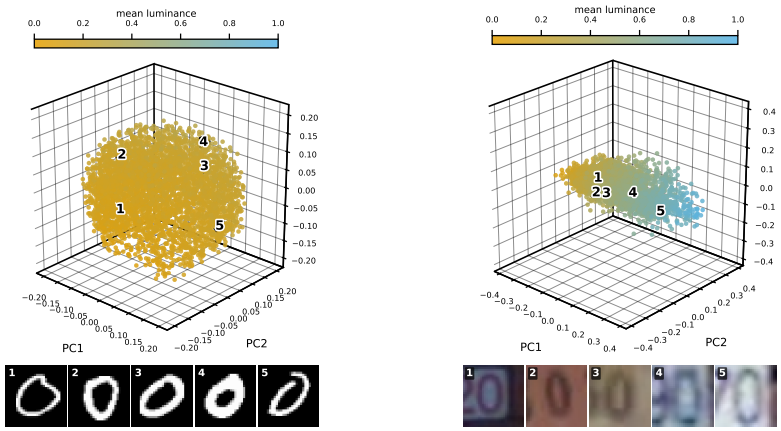
(g) SVHN digit 0 (h) SVHN digit 1 (i) SVHN digit 8

Figure 27: Raw-pixel GMI vs. N for per-class subsets omitted from Figure 5. Row 1: Fashion-MNIST. Row 2: CIFAR-10. Row 3: SVHN class subsets.

K IMAGE APPENDIX RESULTS

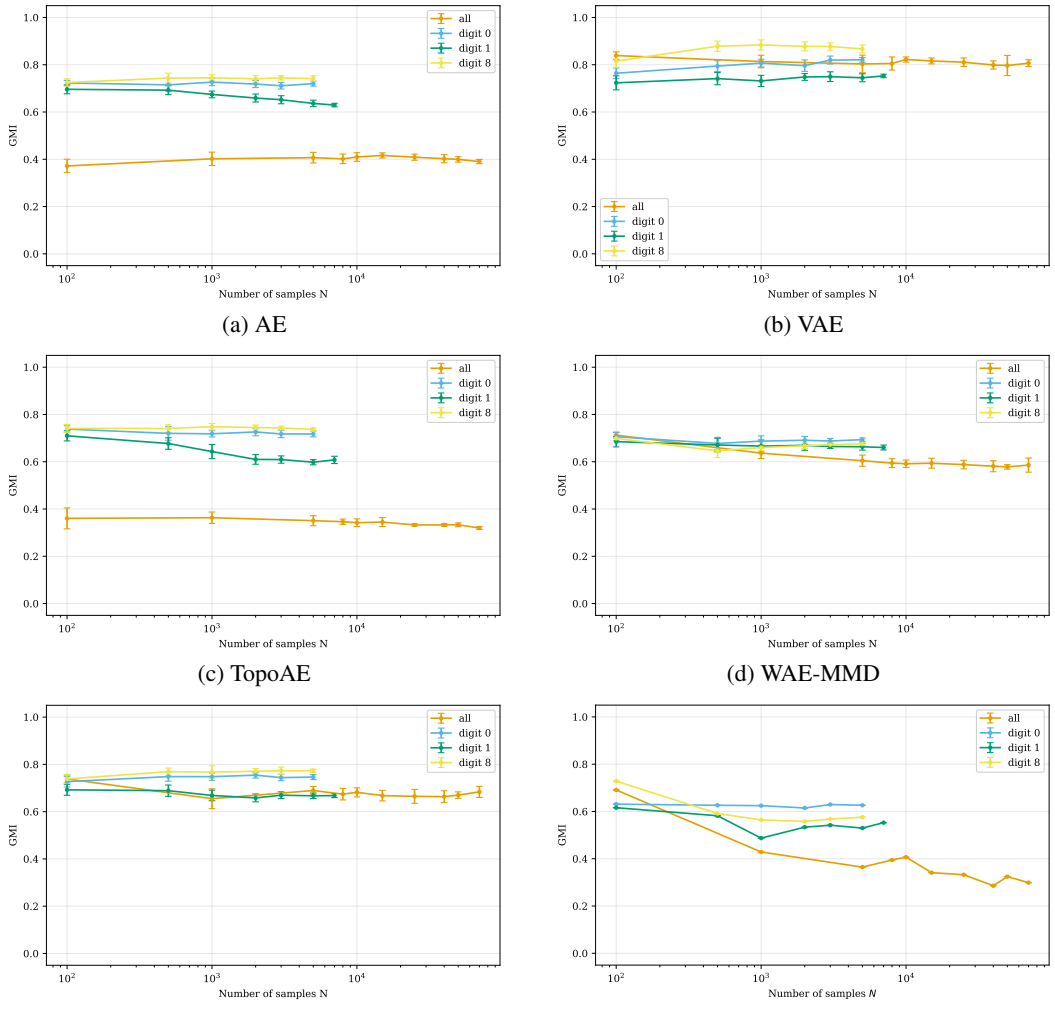
Figures 26 and 27 report per-class raw-pixel asymptotics omitted from the main text. Figures 29–33 collect encoder asymptotics, ImageNet bars, the WAE-GAN reconstruction correlation, and reconstruction-component sweeps.

Manifolds may not coincide with semantic class labels and class boundaries. Semantic and geometric partitions can diverge (Figure 28): on digit 0, MNIST PC1 tracks stroke shape at near-constant luminance, whereas SVHN PC1 captures brightness variation. SVHN's higher all-class GMI (Figure 5) reflects a shared brightness axis across classes, not cleaner within-class shape geometry.



(a) MNIST digit 0 (b) SVHN digit 0

Figure 28: Within-class digit 0 ($N = 4{,}000$): top-3 PCs coloured by mean luminance with five anchors along PC1; bottom: matching images. SVHN PC1 tracks brightness; MNIST PC1 tracks shape at near-constant luminance.



(a) AE (b) VAE

(c) TopoAE (d) WAE-MMD

(e) WAE-GAN (f) SD-VAE PCA-1024

Figure 29: MNIST latent-space asymptotic GMI ($d = 16$; SD-VAE PCA-projected to 1024-D).

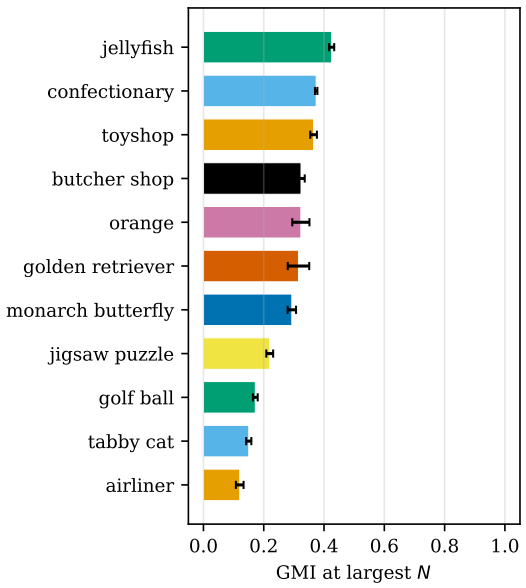


Figure 30: ImageNet SD-VAE individual classes at largest N (PCA-256, 5 seeds).

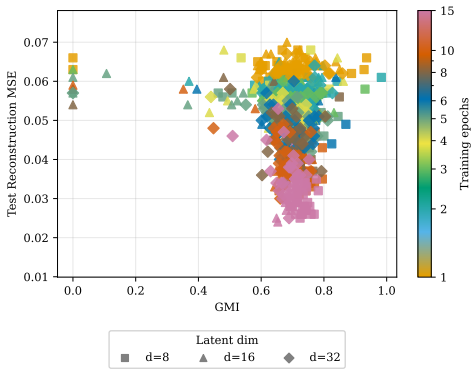
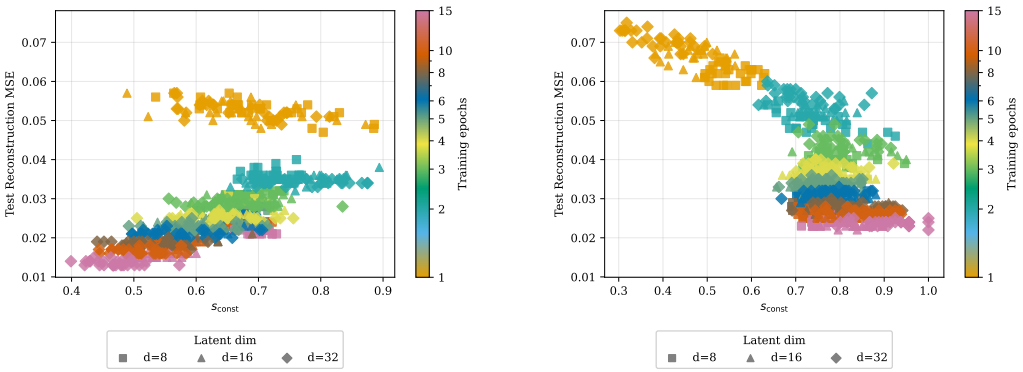
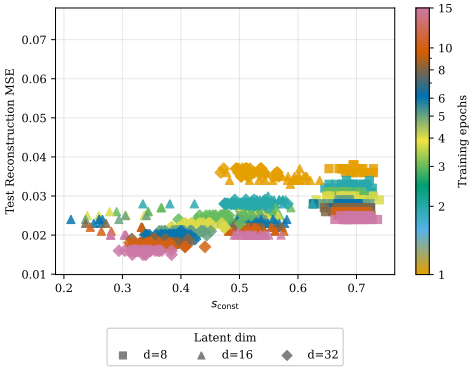


Figure 31: WAE-GAN GMI vs. test reconstruction MSE ($r = -0.14$), omitted from Figure 9 (same sweep; 30 seeds, $d \in \{8, 16, 32\}$).

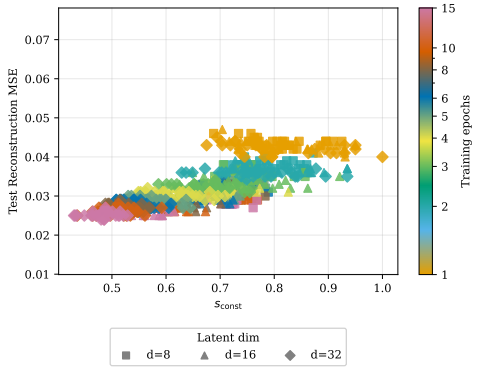


(a) AE s_{const} ($r = 0.54$)

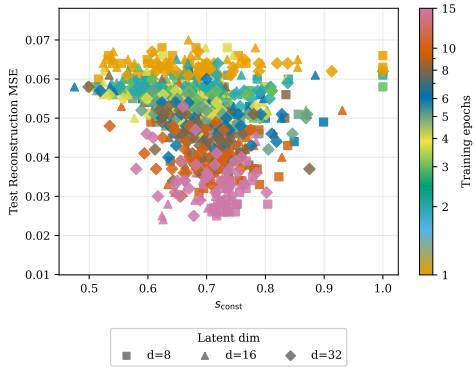
(b) VAE s_{const} ($r = -0.77$)



(c) TopoAE s_{const} ($r = 0.64$)

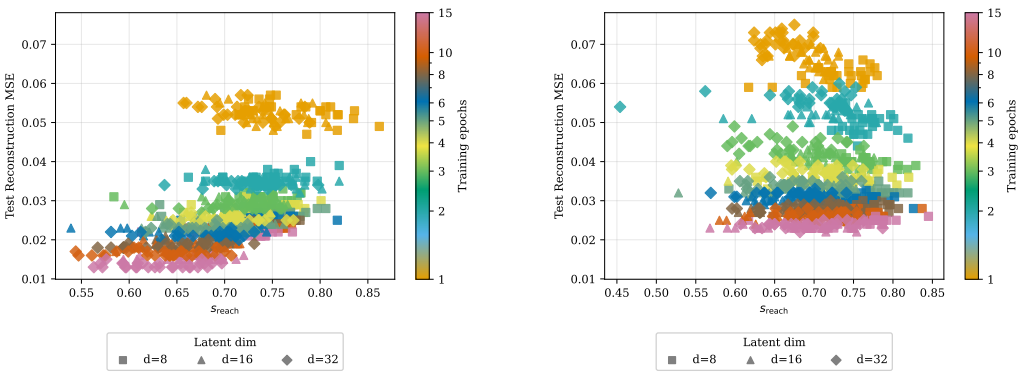


(d) WAE-MMD s_{const} ($r = 0.77$)



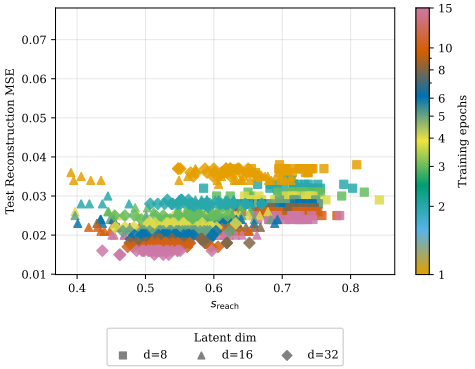
(e) WAE-GAN s_{const} ($r = -0.24$)

Figure 32: s_{const} vs. test reconstruction MSE (same sweep as Figure 9; $d \in \{8, 16, 32\}$).

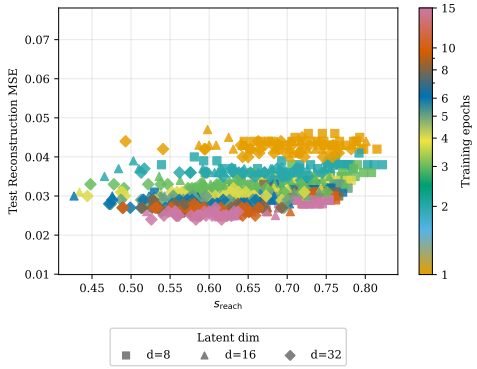


(a) AE s_{reach} ($r = 0.53$)

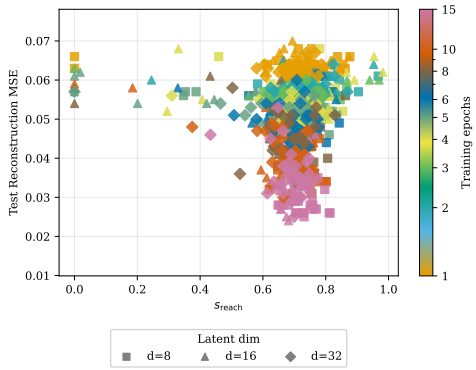
(b) VAE s_{reach} ($r = -0.05$)



(c) TopoAE s_{reach} ($r = 0.53$)



(d) WAE-MMD s_{reach} ($r = 0.44$)



(e) WAE-GAN s_{reach} ($r = -0.02$)

Figure 33: s_{reach} vs. test reconstruction MSE (same sweep as Figure 9; $d \in \{8, 16, 32\}$).

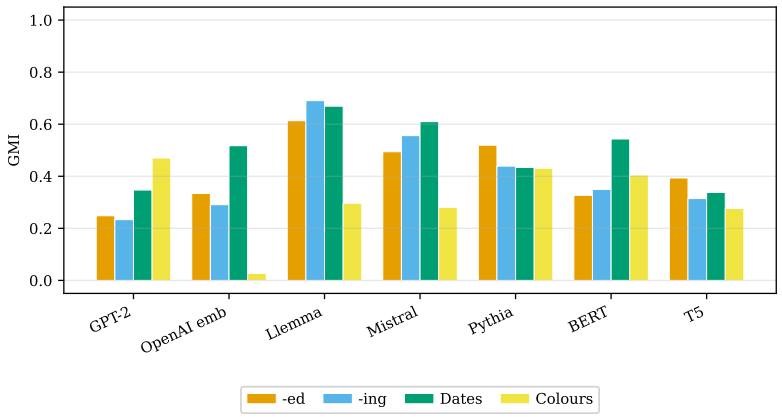
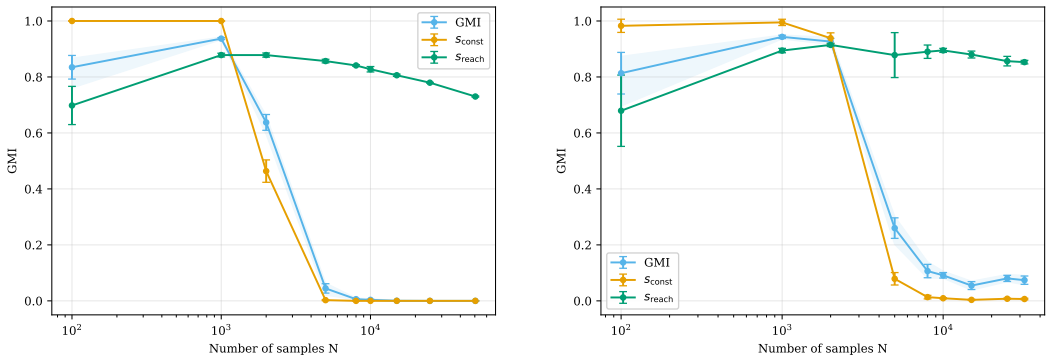


Figure 34: Vocabulary-subset GMI



(a) Pythia 6.9B

(b) Mistral 7B

(c) BERT base

(d) GPT-2

Figure 35: Token embedding GMI vs. N (representative LLMs).

L TOKEN EMBEDDING APPENDIX RESULTS

Figure 35 reports full-vocabulary GMI vs. N on representative LLMs; Figure 34 reports curated vocabulary subsets.

Figure 21 sweeps $\alpha_{\mathrm{FO}} \in [0.01, 0.20]$ on synthetic shapes and image datasets. Full-vocabulary token clouds ($D \in \{768, 4096\}$) can saturate at $d_{\mathrm{LPCA}} \approx k_{\mathrm{patch}} - 1$ when α_{FO} is too low, inflating s_{const}. Figure 36 extends the sweep to $\alpha_{\mathrm{FO}} \in \{0.05, 0.1, 0.175, 0.3, 0.5\}$ on subsampled GPT-2 and Pythia 6.9B tables ($N \in \{1{,}000, 5{,}000, 25{,}000\}$, 3 seeds, native dimension). At small N, low α_{FO} can report misleadingly high GMI; no value in $\{0.05, \ldots, 0.5\}$ yields stable full-vocabulary scores at scale. We retain $\alpha_{\mathrm{FO}} = 0.175$ to match the synthetic and image ablations.

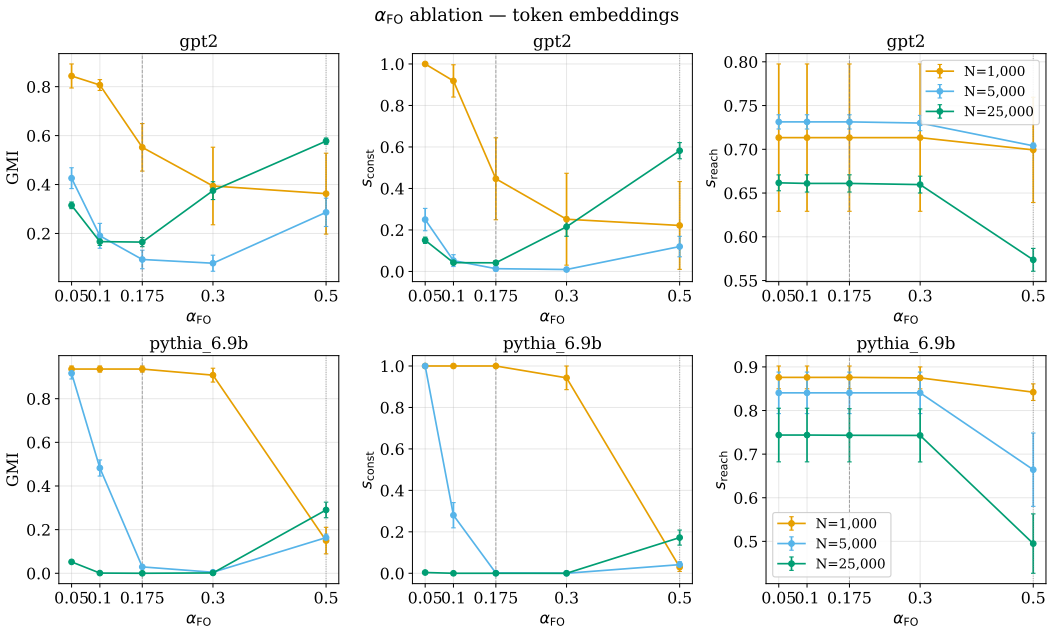
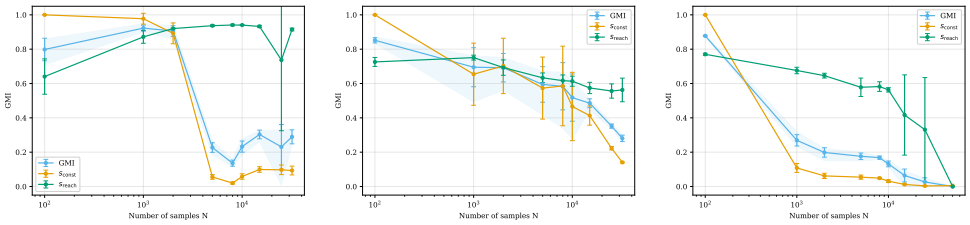


Figure 36: α_{FO} ablation on GPT-2 and Pythia 6.9B embeddings (3 seeds). Dashed: default $\alpha_{FO} = 0.175$.



(a) Llemma 7B (b) T5 base (c) OpenAI 3

Figure 37: Token embedding GMI vs. N for the three embedding tables not shown in Figure 35.

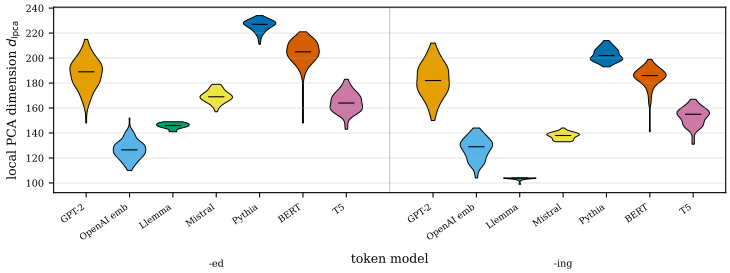


Figure 38: Per-patch local LPCA dimension distributions at FPS patch centres on the *-ed and *-ing morphological token subsets (PCA-256), shown for each embedding table and grouped by suffix. Black bars mark medians.

Figure 37 shows GMI vs. N for the three embedding tables omitted from Figure 35, and Figure 38 shows per-patch geometry on the morphological subsets. Extreme FPS centres on *-ing/*-ed subsets are listed in Section M.

M TOKEN EMBEDDING TABLES

Table 11 reports PCA-256 variance retained on curated token subsets. The remaining tables list FPS centres with extreme local LPCA and reach across embedding tables: for the *-ing subset, Tables 12, 13, 14, and 15; for the *-ed subset, Tables 16, 17, 18, and 19.

Table 11: Fraction of variance retained by PCA-256 on token subsets across all seven embedding tables (same projection used in Figures 34 and 38). Full vocabulary uses PCA-256 for Pythia, Mistral, and Llemma only; all other columns use PCA-256 for every model. Dates and colours ($N \lesssim 366$) are mean-pooled phrase embeddings and are effectively full-rank in PCA-256.

Model	Full vocab	*-ed	*-ing	Dates	Colours
GPT-2	—	0.76	0.77	1.00	1.00
OpenAI 3	—	0.83	0.84	0.96	1.00
Llemma 7B	0.26	0.65	0.79	1.00	1.00
Mistral 7B	0.23	0.55	0.66	1.00	1.00
Pythia 6.9B	0.25	0.40	0.44	1.00	1.00
BERT base	—	0.74	0.77	1.00	1.00
T5 base	—	0.84	0.86	1.00	1.00

Table 12: Lowest d_{LPCA} (*-ing suffix; PCA-256 FPS centres; value in parentheses at 3 s.f.; duplicate tokens seen across models are shown in red.)

	GPT-2	OpenAI 3	Llemma	Mistral	Pythia	BERT	T5
1	demanding (150)	amazing (104)	string (99)	nothing (133)	something (193)	uising (141)	fascinating (131)
2	evolving (154)	understanding (106)	evening (102)	engineering (133)	emitting (194)	working (164)	demanding (135)
3	describing (154)	trying (108)	networking (103)	aining (133)	processing (194)	living (165)	decreasing (144)
4	warning (158)	grinning (113)	stelling (103)	having (134)	startling (195)	running (168)	planning (144)
5	shocking (159)	stimulating (113)	unning (104)	string (134)	everything (195)	building (172)	lightning (144)
6	publishing (160)	excruciating (114)	seeing (104)	planning (134)	monitoring (196)	following (174)	learning (145)
7	looking (161)	serving (115)	surprising (104)	padding (135)	according (196)	making (175)	imaging (145)
8	leading (164)	deteriorating (115)	sampling (104)	setting (135)	signalling (196)	looking (175)	conditioning (145)
9	everything (164)	irling (115)	burning (104)	including (135)	adding (197)	playing (178)	lasting (146)
10	during (164)	emanating (115)	landing (104)	during (136)	reporting (197)	breathing (178)	marketing (146)

Table 13: Highest d_{LPCA} (*-ing suffix; PCA-256 FPS centres; value in parentheses at 3 s.f.; duplicate tokens seen across models are shown in red.)

	GPT-2	OpenAI 3	Llemma	Mistral	Pythia	BERT	T5
1	turing (212)	jinping (144)	burning (104)	conting (144)	lightning (214)	dating (199)	spring (167)
2	sterling (211)	cunning (142)	ughing (104)	binding (141)	speaking (213)	missing (197)	standing (166)
3	downing (206)	tearing (142)	unning (104)	ceiling (141)	spring (213)	winning (194)	missing (163)
4	boeing (206)	kidnapping (142)	fixing (104)	selling (141)	warning (212)	crossing (194)	pudding (163)
5	nesting (204)	itching (141)	sampling (104)	wedding (140)	padding (212)	gathering (193)	ruling (162)
6	conting (203)	wasting (140)	during (104)	breaking (140)	wrestling (210)	helping (193)	willing (161)
7	rating (203)	string (140)	surprising (104)	umbing (140)	pending (210)	anything (193)	string (160)
8	neighbouring (202)	selling (139)	forcing (104)	incoming (140)	nsstring (210)	spending (192)	wyoming (160)
9	lightning (202)	taking (138)	seeing (104)	morning (139)	starting (210)	evening (192)	greeting (160)
10	blooming (201)	smoking (138)	acting (104)	reading (139)	dissenting (209)	setting (191)	according (159)

Table 14: Lowest τ (*-ing suffix; PCA-256 FPS centres; value in parentheses at 3 s.f.; duplicate tokens seen across models are shown in red.)

	GPT-2	OpenAI 3	Llemma	Mistral	Pythia	BERT	T5
1	shocking (0.32)	warning (0.348)	string (0.5)	string (0.423)	something (0.351)	uising (0.382)	fascinating (0.413)
2	describing (0.358)	anything (0.348)	during (0.549)	nothing (0.439)	startling (0.362)	traveling (0.402)	learning (0.448)
3	everything (0.365)	during (0.36)	seeing (0.56)	setting (0.448)	according (0.363)	falling (0.412)	decreasing (0.452)
4	including (0.38)	refusing (0.361)	surprising (0.561)	indicating (0.45)	processing (0.363)	adding (0.414)	something (0.455)
5	finding (0.384)	asking (0.366)	making (0.567)	including (0.465)	building (0.366)	broadcasting (0.415)	marketing (0.457)
6	starting (0.387)	adding (0.366)	warning (0.567)	fascinating (0.466)	regarding (0.366)	fishing (0.418)	delivering (0.475)
7	banning (0.388)	moving (0.369)	forcing (0.571)	working (0.466)	setting (0.368)	showing (0.419)	including (0.476)
8	testing (0.398)	driving (0.37)	sampling (0.577)	having (0.468)	thinking (0.368)	surrounding (0.419)	cleaning (0.477)
9	mining (0.406)	amazing (0.371)	fixing (0.58)	padding (0.475)	moving (0.368)	considering (0.422)	writing (0.484)
10	winning (0.407)	taking (0.377)	networking (0.58)	breaking (0.48)	warning (0.369)	saving (0.422)	beginning (0.485)

Table 15: Highest τ (*-ing suffix; PCA-256 FPS centres; value in parentheses at 3 s.f.; duplicate tokens seen across models are shown in red.)

	GPT-2	OpenAI 3	Llemma	Mistral	Pythia	BERT	T5
1	ceiling (0.487)	viking (0.525)	stelling (0.609)	ceiling (0.537)	wedding (0.416)	following (0.541)	wyoming (0.579)
2	conting (0.48)	lightning (0.513)	gering (0.609)	conting (0.537)	boeing (0.415)	spring (0.532)	boeing (0.576)
3	turing (0.469)	jinping (0.498)	imming (0.596)	wedding (0.534)	nursing (0.409)	working (0.525)	sterling (0.565)
4	boeing (0.468)	hawking (0.492)	ipping (0.593)	umbing (0.518)	lightning (0.408)	running (0.514)	pudding (0.561)
5	downing (0.468)	fledgling (0.483)	acting (0.59)	during (0.517)	shipping (0.407)	during (0.51)	wellbeing (0.561)
6	recurring (0.464)	rohing (0.481)	landing (0.59)	lacking (0.506)	wiping (0.404)	living (0.508)	string (0.556)
7	sterling (0.463)	sporting (0.474)	unning (0.587)	logging (0.505)	signalling (0.404)	building (0.506)	fitting (0.554)
8	fitting (0.463)	piercing (0.473)	evening (0.587)	incoming (0.502)	craving (0.403)	including (0.497)	spring (0.551)
9	landing (0.463)	lansing (0.467)	ughing (0.583)	listing (0.502)	lasting (0.402)	holding (0.495)	evening (0.549)
10	lightning (0.462)	farming (0.465)	atting (0.581)	morning (0.502)	matching (0.402)	setting (0.492)	ruling (0.547)

Table 16: Lowest d_{LPCA} (*-ed suffix; PCA-256 FPS centres; value in parentheses at 3 s.f.; duplicate tokens seen across models are shown in red.)

	GPT-2	OpenAI 3	Llemma	Mistral	Pythia	BERT	T5
1	threatened (148)	arranted (110)	speed (141)	occurred (157)	realized (211)	orted (148)	inspired (143)
2	agreed (159)	underrated (111)	unsigned (141)	realised (161)	straightened (213)	jared (173)	concerned (147)
3	detailed (162)	sighted (111)	disabled (141)	ished (161)	ained (215)	renamed (183)	added (154)
4	stunned (163)	motivated (113)	called (143)	refused (162)	shouted (215)	called (189)	sophisticated (155)
5	determined (166)	exalted (114)	based (144)	undred (164)	grinned (217)	stared (189)	recommended (156)
6	interested (166)	hindered (114)	created (144)	handed (165)	indeed (218)	responded (189)	printed (156)
7	continued (167)	fashioned (115)	erved (145)	signed (165)	suspected (218)	asked (191)	associated (156)
8	crashed (168)	inquired (115)	undred (145)	allowed (165)	analyzed (219)	considered (192)	declined (156)
9	compared (168)	toppled (116)	verschied (145)	demonstrated (165)	upregulated (220)	dated (193)	concealed (157)
10	experienced (169)	embarked (116)	alfred (145)	created (165)	favored (220)	included (194)	dedicated (158)

Table 17: Highest d_{LPCA} (*-ed suffix; PCA-256 FPS centres; value in parentheses at 3 s.f.; duplicate tokens seen across models are shown in red.)

	GPT-2	OpenAI 3	Llemma	Mistral	Pythia	BERT	T5
1	unlimited (215)	lockheed (152)	granted (149)	protected (179)	ahmed (234)	fixed (221)	handed (183)
2	sched (212)	infrared (142)	proced (149)	preced (178)	unsigned (233)	covered (219)	dated (180)
3	breed (210)	borrowed (142)	opposed (149)	united (178)	sched (233)	mixed (219)	alfred (178)
4	unified (209)	boiled (140)	bounded (148)	embed (177)	accredited (232)	advanced (218)	schied (176)
5	rated (209)	stared (140)	ended (148)	ingred (177)	linked (232)	disabled (217)	mitglied (176)
6	embed (207)	taxed (140)	oured (148)	limited (176)	undefined (232)	applied (216)	colored (174)
7	ained (205)	punched (140)	induced (148)	reserved (175)	shared (232)	limited (216)	hundred (174)
8	erred (205)	twisted (139)	gebied (148)	colored (175)	exped (232)	shared (215)	indeed (173)
9	bonded (204)	bened (139)	incred (148)	distributed (174)	hered (232)	carried (215)	unused (172)
10	seated (204)	greed (138)	detailed (147)	disabled (173)	geared (232)	prepared (215)	assisted (172)

Table 18: Lowest τ (*-ed suffix; PCA-256 FPS centres; value in parentheses at 3 s.f.; duplicate tokens seen across models are shown in red.)

	GPT-2	OpenAI 3	Llemma	Mistral	Pythia	BERT	T5
1	stunned (0.34)	finished (0.306)	disabled (0.471)	expected (0.399)	grinned (0.323)	stared (0.344)	limited (0.46)
2	famed (0.374)	published (0.324)	tried (0.478)	created (0.41)	analyzed (0.33)	responded (0.362)	inspired (0.466)
3	chanted (0.378)	terrified (0.33)	created (0.479)	demonstrated (0.413)	ained (0.338)	restored (0.375)	recommended (0.475)
4	mentioned (0.378)	defined (0.334)	unsigned (0.482)	updated (0.416)	shouted (0.345)	impressed (0.384)	based (0.476)
5	apologized (0.397)	chuckled (0.345)	speed (0.487)	refused (0.419)	affirmed (0.348)	orted (0.387)	printed (0.481)
6	crashed (0.398)	accumulated (0.351)	parsed (0.493)	named (0.421)	upregulated (0.349)	disappeared (0.389)	finished (0.482)
7	answered (0.4)	refused (0.352)	bounded (0.493)	linked (0.423)	embedded (0.351)	disabled (0.395)	colored (0.488)
8	unidentified (0.4)	changed (0.353)	detailed (0.495)	disabled (0.424)	extended (0.355)	confused (0.397)	wanted (0.489)
9	invented (0.401)	packed (0.353)	granted (0.497)	limited (0.425)	funded (0.355)	situated (0.397)	associated (0.49)
10	loved (0.403)	monitored (0.357)	situated (0.499)	based (0.426)	titled (0.355)	credited (0.4)	qualified (0.494)

Table 19: Highest τ (*-ed suffix; PCA-256 FPS centres; value in parentheses at 3 s.f.; duplicate tokens seen across models are shown in red.)

	GPT-2	OpenAI 3	Llemma	Mistral	Pythia	BERT	T5
1	prepared (0.479)	paramed (0.482)	gebied (0.55)	chied (0.484)	indeed (0.401)	called (0.506)	hundred (0.597)
2	shred (0.473)	greed (0.481)	incred (0.541)	ingred (0.482)	ahmed (0.4)	based (0.487)	schied (0.579)
3	breed (0.468)	infrared (0.476)	alfred (0.535)	undred (0.473)	randomised (0.4)	shared (0.484)	alfred (0.577)
4	graduated (0.464)	unprepared (0.473)	proced (0.53)	acknowled (0.47)	exped (0.398)	passed (0.484)	indeed (0.568)
5	randomized (0.461)	lockheed (0.472)	undred (0.529)	preced (0.466)	modelled (0.394)	closed (0.481)	doved (0.566)
6	detailed (0.461)	unfocused (0.472)	opposed (0.528)	protected (0.46)	buffered (0.394)	raised (0.479)	breed (0.566)
7	sched (0.46)	bened (0.471)	indeed (0.526)	prepared (0.459)	geared (0.393)	followed (0.473)	embed (0.561)
8	creed (0.459)	mohammed (0.47)	verschied (0.525)	fried (0.459)	hered (0.393)	united (0.471)	headed (0.561)
9	phased (0.458)	watershed (0.47)	succeed (0.52)	united (0.458)	polarized (0.393)	speed (0.469)	mitglied (0.56)
10	cleared (0.458)	protracted (0.47)	raised (0.518)	deemed (0.457)	accustomed (0.393)	played (0.468)	dated (0.549)