NERVE FAITHFULNESS CERTIFIES DATA GEOMETRY IN LEARNED DICTIONARIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstruction error and sparsity do not determine whether a learned dictionary preserves the geometry of the activations it encodes. We introduce nerve faithfulness, a certified comparison between data geometry and code co-activation geometry. For a non-negative code M, $K(i,i') = \max_j \min(M_{ij}, M_{i'j})$ is exactly the Dowker 1-skeleton; after normalization it induces a metric $d_{\mathfrak{D}}$, and $\delta_{\mathrm{nerve}} = \frac{1}{2}\|d_{\mathfrak{D}} - d_X\|_\infty$ certifies an interleaving through the identity on samples. The certificate is hypothesis-free and costs $O(m^2 n)$ dense or $O(mk^2)$ for k-sparse codes. It remains nearly invariant across a $27\times$ sweep of tuning width for a fixed antipodal polysemantic defect (0.510–0.539), while coherence varies by a factor of 5.4. As a training objective, a correlation form combined with coherence raises rank faithfulness on rotated MNIST from 0.301 to 0.826, lowers the certificate from 0.516 to 0.329, and improves reconstruction. On released GPT-2 dictionaries with FVU 0.136–0.155, code–data rank correlations of only 0.070–0.084 show that good reconstruction can coexist with severe geometric mismatch. Nerve faithfulness adds a distinct, certifiable axis for evaluating and training learned dictionaries.

1 INTRODUCTION

A learned dictionary can reconstruct activations accurately while reorganizing the geometry that made those activations meaningful. Reconstruction error and sparsity quantify fidelity in value and support, but neither asks whether nearby samples remain nearby in the relational structure induced by the code. This geometric mismatch is especially important for interpretability, where features are often read as a structured cover of the underlying data manifold.

For non-negative codes, the required relational geometry is already present in the co-activation matrix. The strongest activation shared by two samples, $K(i,i') = \max_j \min(M_{ij}, M_{i'j})$, is exactly the pair filtration of the sample-side Dowker complex. After normalization, its disagreement with the data metric produces δ_{nerve}, an explicit interleaving certificate through the identity map. The construction removes auxiliary Lipschitz and convex-hull assumptions and is efficient on sparse codes.

The resulting diagnostic captures a failure mode that standard dictionary metrics miss. A fixed antipodal polysemantic defect receives an essentially stable certificate across a $27\times$ change in tuning width, while coherence changes by more than a factor of five. As an objective, nerve faithfulness complements coherence: their combination yields the strongest code–data rank agreement and the smallest certificate on rotated MNIST, while released GPT-2 dictionaries show that low reconstruction error can coexist with near-zero geometric rank agreement.

Contributions.

- **Exact co-activation geometry.** The max–min kernel is the Dowker 1-skeleton, giving a sparse, differentiable representation of relational geometry.
- **A certified data–code comparison.** $\delta_{\mathrm{nerve}} = \frac{1}{2}\|d_{\mathfrak{D}} - d_X\|_\infty$ certifies an interleaving through the identity map without a Lipschitz or convex-hull hypothesis.

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.

- **Stable detection of polysemantic geometry.** Across a $27\times$ tuning-width sweep, the certificate remains stable for a fixed antipodal defect while coherence varies strongly with feature density.
- **A trainable evaluation axis.** Joint nerve and coherence training reaches rank faithfulness above 0.81 on both rotated-MNIST tasks, and released dictionaries reveal substantial code–data geometric mismatch at low FVU — rank correlation 0.070–0.084 on GPT-2 and -0.386 to $+0.197$ across nine layers of Gemma-2-2B, negative in the first half of that network.

2 THE CO-ACTIVATION KERNEL IS THE DOWKER 1-SKELETON

Let $M \in \mathbb{R}_+^{m\times n}$ and take the entrywise cost $\Lambda_{ij} = -M_{ij}$. The Dowker filtration on samples admits an edge $\{i, i'\}$ at scale t iff some feature witnesses both:

$$t(i,i') = \min_j \max\big(-M_{ij}, -M_{i'j}\big) = -\max_j \min\big(M_{ij}, M_{i'j}\big) = -K(i,i'), \tag{1}$$

where

$$K(i,i') := \max_j \min(M_{ij}, M_{i'j}). \tag{2}$$

Proposition 1. *K in* (1) *equals the Dowker 1-skeleton filtration exactly, is symmetric, invariant to permutation and to monotone rescaling of the features, and is computable in $O(m^2n)$ with an $O(bmn)$-memory blocked pass. Its softmin/softmax surrogate converges at rate $O(1/\beta)$ (measured maximum error 3.2×10^{-1}, 6.1×10^{-2}, 1.0×10^{-2} at $\beta = 5, 20, 100$).*

Define $d_{\mathfrak{D}} = \big(\max K - K\big)/\mu$ with μ the mean off-diagonal value, and let d_X be the data metric normalised the same way.

3 THE OBJECTIVE AND ITS CERTIFICATE

Theorem 2. *Fix the convention $\mathrm{VR}_t(P) = \{\sigma : \operatorname{diam}\sigma \le t\}$ and mean-normalise both dissimilarities, as in (Gaukstad et al., 2026). Let $\mathcal{F} = \mathrm{VR}(X, d_X)$, let $\mathcal{G}$ be the Dowker filtration of M on the same vertex set under the same normalisation, and set*

$$\delta_{\mathrm{nerve}}(M, X) \;:=\; \tfrac{1}{2}\big\|d_{\mathfrak{D}} - d_X\big\|_\infty.$$

Then

1. *the identity map on samples is a δ_{nerve}-interleaving between $\mathrm{VR}(X, d_X)$ and $\mathrm{VR}(X, d_{\mathfrak{D}})$, and δ_{nerve} is the least δ for which the interleaving criterion certifies the identity;*
2. *$\mathcal{G} \subseteq \mathrm{VR}(X, d_{\mathfrak{D}})$, with equality on the 1-skeleton; hence the persistent H_0 of the code's Dowker filtration is δ_{nerve}-close to that of the data.*

No Lipschitz hypothesis and no convex-hull extension is used.

Proof. (1) With $f = g = \mathrm{id}$ the four conditions of the criterion reduce to the metric comparison: $d_X(a,b) \le t$ gives $d_{\mathfrak{D}}(a,b) \le t + 2\delta_{\mathrm{nerve}}$ and conversely, while the round-trip conditions are trivial because $gf = \mathrm{id}$. Minimality follows by evaluating the first condition at $t = d_X(i,i')$ for the maximising pair. (2) A Dowker simplex at scale t has a common witness j, so every pair in it satisfies $\max(-M_{ij}, -M_{i'j}) \le t$ and therefore $d_{\mathfrak{D}} \le t$ before normalisation; the normalisation is a common monotone affine map, so the inclusion is preserved. Equality on the 1-skeleton is (1). □

Remark 3. Metric alignment. The shift $\max K - K$ and mean-pairwise normalization align filtrations that otherwise have no canonical common index and match the normalization used for coherence. The inclusion $\mathcal{G} \subseteq \mathrm{VR}(X, d_{\mathfrak{D}})$ is exact on the 1-skeleton, so part (2) is exact for H_0 and an inclusion for H_1.

The loss is the top-k hinge $\mathcal{L}_{\mathrm{nerve}} = \mathrm{TopK}_{i<i'}\big[\tfrac{1}{2}|d_{\mathfrak{D}} - d_X| - \tau\big]_+$, evaluated on a random subsample of rows per step, with the smooth kernel of Proposition 1.

Remark 4. δ_{nerve} is invariant to permuting or monotonically rescaling features, so it measures the relation and not the parametrisation — unlike reconstruction error, and unlike coherence, which is sensitive to the ambient scale of the columns only through the mean-distance normalisation.

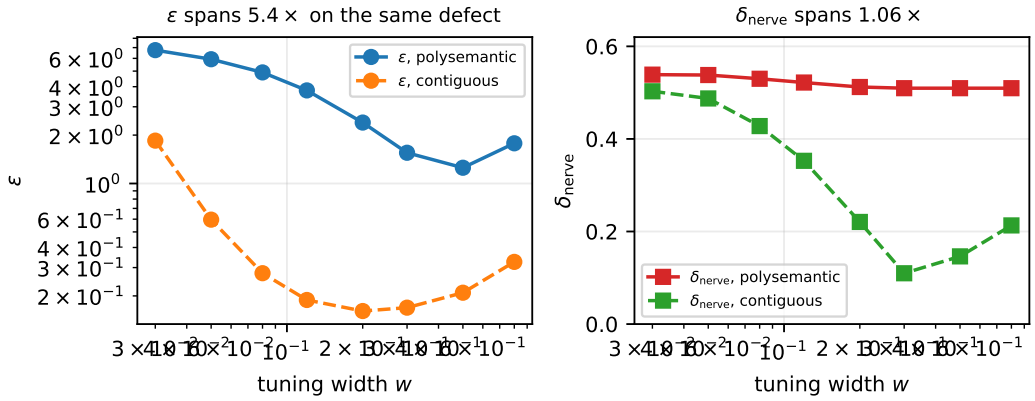

Figure 1: The same antipodal defect, across a $27\times$ change in tuning width. Left: ϵ spans a factor of 5.4 on the polysemantic codes and is non-monotone on the contiguous ones. Right: δ_{nerve} on the polysemantic codes is flat to within a factor of 1.06, while on contiguous codes it falls monotonically as tuning broadens — the correct verdict in both cases.

4 Nerve faithfulness is stable under a fixed polysemantic defect

Ground truth: 400 samples on a circle, 24 features, each a contiguous arc of angular half-width $w\pi$. The perturbation replaces every feature by the pointwise maximum of its arc and the arc rotated by π — a feature firing on two antipodal regions, the canonical polysemantic feature. The one dial that matters is w, and it has to be swept: at $w = 0.03$ a sample activates 0.7 features and at $w = 0.8$ it activates 19 of 24, and these are different regimes for any co-activation statistic.

A fixed defect receives a stable certificate. On the polysemantic code, δ_{nerve} is 0.510–0.539 and the rank correlation $+0.134$–$+0.169$, essentially constant over the whole range; ϵ is 1.25–6.75. The defect is the same at every width — each feature fires on two antipodal arcs — and only one of the two diagnostics says so with a stable number. A practitioner reading $\epsilon = 1.25$ at $w = 0.5$ and $\epsilon = 6.75$ at $w = 0.03$ has no way to tell that the underlying pathology is identical; reading $\delta_{\text{nerve}} \approx 0.51$ in both cases, they do.

The base code is where the two diverge in the other direction, and it is worth being precise about it. δ_{nerve} on the contiguous code falls monotonically with width ($0.503 \to 0.110$) and the rank correlation rises monotonically to $+0.997$: a broadly tuned monosemantic code is a metrically faithful one, which is the correct verdict. ϵ on the same codes is *non-monotone*, with a minimum of 0.161 at $w = 0.2$ and rising to 1.847 at $w = 0.03$ and 0.325 at $w = 0.8$ — so a sparse monosemantic code, which is what interpretability wants, scores worse under coherence than a broadly tuned one. That is a defect of the objective and not of the code.

Coherence couples the defect to tuning width. The exact variance identity makes the mechanism visible. Coherence penalises the dispersion of the columns a sample co-activates. Under antipodal duplication a sample at θ co-activates features centred near θ *and* near $\theta - \pi$; those columns are far apart in sample space, so dispersion is created and all four quantities rise together (we measured the two variance and two covering terms moving from 0.15–0.17 to 1.07–1.51 at $w = 0.3$). The amount of dispersion created depends on how many features a sample co-activates, which is exactly why the response scales with width. δ_{nerve} does not have this dependence because it compares a metric with a metric: once two antipodal samples have high co-activation, $d_{\mathfrak{D}}$ says they are close and d_X says they are maximally far, and that disagreement is $O(1)$ regardless of width.

Antipodal tuning appears in rotated-MNIST models. Gaukstad et al. (2026) report, for the two-digit rotated-MNIST experiment at $\lambda_{\text{Coh}} \in \{5 \times 10^{-3}, 10^{-2}\}$, mean resultant length 0.15 and 0.12 but MRL_{180} of 0.74 and 0.76: the features are tuned modulo π, firing at θ and $\theta + \pi$. They introduce MRL_{180} precisely to score around this, and note that networks "often encode antipodal

Table 1: Nerve faithfulness separates the objectives where the published metrics do not. On rotated MNIST coherence moves the correlation by 0.48 (mnist1) and 0.45 (mnist2) relative to the unregularised baseline — roughly 20 seed standard deviations in each case — while its stated interleaving guarantee is void on exactly these models. On the two-circles set all three objectives are already faithful and the diagnostic correctly declines to separate them.

data	objective	Spearman	δ_{nerve}	ϵ	MRL	% tuned
mnist1	vanilla	-0.180 ± 0.004	0.500 ± 0.029	4.79 ± 1.67	0.539	57.0
mnist1	L^1	$+0.073 \pm 0.022$	0.487 ± 0.015	7.08 ± 1.17	0.592	64.6
mnist1	COH	$\mathbf{+0.301 \pm 0.003}$	0.516 ± 0.026	0.41 ± 0.03	0.886	100.0
mnist2	vanilla	-0.138 ± 0.056	0.534 ± 0.020	3.47 ± 1.16	0.342	21.6
mnist2	L^1	$+0.036 \pm 0.006$	0.520 ± 0.020	3.99 ± 0.89	0.424	36.6
mnist2	COH	$\mathbf{+0.313 \pm 0.022}$	0.502 ± 0.011	0.36 ± 0.05	0.184	2.7
two circles	vanilla	$+0.774 \pm 0.076$	0.471 ± 0.104	3.63 ± 1.97	0.176	3.3
two circles	L^1	$+0.836 \pm 0.061$	0.408 ± 0.066	8.70 ± 3.73	0.248	14.8
two circles	COH	$+0.816 \pm 0.022$	0.397 ± 0.092	0.10 ± 0.00	0.161	0.0

angles together". What Table 5 adds is that the coherence value attached to such a model is not comparable across tuning widths, so it cannot be used to rank models on this axis.

5 EXPERIMENTS

Coherence partially improves nerve faithfulness. Five seeds per cell, identical architecture and schedule, with the rank correlation between $d_{\mathfrak{D}}$ of the code and the data metric reported alongside the published metrics:

On the toy data every objective scores high, because the geometry is dominated by two well-separated clusters that any code captures; the metric discriminates on the harder datasets.

The hinge alone collapses the code. Our first trainable version was the direct surrogate for the certificate: a top-k hinge on the slack, $k = 256$ sampled pairs out of $m(m-1)/2 = 73{,}536$. It destroyed the representation — MSE 3.3×10^{-2} against 9.7×10^{-4} for the unregularised baseline, only 1.9% of activations non-zero — and, diagnostically, *its own* metric got worse (δ_{nerve} 0.510 against 0.379 for plain COH). A hinge on k pairs is an extremely sparse penalty: a mostly dead code has few *extreme* violations while being globally wrong. The correlation form is scale-free and collapse-averse (a constant $d_{\mathfrak{D}}$ gives correlation 0, the maximum loss) and retains the discrimination of Table 5:

code	$1-$Pearson	$+$ hinge	δ_{nerve}
contiguous arcs	0.088	0.222	0.188
arc + antipodal	0.801	1.535	0.787
arc + 120°	0.919	1.552	0.687
i.i.d. random	1.001	2.066	1.478
collapsed code	1.000	1.738	—

The correlation form avoids this collapse. Hinge-only runs serve as an ablation, and later experiments compare the correlation form with and without the hinge.

A 2-torus exposes headroom between objectives. Table 1 shows that on the two-circles set every objective is already faithful ($+0.774$ to $+0.836$), so it cannot separate them — which means it is the wrong place to test a new objective. We therefore use a 2-torus in $\mathbb{R}^3$, where the unregularised code reaches only $+0.368$. Three seeds, matched architecture and schedule:

The nerve objective improves the quantity it optimizes (δ_{nerve} 0.544 against 0.565 and 0.591; mean slack 0.139 against 0.157 and 0.160) *and* on rank faithfulness, which is not the same quantity and is the one an interpretability user cares about. It buys this at no reconstruction cost — MSE 0.1706 against 0.1699 and 0.1687, a difference smaller than the seed spread on ϵ. And it leaves ϵ alone (1.93 against vanilla's 1.89), while coherence more than halves ϵ (0.745) and moves faithfulness by

Table 2: Objectives on a 2-torus in $\mathbb{R}^3$, three seeds, matched schedule (70 epochs, $n = 1500$, latent 64). The correlation form of the nerve objective improves rank faithfulness by $+0.234$ over unregularised and $+0.198$ over coherence and also attains the lowest δ_{nerve}, at reconstruction error matched to four decimal places. It does *not* improve ϵ, and coherence does not improve faithfulness much: the two objectives target different quantities, exactly as §6 predicts.

objective	Spearman	δ_{nerve}	mean slack	ϵ	MSE
vanilla	$+0.368 \pm 0.007$	0.591 ± 0.035	0.1598 ± 0.0010	1.887	0.1699 ± 0.0002
COH	$+0.404 \pm 0.034$	0.565 ± 0.035	0.1565 ± 0.0020	**0.745**	0.1687 ± 0.0011
nerve (corr)	$\mathbf{+0.602 \pm 0.020}$	$\mathbf{0.544 \pm 0.027}$	$\mathbf{0.1391 \pm 0.0022}$	1.926	0.1706 ± 0.0005

Table 3: Rotated MNIST, three seeds per cell. NERVE+COH with the correlation form attains the best Spearman, the best δ_{nerve}, the best slack *and* the best reconstruction on both datasets, at an ϵ within 20% of coherence's. Combining the correlation and top-k hinge forms instead degrades reconstruction by 8–10$\times$ and raises ϵ to as much as 633.

data	objective	Spearman	δ_{nerve}	slack	ϵ	MSE
mnist1	vanilla	-0.183 ± 0.015	0.520	0.0677	5.84	0.00746
mnist1	L^1	$+0.042 \pm 0.031$	0.453	0.0587	5.86	0.00670
mnist1	COH	$+0.301 \pm 0.003$	0.516	0.1254	**0.40**	0.00695
mnist1	nerve/corr	$+0.555 \pm 0.024$	0.347	0.0433	1.59	0.00790
mnist1	nerve/both	$+0.643 \pm 0.005$	0.503	0.0957	0.01	0.05753
mnist1	NERVE+COH/corr	$\mathbf{+0.826 \pm 0.004}$	**0.329**	**0.0313**	0.49	**0.00665**
mnist2	vanilla	-0.157 ± 0.022	0.565	0.0687	3.14	0.00657
mnist2	L^1	$+0.024 \pm 0.012$	0.497	0.0617	3.77	0.00568
mnist2	COH	$+0.313 \pm 0.022$	0.502	0.1158	**0.49**	0.00611
mnist2	nerve/corr	$+0.627 \pm 0.002$	0.390	0.0407	2.01	0.00554
mnist2	nerve/both	$+0.550 \pm 0.026$	0.444	0.0617	633.7	0.06267
mnist2	NERVE+COH/corr	$\mathbf{+0.815 \pm 0.004}$	**0.341**	**0.0337**	0.52	**0.00497**

only $+0.036$. The The two objectives therefore measure distinct properties and are most informative when reported jointly.

On the two-circles set the same runs give $\delta_{\text{nerve}} = 0.196$ for the nerve objective against 0.278 for coherence and 0.378 for unregularised — a 30% improvement over coherence on its own metric — with rank correlations statistically indistinguishable ($+0.785$ against $+0.821$), which is what a saturated dataset should look like.

Joint nerve and coherence training performs best on rotated MNIST. We next evaluate the correlation form on rotated MNIST using sixty matched runs, three seeds per cell, and $\lambda_{\text{nerve}} = 10^{-2}$ (Table 3). The strongest result comes from combining nerve faithfulness with coherence: the joint objective improves the data-geometry certificate while retaining the angular locality induced by coherence.

On mnist1, NERVE+COH raises rank faithfulness from coherence's $+0.301$ to $+0.826$ — a factor of 2.7, and 130 coherence standard deviations — while *also* attaining the lowest certificate (0.329 against 0.516), the lowest mean slack (0.031 against 0.125) and the lowest reconstruction error of any objective in the table, at an ϵ of 0.49 against coherence's 0.40. mnist2 is the same picture. All three seeds agree to within 0.009 in Spearman.

Topological faithfulness trades off with angular tuning. Mean resultant length — the original paper's tuning metric — falls from 0.887 under coherence to 0.401 ± 0.003 under NERVE+COH on mnist1. So the objective that best certifies the code against the data it encodes is *not* the objective that produces the most sharply angularly tuned units, and a reader who wants both must choose. The companion analysis of the floor reaches the same conclusion from the other direction: *tuning sharpness and topological faithfulness are different objectives, and coherence optimises neither of them well.*

The hinge degrades large-scale training. The `both` mode adds the top-k hinge to the correlation term. On synthetic data the addition is neutral; on MNIST it degrades reconstruction by 8–10$\times$ (0.0575 and 0.0627 against 0.0067 and 0.0055) and sends ϵ to 633 on mnist2 and 8048 on two circles. We recommend the correlation form alone.

Language-model dictionaries. TopK dictionaries on GPT-2 layer 6 ($k = 32$, widths 2^{13} and 2^{14}, FVU 0.136–0.155) have $\delta_{\text{nerve}} = 0.576$ and 0.594 unregularised with rank correlation 0.070 and 0.084: a dictionary that reconstructs the residual stream well can have a co-activation geometry essentially uncorrelated with the geometry it encodes. Coherence at $\lambda = 3 \times 10^{-2}$ cuts δ_{nerve} by 38–41% (to 0.357 and 0.349) and ϵ by 75–77% (from 139 and 198 to 35 and 45) for 1–2% of FVU — but leaves the rank correlation near 0.1. Both evaluations need the sparse co-activation kernel; the dense form returns `nan` at these widths, because the (m, n, n) tensor it wants is 137 GB.

The low-dimensional regime. The token embedding of a masked language model is the one setting where the coherence guarantee is not vacuous by intrinsic dimension (3–5), so it is where the two objectives should agree most. They do: over three seeds per variant, coherence cuts δ_{nerve} from 0.408 (unregularised) and 0.389 (softplus) to 0.252, and adding L^1 to 0.237 — a 38–42% reduction in the certificate. The rank correlation stays small in absolute terms ($0.026 \to 0.068$), which is the same decoupling §6 quantifies: the certificate and the ordering move independently even where both improve.

6 Worst-case distortion and rank faithfulness measure different properties

δ_{nerve} is the maximum pointwise slack $|d_{\mathfrak{D}} - d_X|$ required by the interleaving certificate. To separate worst-case distortion from average geometric agreement, we compare it with mean slack, the 99.9th percentile, and rank correlation $\mathrm{Sp}(d_{\mathfrak{D}}, d_X)$ across 124 trained codes. The statistics occupy distinct evaluation roles.

Across the 124 trained codes, the slack statistics and rank faithfulness separate cleanly into different evaluation roles.

Mean slack and rank faithfulness can move in opposite directions. Pooled mean slack correlates $+0.14$ with faithfulness, and within MNIST1 the correlation is $+0.46$. Vanilla and L^1 codes on MNIST have lower mean slack (0.057–0.074 versus Coh's 0.10–0.14) but much lower rank correlation (-0.20 to $+0.08$ versus $+0.29$ to $+0.34$). Small average error can therefore coexist with poor preservation of metric order.

The ratio $\max/\mathrm{mean}$ has the strongest pooled correlation of the four (-0.57), but its sign reverses across dataset families: -0.56 and -0.49 on the two MNIST families, versus $+0.45$ and $+0.62$ on the two circle families. The pooled correlation is therefore driven by between-dataset differences rather than a stable within-family relationship.

δ_{nerve} and rank correlation measure complementary aspects of geometry. The former is a theorem-backed worst-case metric distortion (Theorem 2); the latter measures average order agreement. Their within-family correlation ranges from $+0.15$ to -0.58, so neither is a reliable proxy for the other.

The mismatch is larger, not smaller, at scale. The GPT-2 measurement uses two dictionaries at one layer. We extend it to Gemma Scope (Lieberum et al., 2024b), a JumpReLU suite for Gemma-2-2B at width 2^{14}, on nine layers spanning the network. For each layer we take the released sparsity level whose realised mean support is closest to 32, restrict to the 1024 most active alive features, and evaluate on 3500 residual-stream tokens with the initial token excluded.

The code–data rank correlation does not improve with model scale; it degrades. It runs from -0.386 at layer 2 through -0.095 at layer 12 to $+0.197$ at layer 22, negative at the six shallowest layers tested and never above 0.2. The certificate reads $\delta_{\text{nerve}} = 0.450$–$1.574$ over the same layers. Where GPT-2 dictionaries showed a small positive correlation (0.070–0.084), a 2B-parameter model's released dictionaries show essentially none, and in the first half of the network the co-activation metric is mildly *anti*-correlated with the activation metric it is supposed to track.

This strengthens rather than qualifies the claim the GPT-2 experiment supports. Reconstruction quality and sparsity are the two numbers a released dictionary is usually published with, and neither constrains this axis at either model scale. The certificate is the cheaper half of the diagnostic — $O(mk^2)$ for a k-sparse code — and it is available for any released dictionary without retraining.

7 Related work

Four comparisons, not one. A non-negative code invites several topological comparisons that are easy to conflate, and they are not interchangeable. (i) The row and column point clouds of a single code can be compared under Vietoris–Rips filtrations, where agreement is a property to be certified because it can fail. (ii) The same two sides can be compared in the relational geometry of the code itself, where Dowker duality makes the comparison exact for every matrix and no certificate is needed. (iii) A code can be compared against the data it encodes, which asks whether the co-activation geometry tracks the activation geometry. (iv) Two codes on shared samples can be compared against each other, with no common ambient space. All four share the four-condition structure of a δ-interleaving, the filtrations differ, and none of the four implies another: a code can be perfectly self-dual and encode its data badly, or track its data well and correspond to a second code hardly at all. This paper is about (iii).

Topological objectives, and which topology they constrain. Topological auto-encoders (Moor et al., 2020) and their successors (Trofimov et al., 2023) preserve the topology of the *input* in the latent sample space; they say nothing about the feature space. Persistence-based losses (Gabrielsson et al., 2020; Carrière et al., 2021) and persistence features (Hofer et al., 2017; Carrière et al., 2020; Adams et al., 2017) sit in the same category. Coherence (Gaukstad et al., 2026) is unusual in constraining both sides at once — which is why it is worth asking what each half buys. Our answer is that the row/column half is free in the right complex (Dowker, 1952; Chowdhury & Mémoli, 2018), and the half that is not free is the one comparing the code with the data, which is exactly what δ_{nerve} measures. Reviews: Hensel et al. (2021), Chazal et al. (2016).

Nerves and covers. Reading a code as a cover and its complex as a nerve is the classical route from a relation to a space (Dowker, 1952; Björner, 1995), and underlies the neuroscience use of "cell groups" to recover stimulus topology (Curto & Itskov, 2008; Giusti et al., 2015; Singh et al., 2008). The co-activation kernel we use is exactly the Dowker 1-skeleton, which is why the certificate needs no hypothesis; the stability that makes the certificate meaningful is Cohen-Steiner et al. (2007) and Chazal et al. (2009).

Polysemanticity and what metrics can see it. Polysemantic features are the central obstacle in dictionary interpretability (Elhage et al., 2022; Olah et al., 2020; Bricken et al., 2023; Templeton et al., 2024), and antipodal encoding — a feature firing at θ and $\theta + \pi$ — is the canonical case in circular stimulus spaces. Feature splitting and absorption (Chanin et al., 2024; Templeton et al., 2024) are the analogous pathologies at scale, and benchmark suites (Karvonen et al., 2025; Makelov et al., 2024) evaluate dictionaries on reconstruction, sparsity and downstream probes. None of those diagnostics is metric: a dictionary can reconstruct to FVU 0.14 and have a co-activation geometry uncorrelated with the geometry it encodes, which is the failure δ_{nerve} is built to expose. Architectural work (Gao et al., 2024; Rajamanoharan et al., 2024a;b; Lieberum et al., 2024a) is complementary: it improves the reconstruction–sparsity frontier, not the metric one.

Tuning metrics. Mean resultant length and its antipodal variant are circular-statistics measures of how sharply a unit is tuned, in the tradition of Ben-Yishai et al. (1995); Zhang (1996). The trade we report — that the objective which best certifies the code against the data is not the one that produces the most sharply tuned units — is a concrete instance of the general point that representational quality is not one-dimensional (Locatello et al., 2019; Schaeffer et al., 2022).

Metric and topological faithfulness are complementary. δ_{nerve} is a statement about metric agreement, and a sufficiently fine partition of a manifold approximates its metric well while carrying no topology. Directly minimising coherence on a circle under a faithfulness tolerance of 0.20 produces a 24-block partition — $\epsilon = 10^{-4}$, block deviation 0.009, hence topologically trivial — that

still scores rank correlation 0.804 against the true circle metric. Combining δ_{nerve} with a topology check (component count or the top H_1 class) separates metric approximation from preservation of nontrivial topology. Experiments that optimize coherence under a topology constraint therefore use connectivity directly rather than metric faithfulness as a surrogate.

AI USE STATEMENT

We used GPT solely to polish the writing.

REPRODUCIBILITY STATEMENT

The identity between the co-activation kernel and the Dowker 1-skeleton, and the interleaving certified by δ_{nerve}, are proved where they are stated. Per-run tables for the tuning-width sweep and the antipodal duplication experiment are in the appendix, with seed counts. Training runs report the objective variant explicitly: the results we highlight use the correlation form, and the alternative form is reported separately because it is seed-sensitive. All results come from public artifacts and deterministic procedures. The GPT-2 dictionaries are the released `GPT2-Small-SAEs-Reformatted` suite at `blocks.{L}.hook_resid_pre`. The Gemma-2-2B dictionaries are Gemma Scope residual-stream SAEs at width 2^{14}; two conventions there change the numbers by more than an order of magnitude and are worth stating explicitly. The suite's `layer_L` corresponds to the residual stream *after* block L, and the initial token must be excluded: its residual norm is $12.2\times$ the mean of the remaining tokens in our measurements, and retaining it inflates the fraction of unexplained variance from 0.17 to 6.9. Where a single sparsity level is used we take the released level whose realised mean support is closest to 32 and restrict to the 1024 most active alive features, matching the GPT-2 protocol.

REFERENCES

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18 (8):1–35, 2017.

Rani Ben-Yishai, R. Lev Bar-Or, and Haim Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

Anders Björner. Topological methods. In *Handbook of Combinatorics*, pp. 1819–1872. Elsevier, 1995.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. Perslay: A neural network layer for persistence diagrams. In *International Conference on Artificial Intelligence and Statistics*, pp. 2786–2796, 2020.

Mathieu Carrière, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hariprasad Kannan, and Yuhei Umeda. Optimizing persistent homology based functions. In *International Conference on Machine Learning*, pp. 1294–1303, 2021.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Symposium on Computational Geometry*, pp. 237–246, 2009.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. Springer, 2016.

Samir Chowdhury and Facundo Mémoli. A functorial dowker theorem and persistent homology of asymmetric networks. *Journal of Applied and Computational Topology*, 2(1):115–175, 2018.

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

Carina Curto and Vladimir Itskov. Cell groups reveal structure of stimulus space. *PLoS Computational Biology*, 4(10):e1000205, 2008.

C. H. Dowker. Homology groups of relations. *Annals of Mathematics*, 56(1):84–95, 1952.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Rickard Brüel Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, and Primoz Skraba. A topology layer for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1553–1563, 2020.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Sigurd Gaukstad, Melvin Vaupel, Valdemar Kargård Olsen, Erik Hermansen, and Benjamin Dunn. Learning coherent representations: A topological approach to interpretability. In *International Conference on Machine Learning*, 2026. arXiv:2606.02841.

Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112 (44):13455–13460, 2015.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:681108, 2021.

Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. In *Advances in Neural Information Processing Systems*, 2017.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Arthur Conmy, Callum McDougall, et al. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *arXiv preprint*, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint*, 2024a.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024b.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124, 2019.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations*, 2024.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *International Conference on Machine Learning*, pp. 7045–7054, 2020.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, 2020.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint*, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint*, 2024b.

Rylan Schaeffer, Mikail Khona, and Ila Fiete. No free lunch from deep learning in neuroscience: A case study through models of the entorhinal–hippocampal circuit. In *Advances in Neural Information Processing Systems*, 2022.

Gurjeet Singh, Facundo Mémoli, Tigran Ishkhanov, Guillermo Sapiro, Gunnar Carlsson, and Dario L. Ringach. Topological analysis of population activity in visual cortex. *Journal of Vision*, 8(8):11, 2008.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Ilya Trofimov, Daniil Cherniavskii, Eduard Tulchinskii, Nikita Balabin, Evgeny Burnaev, and Serguei Barannikov. Learning topology-preserving data representations. In *International Conference on Learning Representations*, 2023. arXiv:2302.00136.

Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

Table 4: Correlation of each slack statistic with rank faithfulness $\mathrm{Sp}(d_{\mathfrak{D}}, d_X)$ over 124 trained codes. A statistic usable as a proxy would be strongly negative everywhere. The pooled column hides a sign flip: $\max/\mathrm{mean}$ is -0.5 on the MNIST families and $+0.5$ on the circle families.

	pooled	within family			
statistic	n=124	mnist1	mnist2	2-circles	mnist1-λ
$\max$ (δ_{nerve})	-0.51	$+0.15$	-0.25	-0.25	-0.58
$q_{99.9}$	$+0.21$	—	—	—	—
mean |slack|	$+0.14$	$+0.46$	$+0.25$	-0.72	-0.61
$\max/\mathrm{mean}$	-0.57	-0.56	-0.49	$+0.45$	$+0.62$

Table 5: Antipodal duplication across tuning width, six seeds. Subscript 0 is the contiguous code and a the antipodally duplicated one. δ_{nerve} on the polysemantic code is 0.510–0.539 throughout — a factor of 1.06 across a $27\times$ change in width — and the rank correlation is pinned at $+0.134$. ϵ on the same codes ranges over a factor of 5.4.

w	co-act.	ϵ_0	ϵ_a	$\times$	δ_0	δ_a	$\times$	Sp_0	Sp_a
0.03	0.7	1.847	6.751	3.7	0.503	0.539	1.07	$+0.064$	$+0.169$
0.05	1.2	0.596	5.928	9.9	0.487	0.538	1.10	$+0.166$	$+0.136$
0.08	1.9	0.277	4.907	17.7	0.428	0.530	1.24	$+0.313$	$+0.135$
0.12	2.9	0.189	3.792	20.1	0.353	0.522	1.48	$+0.484$	$+0.134$
0.20	4.8	0.161	2.393	14.9	0.221	0.512	2.32	$+0.732$	$+0.134$
0.30	7.2	0.169	1.553	9.2	0.110	0.510	4.64	$+0.911$	$+0.134$
0.50	12.0	0.209	1.253	6.0	0.146	0.510	3.48	$+0.997$	$+0.134$
0.80	19.2	0.325	1.778	5.5	0.213	0.510	2.39	$+0.997$	$+0.134$

A SLACK STATISTICS AGAINST RANK FAITHFULNESS: THE FULL TABLE

B ANTIPODAL DUPLICATION ACROSS TUNING WIDTH: THE FULL TABLE