Memorization Is Folding: Topological Signatures of Noisy-Label Learning

Zhongtian Sun[1,*] **Fan Mo**[2] **Prayag Tiwari**[3] **Kelin Xia**[4]
[1]University of Kent, United Kingdom
[2]University of Oxford, United Kingdom
[3]Halmstad University, Sweden
[4]Nanyang Technological University, Singapore
[*]Corresponding author: zs2566@kent.ac.uk

Abstract

How does noisy supervision change the representations learned by a neural network? Training accuracy measures agreement with training labels but provides little information about representation geometry. We use persistent homology to track loop structure across distance thresholds in neural activations. Experiments span image classification, text classification, and modular arithmetic. On a fixed CIFAR-10/ResNet-18 training probe, equally sized clean-label and corrupted-label subsets show no detectable persistence difference at epoch 10, whereas corrupted-label subsets have higher persistence at epoch 50. The contrast survives uniform-scale controls, even though predictions match the assigned noisy labels for only 7–11% of corrupted probe examples at that checkpoint. Normalisation preserves the crossover between clean and noisy training trajectories on CIFAR-10/ResNet-18, whereas CIFAR-100/ResNet-18 shows no crossover under any of the five evaluated constructions. In a separate noise-rate calibration comparison across five pipelines, mean activation norm achieves a higher mean in-sample R^2 than raw persistence (0.936 versus 0.908). These findings show that noisy supervision leaves a measurable geometric signature even when agreement with corrupted labels is low. Tracking scale-controlled persistence alongside label fitting offers a way to study how networks reorganise their internal representations during training. This provides a basis for developing training diagnostics that complement accuracy.

1 Introduction

Neural networks can fit incorrect labels, but training accuracy alone reveals little about how noisy supervision changes their representations. Examples with conflicting labels may occupy nearby regions of representation space, raising the question of how their geometry evolves during training. This question connects to studies of memorization, intrinsic dimension, the information bottleneck, neural collapse, and grokking [Zhang et al., 2016, Arpit et al., 2017, Ansuini et al., 2019, Shwartz-Ziv and Tishby, 2017, Papyan et al., 2020, Power et al., 2022]. Persistent homology offers a way to examine these changes through loop structure in activation point clouds. We use *folding* as a geometric intuition for how training with conflicting labels may rearrange neighbourhoods in representation space.

As the pairwise distance threshold increases, points connect, loops form, and those loops eventually fill in. Each loop persists over an interval of thresholds, and total persistence sums the lengths of these intervals. Its interpretation requires attention to activation scale: uniformly enlarging a cloud increases raw persistence while preserving its shape. The analysis therefore combines persistence with activation scale, normalised point clouds, and geometric measurements that do not use topology.

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

Within the same trained model, how does persistence differ between subsets whose labels were changed and subsets whose labels were left unchanged? Which differences remain after controlling for uniform rescaling?

On a fixed CIFAR-10 training probe, we compare equally sized groups of images whose labels were changed or left intact within each ResNet-18 run. We detect no persistence difference at epoch 10. At epoch 50, persistence is higher in the group with changed labels across eight seeds at each of two noise settings, and the difference survives controls for uniform rescaling. At that checkpoint, among all images with changed labels in the probe, only 7–11% of predictions match the assigned incorrect label, while 71–79% match the original label.

The broader experiments distinguish estimating noise amount from characterising its effects on representations. Across five pipelines, mean activation norm achieves a higher mean in-sample R^2 than raw persistence for noise-rate calibration. Scale-controlled persistence instead measures differences that survive uniform rescaling. Normalisation preserves clean–noisy crossover on CIFAR-10/ResNet-18, whereas CIFAR-100/ResNet-18 shows no crossover under the five evaluated constructions. An exploratory comparison of random and class-pair relabelling further examines responses to conditions that differ in both realised corruption rate and relabelling structure. These results motivate interpreting persistence in relation to the training stage, activation scale, and comparison being made.

The contributions are:

1. **A training-subset contrast under noisy supervision.** A fixed-probe comparison identifies higher persistence in corrupted subsets late in training, including under uniform-scale controls, while predictions for most corrupted examples still match their original labels. This provides a group-level view of the response to corruption alongside direct measurements of label fitting.
2. **An analysis of activation scale and loop structure.** Normalised point clouds, persistence ratios, and local geometric measurements distinguish differences that survive uniform rescaling from responses sensitive to scale. This clarifies how geometric changes contribute to the observed persistence response.
3. **Evidence for choosing measurements by task and pipeline.** Crossover comparisons show that persistence trajectories depend on the pipeline and distance construction. A common calibration protocol finds a stronger average in-sample fit for activation norm than raw persistence, distinguishing the measurement of noise amount from the analysis of its representational effects.

2 Setup and Observable

Let $\mathbf{z}_i(t) \in \mathbb{R}^d$ denote the penultimate representation of sample i at epoch t, extracted in evaluation mode. The trajectory analysis uses test-set activations to track representation changes on held-out examples during training with noisy labels. These trajectories describe the response on held-out data rather than directly localizing memorization to corrupted training examples. The fixed training probe in Section 3.1 compares subsets with clean and corrupted labels within the same trained model. Computing global persistence does not require per-sample corruption labels, whereas the subset comparison uses known corruption membership.

We compute Vietoris–Rips persistent homology with Ripser [Bauer, 2021], using `maxdim` $= 1$ and no explicit filtration cap. At each distance threshold, the Rips complex connects points within that distance and fills each clique with a simplex. H_0 describes connected components, while H_1 describes loops that are not filled by triangles. A persistence bar (b, d) records the thresholds at which a feature appears and disappears. Our raw observable is

$$\Phi(t) = \sum_{(b,d)\in\mathcal{F}_1(t)} (d - b), \tag{1}$$

where $\mathcal{F}_1(t)$ contains the finite H_1 bars whose lifetimes exceed $0.01L_{\max}(t)$, and $L_{\max}(t)$ is the longest finite H_1 lifetime at that checkpoint. The sum is zero if no bar is retained. For analyses using this filtered observable, the relative cutoff is fixed across runs, noise levels, and epochs. Some original extensions use unfiltered finite lifetimes; these exceptions are specified in Appendix B. The

Table 1: Notation. Normalised constructions are named explicitly when reporting results.

Symbol	Meaning
$Z(t)$	Probe activation cloud at epoch t.
$\Phi(t)$	Filtered total finite H_1 lifetime on raw activations, unless qualified.
Φ_{H_0}	Sum of finite H_0 lifetimes, excluding the infinite bar.
$\mathcal{F}_1, L_{\max}$	Retained H_1 bars and maximum finite H_1 lifetime, respectively.
sf	Selection fraction for symmetric uniform relabelling
$\Delta\Phi$	Noisy minus clean persistence under the same construction.
$t^\star$	First sampled checkpoint from which noisy persistence remains above its paired clean reference for all seeds and subsequent checkpoints, after initially lying below it.

relative cutoff excludes bars with short relative lifetimes without assuming that they are sampling noise. With no explicit filtration cap, feature deaths are not truncated at a manually chosen distance.

Total persistence accounts for both the number of retained loops and their lifetimes, capturing how long each loop persists across distance thresholds rather than counting loops alone. We focus on H_1 to study loop structure and compare total persistence with alternative summaries and geometric measurements that do not use topology.

Separating scale and structure. The raw analysis uses Euclidean distances between unnormalised activations. Persistence is also computed using unit-normalised activations, cosine dissimilarity, per-feature standardization (z-score), and centered root-mean-square (RMS) rescaling. Unit normalisation and cosine dissimilarity emphasize angular relations, centered RMS rescaling removes a common scale, and z-score standardization adjusts relative feature scales. These are sensitivity analyses, not interchangeable observables. The construction is specified for each result.

Additional summaries include the number of retained H_1 bars, Φ_{H_1}/Φ_{H_0}, and $\Phi/\operatorname{diam}(Z)$, where $\Phi_{H_1} = \Phi$ and $\operatorname{diam}(Z)$ is the maximum pairwise distance under the stated construction. When their denominators are nonzero, both ratios are invariant to uniform positive rescaling, but not to arbitrary anisotropic transformations.

The distinction matters even within a single trained model. Multiplying every activation by the same positive constant multiplies all raw Euclidean birth and death scales by that constant. Since the lifetime cutoff is relative to the longest bar, this operation preserves which bars pass the cutoff while scaling their summed lifetime. An increase in raw Φ can therefore occur without a change in the ordering of pairwise distances or in the loop structure across corresponding scales. Conversely, changing relative distances can alter persistence after a common scale is removed. The analyses below use normalisation to test sensitivity to representation scale and compare relabelling conditions with their realised corruption rates stated explicitly.

Sampling and global versus per-class analysis. The default global computation uses $n = 300$ uniformly sampled activation vectors. Global PH is the primary analysis because it includes cross-class structure. In the classwise protocol, per-class PH is computed only for datasets with at most 20 classes, using classes with at least 10 probe examples and a sampling budget of 50 points per class. Per-class summaries are summed. This rule was introduced after the original results were available; the original classwise analyses use the separate protocols described in Appendix B.

Per-class results use the point counts specified in their appendices, and must not be compared numerically with global sums as if point counts and aggregation were matched. The original experiments used different saved-probe sizes for some extensions; a larger per-class sum alone does not establish a stronger diagnostic. At $n = 300$, the original subsampling study gives rank fidelity $\rho = 0.973$ against a 2,000-point reference across 433 configurations, mean subsampling CV 0.022, and runtime 0.058 s per call versus 2.944 s at 2,000 points. A separate sensitivity analysis is reported in Appendix C.

The original suite spans MNIST [LeCun et al., 1998], FashionMNIST [Xiao et al., 2017], CIFAR-10/100 [Krizhevsky and Hinton, 2009], AG News [Zhang et al., 2015], Tiny-ImageNet [Deng et al., 2009], ImageNette [Howard, 2019], and modular arithmetic [Power et al., 2022]. Architectures include MLPs, ResNets [He et al., 2016], ViTs [Dosovitskiy et al., 2020], DistilBERT [Sanh et al.,

Table 2: Corrupted/clean persistence ratios at epoch 50 on the fixed CIFAR-10/ResNet-18 training probe, with eight seeds per noise setting. Each entry is the median corrupted-subset statistic divided by the median clean-subset statistic across seeds. Corrupted subsets have larger values in all eight seeds except for diameter-normalised persistence at $sf = 0.25$, where the direction agrees in seven seeds. Exact two-sided paired Wilcoxon tests give $p = 0.0078$ for the eight-of-eight comparisons and $p = 0.0156$ for the seven-of-eight comparison.

sf	Raw Φ	Φ_{H_1}/Φ_{H_0}	$\Phi/\operatorname{diam}(Z)$	Loop count
0.25	1.29	1.23	1.44	1.24
0.50	1.37	1.28	1.46	1.36

2019], and small Transformers [Vaswani et al., 2017]. Noise conditions include symmetric uniform relabelling, Patrini class-pair relabelling, cyclic asymmetric noise, a pixel-distance-based instance-dependent construction in the ViT experiment, and CIFAR-10N human annotations [Patrini et al., 2017, Wei et al., 2021]. The intended instance-dependent condition for ResNet-18 is analysed as a random-relabeling condition in Section 5.

The original suite reports over 900 training runs and 8,000 PH computations. The normalisation reanalysis covers 1,540 saved activation files. These counts refer to different stages of the study, not additional independent training runs. Hyperparameters and experiment-specific seed counts appear in the appendices.

3 Early Response and Late Training-Set Evidence

At epoch 10, raw persistence is associated with noise level in the evaluated MLP, ResNet-18, ViT-Tiny, and DistilBERT pipelines (Appendix D). In the separate five-pipeline calibration comparison, mean activation norm achieves a higher mean in-sample R^2 than raw persistence, 0.936 versus 0.908 in Table 5. Section 5 examines sensitivity to activation scale through normalisation.

The association varies over training and across pipelines. On CIFAR-10/ResNet-18, the correlation between noise level and raw persistence changes from $\rho = -0.969$ at epoch 10 to -0.225 at epoch 50. On MNIST/MLP, it changes from -0.931 to -0.318, while ViT-Tiny/CIFAR-10 retains a correlation of -0.962 at both epochs (Appendix D). These correlations compare models trained at different noise levels. The epoch-10 association alone does not establish that the models have memorised corrupted labels. Crossover is assessed separately by comparing the clean and noisy trajectories over time. The following experiment compares clean and corrupted subsets within the same trained model.

3.1 A fixed training probe

The fixed probe contains 4,000 CIFAR-10 training images, with 400 sampled from each true class using RandomState(12345). The same images are evaluated without augmentation across seeds, conditions, and checkpoints. Under symmetric uniform relabelling at $sf = 0.25$ and 0.5, corruption membership identifies labels that actually changed. The clean and corrupted groups are matched in size, but not in class composition, and PH is computed on 300 points from each group. Each condition uses eight seeds. Paired comparisons use the exact two-sided Wilcoxon signed-rank test.

At epoch 50, corrupted subsets have higher raw persistence and higher values of both scale-invariant ratios (Table 2). The corrupted/clean ratios for raw persistence are 1.29 and 1.37 at the two noise settings. Each summary is higher for corrupted subsets in all eight seeds, except $\Phi/\operatorname{diam}(Z)$ at $\mathrm{sf} = 0.25$, where it is higher in seven seeds.

We detected no difference between the subsets at epoch 10. At epoch 25, only loop count differed, at $sf = 0.25$. At epoch 50, among all images whose labels changed in the fixed training probe, 7–11% of predictions matched the assigned incorrect label and 71–79% matched the original label. Figure 1(a) shows the seed-level persistence ratios across checkpoints. Across six comparisons of global persistence on the test set and the training probe, the smallest p-value was 0.11. In a separate control, we fitted a class-conditional Gaussian to each subset. The excess persistence relative to these controls did not differ detectably between the subsets (all $p \geq 0.25$). Class composition and covariance therefore remain possible explanations for the observed contrast.

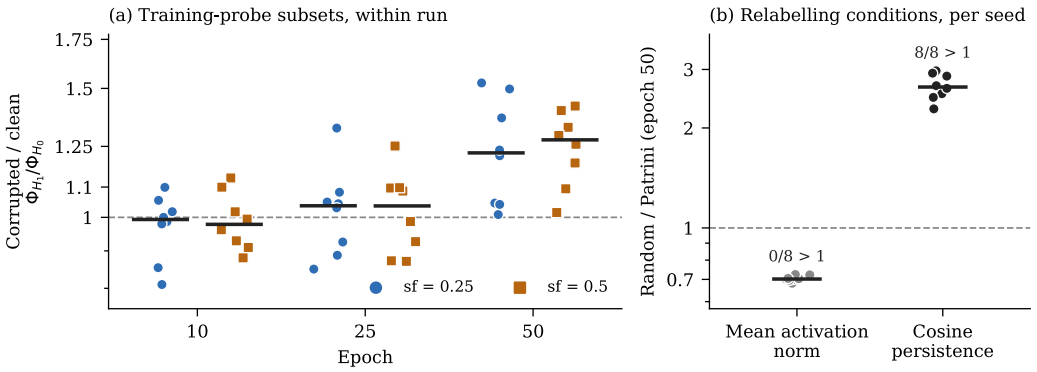

Figure 1: Seed-level persistence comparisons. (a) Corrupted/clean ratios of Φ_{H_1}/Φ_{H_0} on the fixed CIFAR-10/ResNet-18 training probe, with eight seeds per setting. Dots show within-run ratios and horizontal bars show their medians. These differ from the ratios of seed medians reported in Table 2. (b) Epoch-50 random/Patrini ratios for mean activation norm and cosine persistence. Each dot shows one seed's median ratio across three nominal settings. The random arm changes approximately 1.8 times as many labels as the Patrini arm, so the comparison includes differences in both corruption rate and relabelling structure. Both panels use logarithmic vertical axes. Dashed lines mark a ratio of one.

4 Geometric Interpretation and Controlled Comparisons

Conflicting labels can give nearby examples competing training signals. Persistent homology measures loop structure in the sampled activation cloud, but persistence also depends on sampling, class geometry, and distance scale. The following controlled comparisons examine how the placement of label contradictions affects persistence, while the training-probe and normalisation analyses help distinguish these effects.

Scope of the motivating model. Networks composed of affine maps, convolutions, ReLU activations, average or max pooling, and residual addition are piecewise affine. BatchNorm in evaluation mode is a fixed affine map and preserves this property. This argument applies to the corresponding operations in the residual networks studied here, but not to LayerNorm or softmax attention. Piecewise-affine structure alone guarantees neither an H_1 cycle nor a lower bound on its lifetime.

Class-pair controls. We selected class pairs using centroid distances in clean representations. NEAR and FAR each contained five disjoint pairs covering all ten CIFAR-10 classes, with mean centroid distances of 3.32 and 4.94, respectively. Within each original class, we randomly selected half the examples and reassigned them to the paired class. At the prespecified endpoint of epoch 10, the FAR/NEAR ratio for raw persistence was 0.935 ($p = 0.77$). At epoch 50, raw persistence was higher in FAR in all ten seeds, with a FAR/NEAR ratio of 1.445 ($p = 0.002$). The epoch 50 difference in raw persistence may reflect changes in activation scale, relative geometry, or both.

An analysis across class pairs relates centroid distance to the noisy/clean persistence ratio. At $sf = 0.5$, clean centroid distances and noisy/clean ratios are averaged across seeds for each pair before computing the correlation across pairs. At epoch 50, the Spearman correlations are 0.867 for CIFAR-10/ResNet-18, 0.614 for CIFAR-10/ViT-Tiny, and 0.413 for the 60 evaluated CIFAR-100/ResNet-18 pairs. These associations vary over training, as detailed in Appendix U. The appendix also retains the original exploratory experiment, in which pair selection partly used persistence.

Synthetic contradictions and local geometry. The two-dimensional experiment uses two Gaussian clusters with label flips near the boundary (Appendix B). Across 140 individual seed runs with 0–50% of near-boundary labels flipped, the association between flip rate and persistence at epoch 10 is $\rho = 0.33$ ($p = 6.1 \times 10^{-5}$). Averaging within each of the seven flip rates gives $\rho = 1.000$ (exact two-sided $p = 0.0004$). The first analysis captures variation across individual runs, while the second describes the ordering of the rate means. Figure 2 shows the activation projections and representative

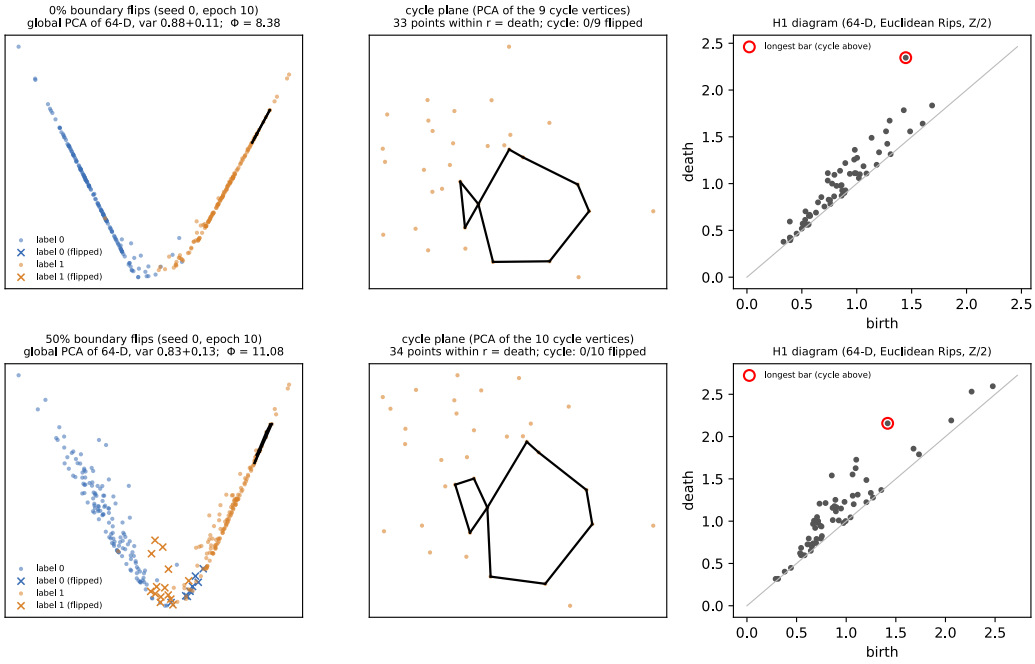

Figure 2: Representation and persistent cycles in the synthetic two-Gaussian task, using seed 0 at epoch 10. The top and bottom rows show 0% and 50% of near-boundary labels flipped, respectively; the boundary region is $|x_0| < 1$. Left: PCA projections of the 300-point subsamples of 64-dimensional activations, fitted separately for each condition. Colour denotes the observed label and crosses mark flipped labels. Black edges show a representative of the longest finite H_1 bar, computed in the original 64-dimensional Euclidean Rips filtration over $\mathbb{Z}/2\mathbb{Z}$. Middle: the same representative projected onto its own first two principal components, with subsample points lying within its death radius of a cycle vertex. Right: persistence diagrams with the represented bar circled. Total persistence is 8.38 and 11.08, respectively. Neither displayed representative contains flipped points. Across 20 seeds, the longest-bar representatives show no detected enrichment for flipped points.

persistent cycles for seed 0 at epoch 10. Total persistence is higher under boundary-label corruption, but the longest-bar representative contains no flipped points in either condition.

Local geometric measurements provide a complementary view. On CIFAR-10/ResNet-18, local PCA dimension with $k = 10$ correlates with noise level at $\rho = 0.946$ at epochs 25 and 50 when the fully randomised condition is excluded ($sf \in \{0, 0.25, 0.5\}$, five seeds). In the exploratory ResNet-18 relabelling comparison, local PCA dimension is higher in the random-relabeling arm in all 24 seed–rate pairs at epoch 50. These pairs comprise eight seeds at three settings, and the arms differ in realised corruption rate as well as relabelling structure. Multiplying activations by 9.3 changes the geodesic-to-Euclidean measure by less than 10^{-9}. Local dimension measures how many directions contribute to neighborhood variation, while persistence describes loop structure across distance thresholds. These measurements characterize the sampled representations without uniquely identifying folding or establishing a new topological hole in the underlying data manifold.

5 Normalisation, Crossover, and Noise Type

Crossover across distance constructions. We report the first persistent crossing: the earliest sampled checkpoint from which noisy persistence remains above its seed-matched clean reference at every subsequent checkpoint for all evaluated seeds, after initially lying below it. On CIFAR-10/ResNet-18 at $sf = 0.5$, this occurs at epoch 20 under cosine, unit normalisation, centered RMS, and z-score constructions, compared with epoch 30 under raw Euclidean distance. On MNIST/MLP, the all-seed criterion is first met at epoch 30 under raw, unit-normalised, and cosine constructions,

Table 3: First persistent crossing under the evaluated constructions: the earliest sampled checkpoint from which all seed-paired noisy trajectories remain above their clean references. "None" denotes no persistent crossing over the evaluated checkpoints. The CIFAR-10/ResNet-18 comparison uses $sf = 0.5$.

Pipeline	Seeds	Raw	Unit / cosine	RMS / z-score
CIFAR-10 / ResNet-18	5	30	20	20
MNIST / MLP	5	30	30	None
CIFAR-100 / ResNet-18	3	None	None	None

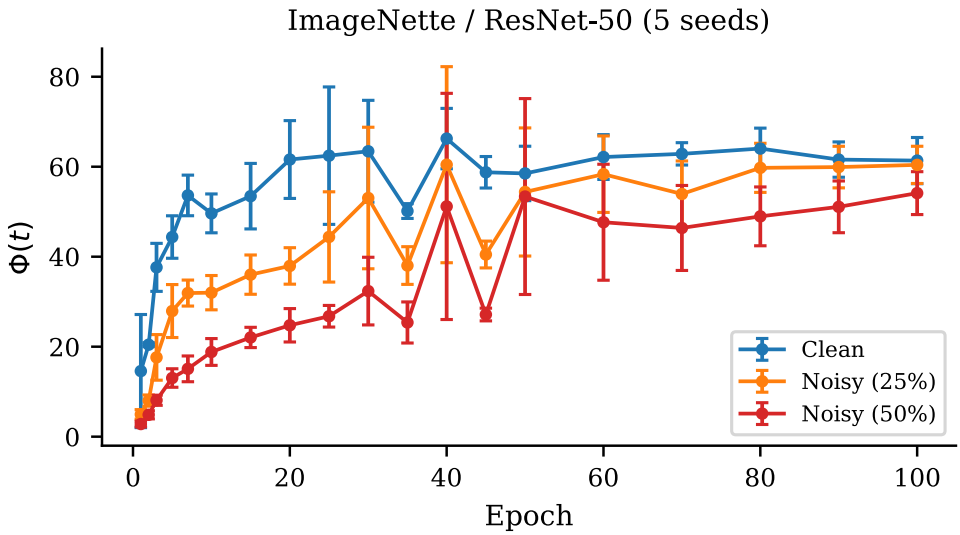

Figure 3: Raw persistence on ImageNette/ResNet-50 over 100 epochs. Curves show the mean across seeds, with error bars indicating the sample standard deviation (SD, ddof = 1).

while no seed crosses under RMS or z-score (Table 3). FashionMNIST shows no crossover under any of the five constructions. These results depend on both the pipeline and the distance construction.

Two trajectory properties help describe these cases: clean persistence declines (C1), and noisy persistence declines less rapidly (C2). For crossover to occur, the initially negative gap must close within the observed horizon. C1 and C2 describe the measured trajectories rather than provide a general crossover criterion, and C2 requires observations from noisy training. In the held-out ResNet-34/CIFAR-10 experiment, the clean trajectory supported a prospective prediction of no crossover, subsequently confirmed in six noisy runs against three clean reference runs (Appendix Q). The recomputed CIFAR-100/ResNet-18 analysis shows no crossover in any of the three seeds under any of the five constructions. Under unit normalisation, the epoch-50/epoch-10 persistence ratios are 0.843 for clean training and 0.434 for noisy training. Persistence declines in both cases, with a larger proportional decline under noisy supervision. Crossover timing refers to sampled checkpoints and does not identify when individual corrupted examples begin to be memorized.

Persistent clean trajectories. On ImageNette/ResNet-50, the mean clean trajectory rises overall over 100 epochs and remains above the noisy means at the displayed checkpoints (Figure 3). The curves fluctuate during training, showing that persistence can change without crossover. The ResNet-34, ViT-Tiny, and Tiny-ImageNet experiments provide further comparisons, although differences in datasets, sampling, and normalisation preclude a general rule based on capacity or depth.

A direct scale check. Under uniform positive rescaling of a fixed geometry, Φ_{H_1} and Φ_{H_0} scale by the same factor, giving a slope of one when $\log \Phi_{H_1}$ is plotted against $\log \Phi_{H_0}$. On CIFAR-10/ResNet-18 at epoch 50, the estimated slope is 1.577 (95% interval $[1.057, 2.052]$) using four noise levels and 2.260 ($[1.903, 2.646]$) after excluding the fully randomized level. Across four vision

Table 4: Random versus Patrini asymmetric relabelling on ResNet-18. The conditions have unequal realised corruption rates, as stated in the text. Each seed contributes the median of its three per-setting random/Patrini ratios. Entries give the number of seeds agreeing with the indicated direction out of eight, with exact two-sided Wilcoxon p-values in parentheses. Arrows indicate the random arm relative to the Patrini arm, not changes over training.

Observable / direction	Epoch 10	Epoch 25	Epoch 50
Raw Φ	$\downarrow$ 8/8 (.0078)	$\downarrow$ 8/8 (.0078)	$\uparrow$ 8/8 (.0078)
Cosine Φ ($\uparrow$)	7/8 (.0156)	8/8 (.0078)	8/8 (.0078)
Φ_{H_1}/Φ_{H_0} ($\uparrow$)	7/8 (.0156)	6/8 (.0391)	8/8 (.0078)
Mean norm ($\downarrow$)	8/8 (.0078)	8/8 (.0078)	8/8 (.0078)

configurations and 14 checkpoints excluding CIFAR-100, slopes differ from one in 52/56 cells after Benjamini–Hochberg false-discovery-rate correction using four levels, compared with 19/56 using three. The latter analysis has fewer points per slope estimate. The reported ViT-Tiny slope is much closer to one at 1.029 ($[1.008, 1.055]$). Because the bootstrap resamples individual runs, including runs that share a seed, we interpret the slope estimates descriptively. The within-model comparison in Section 3.1 tests sensitivity to uniform rescaling directly using scale-invariant ratios.

Exploratory comparison of relabelling conditions. The ResNet-18 experiment compares the Patrini asymmetric relabelling scheme with a condition originally designed to be instance-dependent. In the latter condition, entropy scores computed with a shuffled data loader were assigned to dataset indices in their original order, breaking the correspondence between scores and examples. The selected examples were then assigned new labels by uniform sampling. We refer to this implementation as the random-relabeling arm. At nominal settings of 0.1, 0.2, and 0.4, the fractions of labels that actually changed were 0.050, 0.100, and 0.200 in the Patrini arm and 0.089–0.090, 0.179–0.182, and 0.359–0.361 in the random-relabeling arm, respectively. The comparison is exploratory because the arms differ in both the fraction of labels changed and the relabelling rule.

For each observable and seed, we take the median of the three per-setting random/Patrini ratios. Table 4 reports exact two-sided Wilcoxon tests of these seed-level ratios against one, using eight seeds. Raw persistence is lower in the random arm at epochs 10 and 25 and higher at epoch 50 in all eight seeds. Cosine persistence is higher in seven, eight, and eight seeds, respectively, while mean activation norm is lower in all eight seeds at each checkpoint (Table 4). The Φ_{H_1}/Φ_{H_0} ratio is higher in seven, six, and eight seeds. At epoch 50, the ratio of pooled median cosine persistence is 2.759. Figure 1(b) shows the seed-level epoch-50 comparison. Across the 48 runs, the partial Spearman correlation between the relabelling arm and cosine persistence was 0.472 after adjustment for mean norm and diameter. This is a descriptive association across runs sharing eight seeds. The two arms differ in both how labels were reassigned and how many labels changed.

On CIFAR-10N, unit-normalised and cosine persistence correlate with corruption rate at $\rho = 0.936$ and 0.929, respectively, compared with $\rho = 0.302$ ($p = 0.15$) for raw persistence. Across these experiments, activation scale and normalised persistence describe different aspects of the response to label noise.

Alternative summaries and sampling. Loop count is invariant to uniform positive rescaling under the relative lifetime cutoff. Its reported noise correlations are $\rho = -0.799$ on CIFAR-10/ResNet-18 and -0.830 on CIFAR-10/ViT-Tiny (both $p < 0.001$), but are not significant on the reported CIFAR-100, FashionMNIST, and MNIST comparisons (Appendix U). Thus the number of retained loops captures part of the response, while their lifetimes retain additional descriptive detail. Separately, the added CIFAR-10/ResNet-18 epoch-10 sampling analysis gives $\rho = -0.850$ at 100 points and -0.945 at each tested size from 200 to 2,000 points. Equal rank correlations do not imply equal persistence values or eliminate finite-sampling effects.

6 Grokking and Decreasing Persistence

Grokking refers to delayed generalization after training accuracy has saturated [Power et al., 2022]. We examine twelve runs of two-layer Transformers trained on modular addition and multiplication

Table 5: In-sample noise-rate calibration across five pipelines at epoch 10. Within each pipeline, ordinary least squares with an intercept predicts the nominal noise setting from each statistic over individual runs. Ranges and means summarize R^2 across pipelines; ranges are not confidence intervals.

Observable	R^2 range	Mean R^2
Mean activation norm	0.890–0.990	0.936
Raw Φ	0.871–0.927	0.908
Point-cloud diameter	0.566–0.972	0.790
Unit-normalised Φ	0.382–0.821	0.606
Φ_{H_1}/Φ_{H_0}	0.115–0.757	0.500

with primes 97 and 113, using three seeds per task-prime combination. Unfiltered total persistence on raw activations decreases across the grokking transition in ten runs and increases by 3.7 and 4.5 in the other two (Figure 4; Appendix J). The mean change is -35.2 across all twelve runs and -43.1 among the ten decreasing runs. We define $\Delta\Phi_{\mathrm{grok}}$ as the difference in mean test-set persistence between checkpoints with training accuracy ≥ 0.99 and test accuracy ≥ 0.90, and those with training accuracy ≥ 0.99 and test accuracy < 0.50. Persistence is computed on 300 test points as the sum of all finite H_1 lifetimes, without a relative cutoff. The results associate delayed generalization with lower raw persistence in most runs, extending the analysis beyond label corruption. However, they do not establish a causal role for unfolding. The evidence remains limited to modular arithmetic, and the observed pattern has not been shown to hold under scale normalisation.

7 Diagnostic Use: Calibration and Representation Comparisons

An early representation statistic can be calibrated against the nominal noise setting within a fixed pipeline. Calibration requires runs at known settings, including a clean reference. A clean run alone cannot determine the slope, although applying the calibration to a new probe does not require individual corruption labels. At epoch 10, we fit a separate ordinary least-squares regression with an intercept within each of five pipelines, predicting the nominal noise setting from each representation statistic over individual runs. Table 5 reports in-sample R^2, not held-out prediction performance. Mean activation norm achieves a mean R^2 of 0.936, compared with 0.908 for raw persistence. Excluding the fully randomised condition gives 0.892 and 0.729, respectively. Appendix K specifies the pipelines and calibration protocol.

Normalised persistence describes representation differences after specified aspects of activation scale are removed. The within-model subset comparison and the normalised trajectories address this descriptive role. Population-level calibration is a separate task from detecting individual corrupted examples or improving robust classification accuracy; its regression R^2 should not be interpreted as detection or classification performance. Two exploratory experiments examine whether persistence can guide training decisions. The checkpoint-selection procedure uses validation accuracy to choose between the persistence-peak checkpoint and the final checkpoint (Appendix G). It therefore uses validation labels and does not establish an advantage over selecting the best checkpoint by validation accuracy. Adaptive label smoothing improves on fixed smoothing at one tested noise level, but its displayed mean accuracy is 0.07 percentage points lower.

Additional results show why persistence should not be treated as an objective to maximize or minimize (Appendix F). A covariance effective-rank proxy correlates with raw persistence on CIFAR-10/ResNet-18 ($\rho = 0.831$), but shows little association on FashionMNIST ($\rho = 0.002$). Label smoothing and weight decay reduce persistence, whereas topology-guided training increases it, without a corresponding monotone ordering of test accuracy. Persistence thus characterizes representation changes, while its value for choosing training interventions requires separate evaluation.

8 Scope and Limitations

The results vary across datasets and architectures. AG News/DistilBERT shows a positive early noise association, whereas the evaluated vision settings show negative associations. Clean persistence remains elevated on ImageNette and Tiny-ImageNet, while the recomputed CIFAR-100/ResNet-

18 trajectories show no crossover. These differences make calibration specific to each pipeline. Understanding the role of depth requires comparisons with matched data and measurement protocols. Sampling also matters. From epochs 10 to 50, clean global persistence on ViT-Tiny/CIFAR-10 rises from 32.1 to 66.1 at $n = 300$, and the classwise sum rises from 260.3 to 403.4. Clean global persistence on Tiny-ImageNet/ResNet-18 rises from 84.4 to 165.6 (Appendices O and P). Classwise PH measures structure within classes, while global PH also includes relations between classes. On Tiny-ImageNet, these analyses use 300 points per class across 50 classes and 300 points overall, respectively. Their different budgets and aggregation rules must be accounted for when comparing discrimination.

The late training-probe contrast distinguishes clean and corrupted groups at a checkpoint where most corrupted examples are predicted as their original labels. The class-conditional Gaussian control detects no group difference in excess persistence, leaving the interpretation of folding to be assessed alongside class geometry, anisotropy, and sampling. In the exploratory ResNet-18 relabelling comparison, noise structure and realised corruption rate vary between arms, leaving their individual contributions to the persistence contrast unresolved. Several questions remain open. Sensitivity to the fixed cutoff of $0.01L_{\max}$ requires a threshold sweep, and finer checkpoint sampling would improve the timing resolution of the 14-checkpoint study. The current summaries describe groups of examples, with classwise analysis additionally requiring class labels. Extending the piecewise-affine account to LayerNorm and attention, and testing normalised persistence during grokking beyond modular arithmetic, remain directions for further work.

9 Related Work

Persistent homology has been used to study neural representations and network structure [Naitzat et al., 2020, Rieck et al., 2018, Hofer et al., 2017], filter noisy labels and regularize classifiers [Wu et al., 2020, Chen et al., 2019], and investigate intrinsic dimension and generalization [Birdal et al., 2021]. This work examines how persistence evolves during training with noisy labels, using comparisons between clean and corrupted training subsets and controls for activation scale. The ability of neural networks to fit random labels, and their tendency to learn patterns before memorizing noise, are well established [Zhang et al., 2016, Arpit et al., 2017]. Risk correction and sample filtering seek to improve learning under label noise [Natarajan et al., 2013, Li et al., 2020]. Our study uses persistence to characterize the accompanying representation changes at the population level.

Other accounts of training dynamics include the information bottleneck, double descent, neural collapse, and grokking [Shwartz-Ziv and Tishby, 2017, Nakkiran et al., 2021, Papyan et al., 2020, Power et al., 2022]. Related approaches to characterizing representations include intrinsic dimension [Ansuini et al., 2019], representation topology divergence [Barannikov et al., 2022], and CKA [Kornblith et al., 2019], alongside further topological analysis [Ballester et al., 2026] and ricci curvature analysis [Sun and Harit, 2025]. Our experiments examine persistence alongside activation scale and local geometry to understand which aspects of the response to noisy supervision each measure captures.

10 Discussion and Conclusion

Noisy supervision can leave a measurable signature in learned representations even when most predictions for relabelled examples still match their original classes. In our fixed-probe comparison, subsets with corrupted labels showed higher loop persistence late in training, and the difference persisted after controlling for uniform rescaling. The result shows why predictions and representation geometry are worth examining together when studying how models learn from conflicting labels. The choice of measurement depends on the question. Across the five pipelines we evaluated, mean activation norm achieved a higher mean in-sample R^2 than raw persistence for predicting the nominal noise setting. To study how representations respond during training, our results motivate following a fixed probe across checkpoints and reading scale-controlled persistence alongside agreement with the original and assigned labels, where both are known. These findings motivate a broader principle for studying learning: examine how a model organises the data alongside how well it fits the labels, particularly when supervision is imperfect.

Acknowledgements and Disclosures

Funding. This research received no specific funding. **Competing interests.** The authors declare no competing interests.

References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

Rubén Ballester, Carles Casacuberta, and Sergio Escalera. Topological data analysis for neural networks. In *SpringerBriefs in Computer Science*, pages 1–103. Springer, 2026.

Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1607–1626. PMLR, 2022.

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

Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in neural information processing systems*, 34:6776–6789, 2021.

Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2573–2582. PMLR, 2019.

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*, pages 248–255. IEEE, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. *Advances in neural information processing systems*, 30, 2017.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019. URL `https://github.com/fastai/imagenette`.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbsch, and Karsten Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology. *arXiv preprint arXiv:1812.09764*, 2018.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT: a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Zhongtian Sun and Anoushka Harit. Ricciflowrec: A geometric root cause recommender using ricci curvature on financial graphs. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 1284–1289, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise. *Advances in neural information processing systems*, 33: 21382–21393, 2020.

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

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

A Impact Statement

Understanding how noisy supervision changes representations may help researchers audit datasets and compare training procedures. Persistence describes groups of examples, so interpreting it as evidence of individual annotation errors risks confusing effects of the training pipeline with label quality. Using it alone to filter data could remove valid examples, including those from underrepresented groups. Any future use of persistence in data selection should be evaluated alongside annotation review, with particular attention to whether it disproportionately removes examples from underrepresented groups.

B Training Hyperparameters

Symmetric uniform relabelling. A fraction sf of training examples is selected for uniform relabelling over all classes, including the original class. Thus sf denotes the selection fraction rather than the realised fraction of changed labels. In the fixed training probe, corruption membership is determined by whether the assigned label differs from the original label.

Table 6: Training hyperparameters for the architectures used in the main experiments.

Architecture	Optimizer	LR	Schedule	Weight Decay	Batch	Epochs
MLP	Adam	0.001	Cosine	0	128	50
ResNet-18/50/101	SGD+mom.	0.1	Cosine	5×10^{-4}	128	50/100
ViT-Tiny/Large	AdamW	0.001	Cosine	0.01	128	50/100
DistilBERT	AdamW	2×10^{-5}	Lin. warmup	0.01	32	50

Synthetic 2D setup. We train a 2-hidden-layer MLP with 64 ReLU units per layer (input dimension 2, output dimension 2) for 100 epochs using Adam with learning rate 0.01. Boundary points are defined as points satisfying $|x_0| < 1.0$, that is, within distance 1.0 of the optimal vertical separator between the two clusters. The two 2D Gaussian clusters are centred at $(-2, 0)$ and $(+2, 0)$ with unit variance, and we use 20 seeds per flip rate. The flip fraction is defined over the near-boundary points, not over the entire training set.

Persistence conventions. The relative lifetime filter in Eq. 1 is used for the normalisation, slope, crossover, training-probe, NEAR/FAR, revised classwise, pair-distance, synthetic, ViT noise-type, exploratory class-pair, ImageNette, Tiny-ImageNet, and CIFAR-10N analyses. Grokking, TGCS, and the original large-scale corruption and pretrained analyses sum all finite lifetimes. The H_0 summary excludes the infinite bar. Ripser uses coefficients in $\mathbb{Z}/2\mathbb{Z}$. Ratios are set to zero when their denominators are zero.

Normalisation. In the normalisation grid, each transformation is applied to the selected 300-point subsample. Unit normalisation divides each activation vector by its Euclidean norm. Z-score standardisation subtracts each feature's mean and divides by its standard deviation, computed over the subsample with divisor n. RMS rescaling centres the cloud and divides by the root mean squared norm of the centred vectors. All three denominators are bounded below by 10^{-12}. Cosine dissimilarities are computed directly from the selected activations.

Classwise protocols. The revised protocol groups saved test activations by true labels and sums persistence across eligible classes. It uses a budget of 50 points per class, includes classes with at least 10 examples, and skips classwise analysis for datasets with more than 20 classes. This protocol was introduced after the original results were available.

The original ViT protocol groups activations by noisy labels from an augmented training loader. It includes classes with at least 30 examples and samples up to 300 points per class without a fixed subsampling seed. The two protocols therefore differ in data source, class labels, eligibility, and sampling budget.

Statistical summaries. Table 2 reports ratios of seed medians, with tests applied to paired values within each run. For Table 4, the three per-setting ratios are reduced to a median within each seed. These seed-level medians are tested against one using an exact two-sided Wilcoxon signed-rank

test. Pooled analyses across seeds and noise settings are descriptive, as observations sharing a seed are dependent. Slope intervals use 2,000 bootstrap resamples of runs without clustering by seed. Benjamini–Hochberg correction is applied across the cells within each reported slope analysis.

Gaussian controls. Gaussian controls are fitted using the class-conditional means and covariances of the observed subset. Grid controls use the 300-point subsample and five Gaussian draws. Training-probe controls use the matched subset grouped by true labels and ten draws, with persistence evaluated on 300 points. The comparison between subsets tests the difference in excess persistence relative to their fitted controls.

Data corrections. The CIFAR-100/ResNet-18 results reported here use reruns that corrected a dataset-loading error in the original experiments. We excluded the affected original outputs and the old Tiny-ImageNet/ResNet-50 and ResNet-101 noise-condition outputs. The Tiny-ImageNet/ResNet-18 analysis in Appendix P and the CIFAR-100 checkpoint-selection experiment were run separately.

C Subsampling Sensitivity

Table 7 compares persistence rankings from subsamples with those from a 2,000-point reference across 433 configurations. At $n = 300$, the Spearman correlation is 0.973, with a runtime of 0.058 s per call compared with 2.944 s for the reference. The mean subsampling coefficient of variation is 0.022, with a maximum of 0.086.

Table 7: Subsampling sensitivity across 433 configurations. ρ measures rank agreement with the 2,000-point reference. Dashes indicate unreported runtimes.

n	50	100	200	300	500	750	1000	1500
ρ	0.908	0.922	0.960	0.973	0.979	0.982	0.986	0.998
Time (s)	—	—	—	0.058	—	—	0.612	—

A separate CIFAR-10/ResNet-18 experiment at epoch 10 varies the global point count over $n \in \{100, 200, 300, 500, 1000, 2000\}$. Across three noise levels and five seeds, the Spearman correlation with noise level is -0.850 at $n = 100$ and -0.945 at every tested size from $n = 200$ onward. The repeated correlation describes the strength of the rank association, while persistence magnitudes and sampling variability may still differ. This experiment measures association with noise level, whereas Table 7 measures agreement with a larger reference cloud. Both analyses vary point count while leaving sensitivity to the lifetime cutoff for separate evaluation.

D Raw Persistence Associations over Training

Table 8 reports Spearman correlations between raw persistence and noise level at epochs 10 and 50. These correlations describe how noise conditions are ordered at each checkpoint. Section 5 examines when noisy persistence overtakes its clean reference under different distance constructions. The original CIFAR-100/ResNet-18 row is omitted following recomputation, with the updated crossover analysis reported in that section.

Table 8: Spearman correlations between noise level and raw persistence at epochs 10 and 50. Changes in correlation describe changes in noise ordering, whereas crossover concerns noisy persistence relative to a clean reference. Values use the raw, 300-point reanalysis consistently across pipelines.

Dataset	Architecture	ρ_{10}	ρ_{50}
CIFAR-10	ResNet-18	-0.969	-0.225
MNIST	MLP	-0.931	-0.318
CIFAR-10	ViT-Tiny	-0.962	-0.962
FashionMNIST	MLP	-0.954	-0.876
AG News	DistilBERT	$+0.870$	$+0.870$

E Large-Scale Results

The original extended CIFAR-100/ResNet-50 experiment reports a crossover at epochs 60–70. The corresponding original ResNet-18 crossover has been superseded by the no-crossover recomputation, so these observations do not establish a monotone delay of crossover with depth. A matched analysis using the same extraction, normalisation, and checkpoint protocol is needed for that comparison.

F Negative Results

ACRP as a PH proxy. Activation Covariance Rank Proxy (ACRP) is the effective rank of the activation sample covariance matrix, defined as $(\mathrm{tr}(\Sigma))^2/\|\Sigma\|_F^2$, where Σ is the sample covariance of the penultimate-layer activations. This quantity, also known as the participation ratio, captures how activation variance is distributed across feature dimensions, with higher values indicating more uniform spread. We tested ACRP as a Ripser-free proxy for monitoring topological complexity in real time. It correlates with Φ on some datasets, including CIFAR-10/ResNet-18 ($\rho = 0.831$) and CIFAR-100 runs ($\rho = 0.504$, $n = 168$), but not on others, including FashionMNIST ($\rho = 0.002$) and MNIST ($\rho = -0.321$), so it is not a universal proxy for persistent homology.

G TGCS Algorithm

Pseudocode for the topology-guided checkpoint selection (TGCS) procedure used in the appendix analysis.

Topology-Guided Checkpoint Selection (TGCS)

Inputs: training procedure, validation set $\mathcal{V}$, subsample size $n = 300$, and final epoch T.

1. Train for T epochs, computing $\Phi(t)$ at each checkpoint.
2. Let $t_{\mathrm{peak}} = \arg\max_t \Phi(t)$.
3. Retain the checkpoints at t_{peak} and T.
4. Evaluate both on $\mathcal{V}$ and return the one with higher validation accuracy.

H Compute Budget

Table 9: Training used the GPU partition, including A100 and older GPU models; persistent-homology computations and response-period analyses used CPUs. GPU-hours and CPU-hours are reported separately.

Component	GPU-hours	CPU-hours
Original training (900+ runs)	∼2,400	—
Grokking experiments (12 jobs)	∼48	—
Realistic noise (39 jobs)	∼195	—
TGCS evaluation (40 jobs)	∼120	—
PH computation	—	∼50

I Regularisation Methods

We compare six training methods on CIFAR-10/ResNet-18: the baseline, dropout, label smoothing, high weight decay, topology-guided training (TGT), and validation-based early stopping. TGT uses Φ to guide learning-rate and weight-decay adjustments. Results are averaged over noise levels. At epoch 50, label smoothing and high weight decay reduce Φ by approximately 50% relative to the baseline, whereas TGT increases it. Test accuracy does not follow a monotone ordering with persistence across these methods (Table 10).

Table 10: Regularisation methods on CIFAR-10/ResNet-18 (averaged over noise levels).

Method	Φ (ep50)	Test accuracy
Baseline	357.0	52.5%
Dropout	351.9	52.3%
Label smoothing	174.0	52.8%
Weight decay (high)	163.6	46.8%
TGT	618.6	50.1%
Val. early stop	—	51.4%

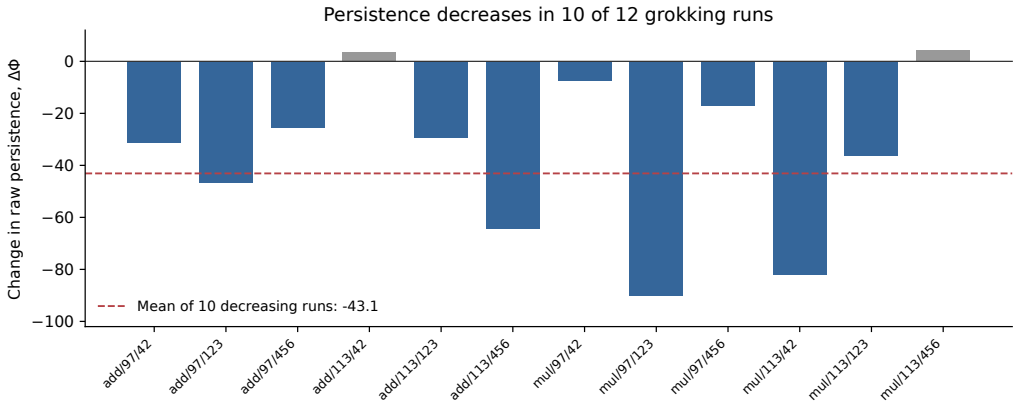

Figure 4: Changes in raw persistence across the grokking transition for all twelve runs. Persistence decreases in ten runs and increases by 3.7 and 4.5 in the remaining two. The mean change is -35.2 across all runs and -43.1 among the ten decreasing runs. The dashed red line marks the latter mean.

J Grokking Results

The experiment evaluates two-layer Transformers on modular addition and multiplication with primes 97 and 113, using three seeds per task and prime. For each run, $\Delta\Phi_{\text{grok}}$ is the mean test-set persistence over checkpoints with training accuracy ≥ 0.99 and test accuracy ≥ 0.90, minus the mean over checkpoints with training accuracy ≥ 0.99 and test accuracy < 0.50 (Table 11). Persistence is computed on 300 test points by summing all finite H_1 lifetimes without the relative filter. Ten runs show a decrease, and two show increases of 3.7 and 4.5.

Table 11: Changes in raw persistence across twelve grokking runs. The mean change is -35.2 across all runs and -43.1 among the ten decreasing runs.

Task	Prime	Seed	$\Delta\Phi_{\text{grok}}$	Decreases?
Addition	97	42	-31.2	Yes
Addition	97	123	-46.9	Yes
Addition	97	456	-25.7	Yes
Addition	113	42	$+3.7$	No
Addition	113	123	-29.4	Yes
Addition	113	456	-64.3	Yes
Multiplication	97	42	-7.4	Yes
Multiplication	97	123	-90.1	Yes
Multiplication	97	456	-17.3	Yes
Multiplication	113	42	-82.1	Yes
Multiplication	113	123	-36.5	Yes
Multiplication	113	456	$+4.5$	No

K Calibration Targets

Table 5 uses five pipelines: CIFAR-10/ResNet-18, CIFAR-10/ViT-Tiny, MNIST/MLP, FashionMNIST/MLP, and CIFAR-100/ResNet-18. Within each pipeline, ordinary least squares with an intercept predicts the nominal noise setting from one epoch-10 representation statistic, using individual runs as observations. Fitting and evaluation use the same observations; the reported R^2 values are in-sample. The statistics are mean activation norm, raw persistence, point-cloud diameter, unit-normalised persistence, and Φ_{H_1}/Φ_{H_0}. Ranges and means in Table 5 summarize the five pipeline-specific R^2 values. Mean activation norm and raw persistence achieve mean R^2 values of 0.936 and 0.908. After excluding the fully randomised condition, the corresponding means are 0.892 and 0.729. These fits require runs at known noise settings and do not establish held-out calibration accuracy.

L Φ-Guided Adaptive Label Smoothing

We test whether the early noise estimate can be used to drive adaptive regularisation. At epoch 10, we compute $\hat{s}$ from $\Phi(10)$ and then set the label-smoothing strength to $\hat{s} \times 0.8$ for the remaining epochs. Results on CIFAR-10/ResNet-18 (4 noise levels, 3 seeds each) are shown below.

Table 12: Φ-guided adaptive vs. fixed label smoothing (0.1). Bold = best per row.

sf	Baseline	Fixed (0.1)	Adaptive	Winner
0.00	**86.92**	86.74	86.77	Baseline
0.10	84.96	**85.62**	85.25	Fixed
0.25	82.64	83.16	**83.47**	Adaptive
0.50	77.00	**77.50**	77.28	Fixed
Mean	82.88	**83.26**	83.19	Fixed

Adaptive smoothing wins at sf=0.25 ($+0.31$ percentage points over fixed). However, it does not outperform fixed smoothing on average (-0.07 percentage points using the displayed means). The reported accuracy differences do not isolate whether calibration error or the smoothing rule limits performance. We include this as a proof-of-concept: the topological signal can be used in a regularisation rule, but this rule does not beat the fixed baseline on average.

M Persistence-Peak and Final-Checkpoint Comparisons

Table 13 shows selected results from the 40-run checkpoint comparison. It reports final test accuracy and the positive part of the difference between test accuracy at the persistence peak and at the final checkpoint. Negative differences are recorded as zero. Evaluating the validation-based selection procedure in Appendix G requires a separate comparison with a validation-only baseline using the same checkpoint budget.

Table 13: Selected comparisons between persistence-peak and final checkpoints. $\delta = 100 \max(0, a_{\text{peak}} - a_{\text{final}})$, where a_{peak} and a_{final} are test accuracies on a 0–1 scale at the Φ-peak and final checkpoints, respectively. Reported clean-data and ViT-collapsed entries have $\delta = 0$.

Dataset	Arch	sf	Seed	Final acc	δ (percentage points)
CIFAR-100	ResNet	0.5	42	0.3842	$+10.0$
CIFAR-100	ResNet	0.5	123	0.3904	$+11.0$
FashionMNIST	MLP	0.5	42	0.7265	$+5.0$
FashionMNIST	MLP	0.5	123	0.7263	$+3.6$
FashionMNIST	MLP	0.5	456	0.7285	$+2.5$
MNIST	MLP	0.5	42	0.6804	$+3.3$
CIFAR-10	ResNet	0.5	42	0.7407	$+1.3$
All clean (sf=0.0)				—	0.0
ViT-Tiny (collapsed)				—	0.0

N Depth Ablation

The original depth ablation used activations extracted during training, whereas the canonical tables used saved activations and a different computation pipeline. Its absolute values cannot be interchanged with the canonical reanalysis. Because the CIFAR-100/ResNet-18 baseline was subsequently recomputed and no longer crosses, the original depth figure and the $1.5\times$–$6.4\times$ ratio comparison are withheld pending reconciliation; they do not currently support a general claim that depth delays folding.

O ViT Per-Class PH Investigation

We reanalyse all 280 saved ViT-Tiny activation files using per-class persistent homology. This is a secondary classwise analysis; its aggregate magnitude differs from a single global computation because its grouping and point budget differ.

Table 14: ViT-Tiny/CIFAR-10: per-class Φ by noise level (mean over 5 seeds). The classwise sum is larger than the global sum and has noise correlation $\rho \leq -0.93$; unequal point budgets preclude interpreting the magnitude ratio as diagnostic improvement.

sf	Per-class Φ	Global (n=300)	Global (n=1000)	Epoch
0.00	260.3	32.1	118.8	10
0.25	143.0	18.5	60.3	10
0.50	118.5	13.5	48.9	10
1.00	0.0	0.0	0.0	10
0.00	403.4	66.1	229.0	50
0.25	160.5	25.8	82.3	50
0.50	168.6	22.2	74.5	50
1.00	0.0	0.0	0.0	50

Clean classwise persistence rises from about 260 to 403. Both global and classwise values retain a noise association, and the original classwise analysis reports $\rho \leq -0.93$, $p < 10^{-9}$. A larger aggregate total cannot on its own distinguish inter-class interference from the effect of computing and summing PH over more points. These results therefore establish sensitivity to grouping and sample budget rather than a unique explanation of the original weak signal.

P Tiny-ImageNet Scaling

We train ResNet-18 on Tiny-ImageNet (200 classes, 64×64, 100K train images) with 4 noise levels and 5 seeds each (20 runs total, 50 epochs). For this experiment, we save a larger activation set per checkpoint and use 50 classes $\times$ 300 points for per-class PH analysis.

Table 15: Tiny-ImageNet/ResNet-18: global vs. per-class Φ by noise level (mean over 5 seeds). The larger classwise aggregate discriminates noise with $\rho = -0.969$ ($p < 10^{-12}$) at epochs 10 and 25.

sf	Global (n=300)	Per-class	Val Acc	Epoch
0.00	84.4	4,035.2	31.0%	10
0.25	82.3	3,311.4	25.4%	10
0.50	54.3	2,554.4	19.2%	10
1.00	0.6	26.5	0.5%	10
0.00	165.6	5,332.6	50.3%	50
0.25	94.5	4,638.8	35.8%	50
0.50	75.7	4,934.8	21.3%	50
1.00	4.1	222.7	0.5%	50

Interpretation. Global PH at $n = 300$ separates the reported clean and noisy conditions. Classwise PH uses 50 classes with 300 points each (15,000 points in total), so its larger aggregate is not a

matched-budget comparison. Its noise correlation is $\rho = -0.969$ ($p < 10^{-12}$, $n = 20$ runs) at epochs 10 and 25 and -0.752 at epoch 50. Global clean persistence rises from 84.4 to 165.6. The table supports a persistent clean trajectory, without establishing a universal rule that increasing clean persistence precludes crossover.

Q Prospective Crossover Test on ResNet-34

CIFAR-10/ResNet-34 provides a prospective test of a crossover prediction based on clean training. This configuration was not used to derive C1/C2 and follows the same training recipe as ResNet-18.

Across three initial clean runs, mean raw persistence rose from 7.4 at epoch 1 to 28.2 at epoch 5, peaked at 32.9 at epoch 35, and ended at 31.7 at epoch 50 (Table 16). Clean persistence thus remained elevated throughout the later checkpoints. This trajectory motivated a no-crossover prediction, recorded before noisy training began.

Subsequent training used three seeds at each of $\mathrm{sf} = 0.25$ and $\mathrm{sf} = 0.5$. All six noisy runs remained below the clean reference at every sampled checkpoint, consistent with the prediction. The contrast with ResNet-18 on CIFAR-10 illustrates variation in crossover behavior across architectures under the reported training recipe.

Table 16: Raw persistence on CIFAR-10/ResNet-34, averaged over three seeds per noise setting. The clean trajectory motivated a no-crossover prediction before noisy training began. All six noisy runs remained below the clean reference at every sampled checkpoint.

Epoch	$\mathrm{sf} = 0$ (clean)	$\mathrm{sf} = 0.25$	$\mathrm{sf} = 0.5$
1	7.4	2.6	0.3
5	28.2	16.7	6.0
10	28.2	16.8	10.7
25	29.5	16.5	12.8
35	32.9	17.1	12.5
50	31.7	18.6	16.4

R ViT Noise-Type Test (Asymmetric vs Instance-Dependent)

ViT-Tiny is trained on CIFAR-10 with cyclic asymmetric noise and pixel-distance-based instance-dependent noise at nominal rates of 0.2 and 0.4, using five seeds per condition. Cyclic asymmetric noise maps class c to $(c + 1) \bmod 10$, a different mapping from the Patrini construction. Instance-dependent noise uses pixel-space distances from the true-class centroid and assigns labels uniformly, allowing the original class to be redrawn. The resulting fractions of changed labels are approximately 0.134 and 0.267, respectively. Persistence is computed within each class using a budget of 300 points per class. These results come from a separate protocol from the ResNet-18 relabelling experiment.

Table 17: Per-class Φ on ViT-Tiny under asymmetric vs. instance-dependent noise (mean $\pm$ std over 5 seeds). The two noise types produce no statistically detectable difference in raw persistence at epoch 10 ($p > 0.3$) but different raw persistence at epoch 50 ($p \leq 0.002$).

Epoch	Rate	Asymmetric Φ	Instance Φ	Ratio	t-test p
10	0.2	821.1 ± 42.3	796.7 ± 43.7	0.97	0.395
10	0.4	777.9 ± 32.9	774.8 ± 31.1	1.00	0.885
25	0.2	660.5 ± 58.3	610.9 ± 24.5	0.92	0.117
25	0.4	662.0 ± 28.9	711.1 ± 37.1	1.07	0.048
50	0.2	426.6 ± 29.5	505.4 ± 27.1	**1.18**	**0.002**
50	0.4	485.8 ± 29.5	717.3 ± 49.0	**1.48**	$\mathbf{< 0.001}$

Interpretation. For ViT-Tiny, no difference in raw persistence between the two noise types is detected at epoch 10. At epoch 50, instance-dependent noise yields higher raw persistence at both rates ($p = 0.002$ and $p < 0.001$, respectively, from independent two-sample t-tests with five seeds per arm). The ResNet-18 comparison in Section 5 uses different corruption generators and seed-level paired tests. Its random-relabeling and Patrini arms differ in realised corruption rate.

S Global and Classwise Persistence on AG News

Saved DistilBERT/AG News activations are analysed using global PH with 300 and 1,000 points and classwise PH with 300 points from each of four classes. Table 18 reports raw persistence under clean training and $\mathrm{sf} = 0.5$, averaged over five seeds.

All three analyses show a positive Spearman correlation with noise level ($\rho = 0.870$, $p < 0.002$). At epoch 10, the classwise sum increases from 103.7 under clean training to 371.4 at $\mathrm{sf} = 0.5$, a noisy/clean ratio of approximately 3.6. This sum combines four separate 300-point computations, with a different point budget and aggregation from global PH.

Table 18: DistilBERT/AG News: Φ at clean vs. sf=0.5 (mean over 5 seeds). All three Φ variants show $\rho = +0.870$ ($p < 0.002$) with a consistent *sign reversal* compared to vision: noise raises raw total persistence; scale and loop count are not separated by this table.

Epoch	sf	Global (n=300)	Per-class	Global (n=1000)	Noisy/Clean Ratio
10	0.00	18.1	103.7	56.1	—
10	0.50	53.1	371.4	224.5	2.9× / 3.6× / 4.0×
25	0.00	15.7	83.0	43.9	—
25	0.50	56.7	354.2	178.3	3.6× / 4.3× / 4.1×
50	0.00	12.2	77.6	40.2	—
50	0.50	45.3	285.4	143.8	3.7× / 3.7× / 3.6×

Clean mean persistence declines across the reported checkpoints, while noisy means remain above their clean references in all three analyses. Noisy persistence also declines in the classwise and global $n = 1000$ analyses. For global $n = 300$, it rises slightly at epoch 25 before declining at epoch 50. The positive noise association contrasts with the early negative associations in the evaluated vision settings, motivating calibration within each pipeline.

T Sign-Reversal Diagnosis: Transformer on Vision

To help distinguish modality from architecture as a possible driver of the DistilBERT sign reversal, we train a small Transformer (4 layers, embed dim 128, 4 heads, 8×8 patches) on CIFAR-10 with 3 noise levels ($\mathrm{sf} \in \{0.0, 0.25, 0.5\}$) and 5 seeds each (15 runs total).

Table 19: Small Transformer on CIFAR-10: Φ by noise level and epoch (mean over 5 seeds, $n = 15$ total). Noise *decreases* Φ at all epochs tested, matching the vision pattern on ResNet/ViT, opposite to DistilBERT/text.

Epoch	sf	Global Φ	Per-class Φ	Val Acc	n
10	0.00	130.5	982.3	59.4%	5
10	0.25	79.6	731.2	52.6%	5
10	0.50	62.5	606.7	44.8%	5
25	0.00	129.4	911.2	68.3%	5
25	0.25	73.6	599.9	63.0%	5
25	0.50	71.8	663.4	55.5%	5
50	0.00	122.9	815.2	72.7%	5
50	0.25	61.8	524.6	67.7%	5
50	0.50	68.1	629.4	60.4%	5

Result. The Spearman correlation between Φ and noise level is negative throughout training. At epoch 10, global and per-class PH both give $\rho = -0.945$ ($p < 0.001$). At epoch 25, the correlations are $\rho = -0.775$ ($p = 0.001$) for global PH and $\rho = -0.510$ ($p = 0.05$) for per-class PH. At epoch 50, they are $\rho = -0.548$ ($p = 0.034$) and $\rho = -0.491$ ($p = 0.06$), respectively.

Interpretation. The sign is clearly negative for a vision Transformer at the diagnostic epoch ($p < 0.001$, $n = 15$), in contrast to the positive sign observed on DistilBERT/AG News. The association weakens over training but remains negative, with global correlations significant at all

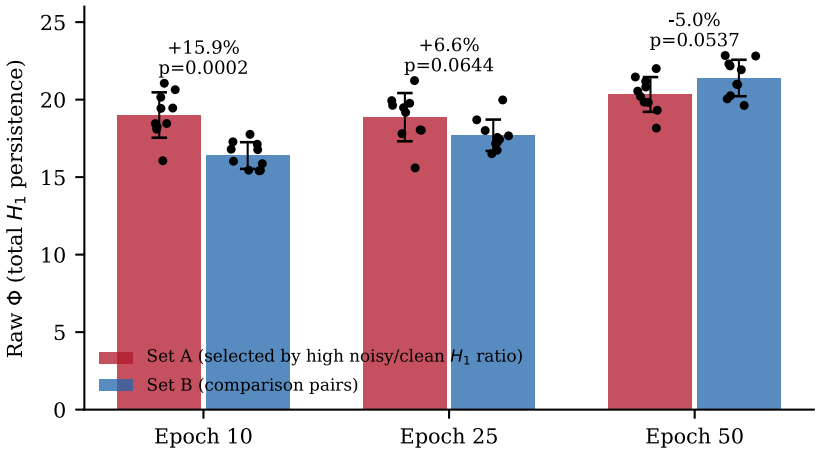

Figure 5: Exploratory class-pair comparison on CIFAR-10/ResNet-18. Set A was selected using a high noisy/clean persistence ratio; Set B contains comparison pairs. Bars show means, error bars show sample standard deviations, and dots show individual runs across ten seeds per condition. Reported p-values use Welch's two-sample t-test. The nominal pairwise flip parameter is 0.5; sequential flips over overlapping pairs change approximately 42.5% of labels, unevenly across classes. These conditions are distinct from the disjoint, centroid-selected NEAR and FAR comparison.

three reported checkpoints. This comparison shows that a Transformer on vision need not share the positive association observed for DistilBERT on text. It does not isolate modality from the many differences between the two training pipelines.

U Additional normalisation and Pair Controls

Original exploratory class-pair experiment. The exploratory CIFAR-10/ResNet-18 experiment applies a nominal pairwise flip parameter of 0.5, with ten seeds per condition. Pairs are processed sequentially on already modified labels and reuse classes. The realised changed-label fraction is approximately 42.5%, with unequal corruption across classes. The first set, selected for a high noisy/clean persistence ratio, contains {horse↔truck, automobile↔truck, automobile↔dog, frog↔ship, cat↔dog}. The comparison set contains {airplane↔automobile, airplane↔ship, bird↔ship, deer↔frog, bird↔truck}.

At epoch 10, raw persistence is 19.0 ± 1.5 in the first condition and 16.4 ± 0.9 in the comparison condition, a relative difference of 15.9% ($t = 4.87$, $p = 0.0002$, Welch's two-sample t-test). The reported dispersions are sample standard deviations across seeds. The difference decreases to 6.6% at epoch 25 ($p = 0.0644$) and -5.0% at epoch 50 ($p = 0.054$), as shown in Figure 5. Because persistence informed selection of the first set, this experiment is exploratory rather than an independent test of the selection criterion. Unlike the centroid-based NEAR and FAR conditions, these sets contain overlapping class pairs.

Pair-distance analysis. The pair-distance analysis uses $sf = 0.5$. For each class pair, clean centroid distance and the noisy/clean persistence ratio are first averaged across seeds; Spearman correlation is then computed across pairs. Both CIFAR-10 analyses use all 45 unordered pairs and five seeds. The CIFAR-100 analysis uses 60 pairs selected with RandomState(0).

Alternative summaries. Scale-invariant bar count correlates with noise at $\rho = -0.799$ on CIFAR-10/ResNet-18 and -0.830 on CIFAR-10/ViT-Tiny (both $p < 0.001$). The corresponding values on CIFAR-100, FashionMNIST, and MNIST are -0.411, 0.000, and 0.361, respectively (none significant). Thus loop count carries some of the signal, but does not replace lifetime-based summaries across these settings. The relative threshold is fixed throughout. A full threshold sensitivity experiment is not reported.

Table 20: Pair-distance correlations reported in the response-period analysis. Both CIFAR-10 configurations have $p < 0.0001$ at epochs 25 and 50. Correlations are computed across pairs after seed averaging at $sf = 0.5$.

Pipeline	Pairs	Epoch 10	Epoch 25	Epoch 50
CIFAR-10 / ResNet-18	45	-0.471	0.823	0.867
CIFAR-10 / ViT-Tiny	45	0.104	0.664	0.614
CIFAR-100 / ResNet-18	60	-0.172	-0.119	0.413

Classwise normalised analysis. For CIFAR-10/ResNet-18, the response-period classwise scale-invariant ratio has noise correlation $\rho = -0.737$ at epoch 10 ($p = 0.002$) and 0.794 at epoch 50 ($p < 0.001$). This documents a change in noise ordering. The first clean–noisy crossing time is determined from the trajectories themselves.

Seed-level relabelling comparisons. Each seed-level ratio is the median of the three per-setting ratios. At epochs 10, 25, and 50, respectively, the random-relabelling arm has higher Φ_{H_1}/Φ_{H_0} than the Patrini arm in 7/8, 6/8, and 8/8 seeds, with paired Wilcoxon p-values of 0.0156, 0.0391, and 0.0078. Mean norm is lower in 8/8 seeds at each checkpoint (median ratios 0.608, 0.580, and 0.702, all $p = 0.0078$). Diameter is also lower in 8/8 seeds, with median ratios of 0.517, 0.448, and 0.790. The two conditions differ in both realised corruption rate and relabelling structure.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction describe the persistence contrasts, their dependence on training stage and normalization, and the noise-rate calibration comparison. The training-subset result concerns groups defined by corruption membership. Folding is presented as a geometric interpretation.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 8 discusses variation across pipelines and sampling protocols, the need for pipeline-specific calibration, checkpoint resolution, the fixed lifetime cutoff and the absence of a systematic threshold sweep. It also describes the scope of the piecewise-affine interpretation and the restriction of the grokking experiments to modular arithmetic.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper is empirical. Section 4 provides a motivating geometric interpretation rather than a formal theoretical result.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper provides the experimental protocol, datasets, architectures, noise models, subsampling procedure, hyperparameters and appendix details needed to reproduce the main results to a reasonable extent.

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

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

   Answer: [Yes]

   Justification: Analysis scripts and selected result files are available at `https://github.com/alvinsun724/Memory-is-folding`. The README explains which analyses these materials support.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Training details, datasets, architectures, noise models, checkpoints, and hyperparameters are specified in the main paper and appendices.

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: The paper reports paired seed-level tests for the main subset and relabelling comparisons, defines the displayed error bars, and identifies pooled analyses and unclustered bootstrap intervals as such.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix H reports hardware categories and approximate GPU-hours and CPU-hours by experiment group.

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work is standard empirical machine learning research on public datasets and model analysis, and it does not present any apparent deviation from the NeurIPS Code of Ethics.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The impact statement in the Appendix A discusses potential uses in dataset auditing and comparisons of training procedures, risks of removing valid or underrepresented examples, and the role of annotation review and assessment across groups.

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The repository README documents the released analysis scripts, including their dependencies, usage, experimental coverage and licence.

13. **New assets**

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

    Answer: [Yes]

    Justification: The released code and result files include documentation of dependencies, usage, experimental coverage, and licensing.

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.