Topological Predictors of Catastrophic Forgetting: A Cross-Architecture Study Using Persistent Homology of Loss Landscapes

Joshua Gutierrez*
Axion Deep Labs Inc
Colorado State University Global
https://www.axiondeep.com

February 2026

Abstract

Catastrophic forgetting remains a central obstacle in continual learning, yet no reliable method exists to assess an architecture's susceptibility before sequential training begins. We investigate whether topological features of neural network loss landscapes—computed via persistent homology on 2D cross-sections around converged minima—correlate with knowledge retention under sequential training. Across 14 architectures spanning CNNs, Vision Transformers, and MLP-based designs trained on Split-CIFAR-100, we compute H_0 (connected components) and H_1 (loops) persistence of Vietoris–Rips filtrations on 50×50 landscape grids and compare them against standard geometry baselines (Hessian trace, sharpness, Fisher information, loss barrier). We find that H_1 total persistence is significantly rank-correlated with retention at 100 steps of sequential training (Spearman $\rho = 0.61$, $p = 0.021$, $n = 14$) and with area under the retention curve ($\rho = 0.65$, $p = 0.012$); both correlations are stable under leave-one-out cross-validation (14/14 folds significant). Fisher information trace shows a significant *negative* correlation with AURC ($\rho = -0.75$, $p = 0.002$). Vision Transformers exhibit both the highest topological complexity and the strongest forgetting resistance, while MLP-Mixer presents a counterexample—high H_0 but zero retention—suggesting that H_1 structure, not H_0 alone, captures the topology–retention relationship. These results are correlational; a single 2D slice and single training seed per architecture limit causal claims. We release all code, configurations, and results to enable reproduction and extension.

*Corresponding author. joshua@axiondeep.com

1 Introduction

Catastrophic forgetting—the abrupt loss of previously learned knowledge when a neural network is trained on new data—is one of the most persistent challenges in continual learning [McCloskey and Cohen, 1989, French, 1999]. Despite decades of mitigation strategies including elastic weight consolidation [Kirkpatrick et al., 2017], progressive neural networks [Rusu et al., 2016], and replay-based methods [Rebuffi et al., 2017], a fundamental question remains open: *can we predict, before sequential training begins, how resistant a given architecture will be to catastrophic forgetting?*

Recent work has established that the geometry of loss landscapes encodes meaningful information about generalization [Li et al., 2018, Keskar et al., 2017], mode connectivity [Garipov et al., 2018], and optimization dynamics. Separately, topological data analysis (TDA)—particularly persistent homology—has emerged as a principled framework for characterizing multi-scale structure in high-dimensional data [Carlsson, 2009, Edelsbrunner and Harer, 2010]. Persistence images [Adams et al., 2017] and related vectorizations have enabled TDA features to serve as inputs to machine learning pipelines.

We bridge these two lines of work by asking: **do topological features of the loss landscape around a converged minimum correlate with that minimum's resistance to catastrophic forgetting?**

Our contributions are:

1. A four-phase experimental protocol for measuring the relationship between loss landscape topology and forgetting dynamics across architectures.

2. Evidence across 14 architectures on Split-CIFAR-100 that H_1 persistence (loop structures) significantly correlates with knowledge retention ($\rho = 0.61$, $p = 0.021$), outperforming H_0 (connected components) and standard geometry baselines, with stability confirmed by leave-one-out cross-validation (14/14 folds significant).

3. Discovery that Fisher information trace is strongly *anti-correlated* with retention ($\rho = -0.75$, $p = 0.002$), suggesting that aggregate parameter sensitivity tracks forgetting vulnerability.

4. Identification of architecture-dependent effects: attention-based models show 40–90× better retention than standard CNNs despite lower initial accuracy.

5. Comparison with standard geometry baselines, several of which exhibit numerical instability at scale, highlighting a practical advantage of topological features.

6. Open-source release of all experimental code, configurations, and raw results.[1]

2 Related Work

Continual Learning and Catastrophic Forgetting. Strategies for mitigating forgetting broadly fall into regularization-based methods [Kirkpatrick et al., 2017, Zenke et al., 2017], replay-based methods [Rebuffi et al., 2017, Shin et al., 2017], and architecture-based methods [Rusu et al., 2016, Mallya and Lazebnik, 2018]. Our work is orthogonal: rather than mitigating forgetting, we investigate whether properties of the pre-trained model *correlate* with forgetting susceptibility.

Loss Landscape Analysis. Li et al. [2018] introduced filter-normalized random directions for meaningful cross-architecture visualization. Keskar et al. [2017] linked sharpness (largest Hessian eigenvalue) to generalization. Garipov et al. [2018] demonstrated low-loss paths between independently trained minima. We adopt filter-normalized 2D cross-sections following Li et al. [2018] and compute topological invariants on the resulting surfaces.

TDA in Machine Learning. Persistent homology has been applied to neural network analysis in several contexts: characterizing decision boundaries [Ramamurthy et al., 2019], analyzing weight-space topology during training [Rieck et al., 2019], and studying activation patterns [Naitzat et al., 2020]. Adams et al. [2017] introduced persistence images for stable vectorization. Our work differs in applying persistent homology directly to loss landscape *surfaces* rather than to weight trajectories or activation manifolds.

[1] `https://github.com/axiondeep/axiondeep-research`

3 Methodology

Our experimental protocol consists of four phases, each producing data consumed by subsequent phases.

3.1 Phase 1: Task A Training

Each architecture is trained to convergence on Task A (first 50 classes of CIFAR-100) using SGD with cosine annealing (100 epochs, 5-epoch linear warmup). All hyperparameters are held constant across architectures: learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, batch size 128, seed 42. The converged checkpoint (best validation accuracy) defines the minimum whose landscape we analyze. Using a single seed means each architecture produces one converged minimum; variance across seeds is not quantified in this study.

3.2 Phase 2: Loss Landscape Topology

We compute persistent homology on a 2D cross-section of the loss landscape around the converged minimum $\boldsymbol{\theta}^*$, proceeding in three stages.

Stage 1: Landscape sampling. Generate two random directions $\mathbf{d}_1, \mathbf{d}_2$ in parameter space with *filter normalization* [Li et al., 2018]: for each convolutional filter (or weight matrix) i, rescale $\mathbf{d}^{(i)}$ so that $\|\mathbf{d}^{(i)}\| = \|\boldsymbol{\theta}^{(i)}\|$. Evaluate the loss on a uniform 50×50 grid over $[-1, 1]^2$:

$$\mathcal{L}(\alpha, \beta) = \mathcal{L}(\boldsymbol{\theta}^* + \alpha\, \mathbf{d}_1 + \beta\, \mathbf{d}_2)\,. \tag{1}$$

Stage 2: Weighted graph construction. Let $V = \{v_{ij}\}$ denote the $50 \times 50 = 2{,}500$ grid vertices. Assign each vertex a filtration value $f(v_{ij}) = \mathcal{L}(\alpha_i, \beta_j)$. Connect vertices by 8-adjacency (cardinal and diagonal neighbors), yielding edge set E. Each edge receives the *lower-star* weight:

$$w(u, v) = \max\big(f(u),\ f(v)\big). \tag{2}$$

This ensures an edge enters the filtration only after both its endpoints have appeared.

Stage 3: Persistent homology. We pass the weighted graph (V, E, w) as a sparse distance matrix to Ripser [Bauer, 2021], which constructs the Vietoris–Rips complex: a k-simplex $[v_0, \ldots, v_k]$ enters the filtration at $\max_{i<j} w(v_i, v_j)$. We compute H_0 (connected components) and H_1 (independent loops) persistence diagrams up to dimension 1.

For each homology dimension k, we record total persistence $\text{Pers}_k = \sum_i (d_i - b_i)$, feature count, and maximum lifetime. Because the landscape directions are random, each run samples a different 2D slice; we log the random seed for reproducibility. A limitation of this design is that **each architecture is characterized by a single random slice**; multi-slice stability analysis is planned but not yet complete (see Section 6).

3.3 Phase 2b: Baseline Geometry Metrics

Alongside topology, we compute four standard geometry metrics at the same minimum:

1. **Hessian trace** via Hutchinson's estimator [Hutchinson, 1989] with 10 Rademacher samples in fp64.
2. **Max Hessian eigenvalue** (sharpness) via 30-iteration power method [Keskar et al., 2017].
3. **Fisher information trace**: $\text{Tr}(\mathbf{F}) = \mathbb{E}\left[\sum_i \left(\frac{\partial \log p}{\partial \theta_i}\right)^2\right]$ over 10 batches.
4. **Loss barrier height**: Maximum loss increase along 10 filter-normalized random directions (20 steps, step size 0.1), normalized by $\sqrt{|\boldsymbol{\theta}|}$.

3.4 Phase 3: Sequential Forgetting Measurement

Starting from the Task A checkpoint, we expand the final classification layer from 50 to 100 outputs (preserving Task A weights) and train on Task B (classes 50–99) with naive SGD at learning rate 0.01 ($\frac{1}{10}$ of Phase 1). No continual learning regularization is applied—this measures the *bare* forgetting dynamics.

Task A test accuracy is evaluated at steps $\{100, 500, 1000, 5000, 10000, 25000\}$. Our primary retention metric is:

$$\text{ret@}k = \frac{\text{acc}_A(\text{step} = k)}{\text{acc}_A(\text{step} = 0)} \tag{3}$$

We also compute the area under the retention curve (AURC) via trapezoidal integration over the evaluation steps. AURC captures cumulative retention across the full trajectory rather than a single snapshot. Note that AURC values are *not* normalized to $[0, 1]$: they scale with both retention magnitude and the number of evaluation steps, so architectures with sustained high retention (e.g., ViT-Small, AURC = 274) produce much larger values than those that forget immediately (e.g., ResNet-18, AURC = 0.2).

3.5 Phase 4: Correlation Analysis

We compute Spearman rank correlations between each metric (topological and baseline) and each retention metric (ret@100, AURC) across all $n = 14$ architectures. To assess robustness, we perform leave-one-out (LOO) cross-validation: for each of the 14 architectures, we drop it from the dataset and recompute ρ and p on the remaining 13. We report the fraction of LOO folds that maintain $p < 0.05$ and the most influential architecture (largest $|\Delta\rho|$ when removed).

4 Experimental Setup

Dataset. Split-CIFAR-100: Task A = classes 0–49 (25,000 train / 5,000 test), Task B = classes 50–99. Standard augmentation (random crop with padding 4, horizontal flip) for training; no augmentation at test time. Images normalized to CIFAR-100 channel statistics.

Architectures. We evaluate 14 architectures spanning five computational paradigms (Table 1). CNN-based architectures are adapted for 32×32 input with 3×3 initial convolutions (stride 1); ViT variants use patch size 4 with 32×32 input. All models are trained from scratch with seed 42.

Landscape Sampling. 50×50 grid (2,500 points) over $[-1.0, 1.0]^2$ in filter-normalized directions. Landscape seed randomized per run to sample different 2D slices; seeds logged for reproducibility. GPU-resident dataset and mixed-precision forward passes for efficiency.

Reproducibility. Global seed 42. All hyperparameters in per-architecture YAML configs. Code available at `https://github.com/axiondeep/axiondeep-research`.

Table 1: Architectures evaluated, with parameter counts and Task A test accuracy after 100 epochs of training.

Architecture	Params	Acc_A	Type
ResNet-18	11.2M	82.0%	CNN
ResNet-50	23.6M	83.6%	CNN
ResNet-18 Wide	44.7M	83.1%	CNN
WRN-28-10	36.5M	84.0%	CNN
DenseNet-121	7.0M	84.5%	CNN
VGG-16-BN	15.0M	78.4%	CNN
EfficientNet-B0	4.1M	76.6%	CNN+SE
MobileNet-V3-S	1.5M	68.6%	CNN+SE
ShuffleNet-V2	1.3M	76.8%	CNN
RegNet-Y-400MF	4.3M	72.2%	CNN+SE
ConvNeXt-Tiny	28.0M	56.7%	Modern CNN
ViT-Small	3.0M	62.2%	Transformer
ViT-Tiny	0.8M	52.7%	Transformer
MLP-Mixer	2.3M	61.5%	MLP

Table 2: Persistent homology features of loss landscapes (50×50 grid). All architectures produce 2,499 finite H_0 features. H_1 features (loops) are resolved at this grid density for 9 of 14 architectures.

Architecture	H_0 Pers	H_1 Pers	H_1 Ct
ConvNeXt-Tiny	29366	0.11	1
ViT-Small	15984	0.32	2
MLP-Mixer	15799	0.00	0
ViT-Tiny	15015	0.18	9
MobileNet-V3-S	14804	1.90	19
EfficientNet-B0	14341	2.12	28
RegNet-Y-400MF	11349	0.02	6
ShuffleNet-V2	10694	0.69	70
VGG-16-BN	10053	0.00	0
WRN-28-10	9029	0.08	14
ResNet-18	8408	0.00	0
DenseNet-121	8020	0.26	2
ResNet-50	6410	0.00	0
ResNet-18 Wide	6027	0.00	0

5 Results

5.1 Topological Features Vary Across Architectures

Table 2 presents persistent homology features computed on each architecture's loss landscape at 50×50 resolution. H_0 total persistence spans a 4.9× range (6,027 to 29,366), and—critically—H_1 features now appear for 9 of 14 architectures (compared to only 1 of 8 at 25×25 resolution in preliminary runs), demonstrating the importance of grid resolution for capturing higher-dimensional topology.

Key observations:

1. H_0 persistence is *not* a proxy for model size. ConvNeXt-Tiny (28M params) has the highest H_0, while ResNet-18 Wide (44.7M) has the lowest.

2. EfficientNet-B0 and ShuffleNet-V2 show the richest H_1 structure (28 and 70 loops respectively), suggesting their architectural motifs (squeeze-and-excitation, channel shuffling) create qualitatively distinct landscape topology.

3. MLP-Mixer, VGG-16-BN, ResNet-18, ResNet-50, and ResNet-18 Wide show zero H_1 even at 50×50 resolution.

5.2 Forgetting Dynamics Are Architecture-Dependent

Table 3 shows Task A retention during sequential Task B training. All architectures exhibit catastrophic forgetting, but the rate varies by two orders of magnitude.

Three patterns emerge (Figure 1):

1. **Transformer advantage.** ViT-Tiny (22.5%) and ViT-Small (9.6%) show the highest ret@100, with ViT-Small uniquely maintaining 1.35% retention at 10,000 steps—the only architecture with measurable knowledge past this horizon.

2. **Lightweight CNN resilience.** ShuffleNet-V2 (17.3%), MobileNet-V3-S (7.6%), and EfficientNet-B0 (7.1%) outperform all large CNNs, suggesting that constrained capacity or specialized operations (channel shuffle, squeeze-and-excitation) may aid retention.

3. **Accuracy does not predict retention.** DenseNet-121 achieves the highest Task A accuracy (84.5%) but near-zero retention, while ViT-Tiny has the lowest accuracy (52.7%) but the highest ret@100.

4. **Non-monotonic forgetting.** ConvNeXt-Tiny shows 0% ret@100 but recovers to 3% at step 1,000 before returning to 0%. This transient recovery likely reflects Task B gradient updates temporarily aligning with Task A feature directions before diverging, rather than a measurement artifact (the same evaluation code is used for all architectures).

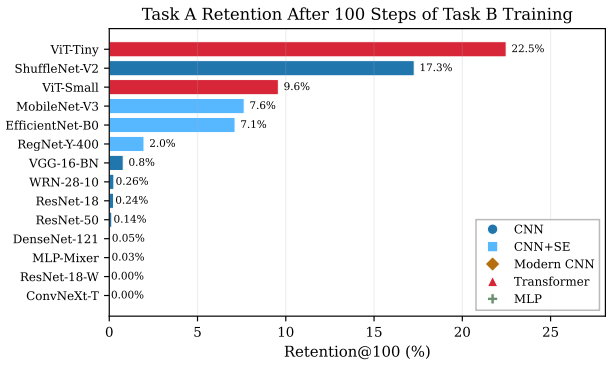

Figure 1: Task A retention after 100 steps of sequential Task B training, sorted by retention. Colors indicate architecture type. Vision Transformers and lightweight CNNs dominate; large standard CNNs retain near-zero knowledge.

5.3 H_1 Persistence Correlates with Retention

Table 4 presents Spearman rank correlations between each metric and retention. The central finding is that H_1 **total persistence significantly correlates with both ret@100** ($\rho = 0.61$, $p = 0.021$) **and AURC** ($\rho = 0.65$, $p = 0.012$), while H_0 does not reach significance for ret@100.

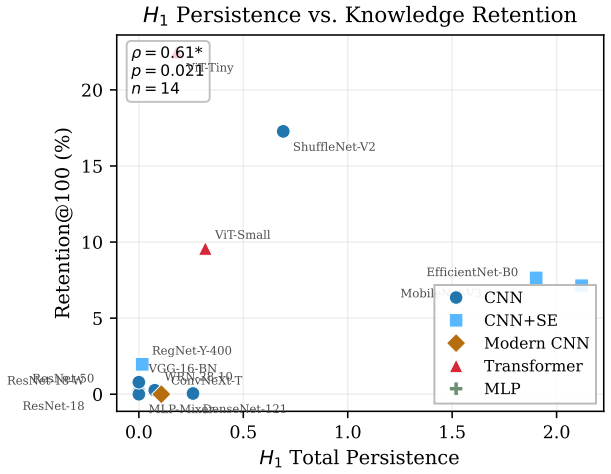

Figure 2: H_1 total persistence vs. retention@100 across 14 architectures. Points colored by architecture type; Spearman $\rho = 0.61$ ($p = 0.021$). Architectures with richer loop structure in their loss landscape retain more Task A knowledge.

This represents a shift from our preliminary $n = 8$ results, where H_0 showed the strongest trend ($\rho = 0.58$, $p = 0.13$). With the expanded architecture set and higher grid resolution, H_1 *emerges as the more informative topological predictor.* This is interpretable: H_1 features correspond to loop structures in the sublevel sets of the loss surface—closed ridges encircling basins. Architectures whose loss landscapes contain persistent loops may have more topologically complex basin structure that resists perturbation from new task gradients.

Fisher information trace shows a significant negative correlation with AURC ($\rho = -0.75$, $p = 0.002$) and marginal significance with ret@100 ($\rho = -0.50$, $p = 0.072$). High Fisher information indicates that model parameters are highly sensitive to the training data—and such sensitivity appears to make the model more vulnerable to catastrophic forgetting.

LOO stability. All three significant correlations are robust under leave-one-out cross-validation. For H_1 vs. ret@100: mean $\rho = 0.60 \pm 0.04$ across 14 folds, with **14/14 folds maintaining** $p < 0.05$. For H_1 vs. AURC: mean $\rho = 0.65 \pm 0.04$ (14/14 significant). For Fisher vs. AURC: mean $\rho = -0.75 \pm 0.03$

Table 3: Task A retention during sequential Task B training. ret@k = Task A accuracy at step k divided by initial accuracy. AURC = area under the retention curve (unnormalized; higher = more cumulative retention).

Architecture	ret@100	ret@1k	ret@10k	AURC
ViT-Tiny	22.48%	5.12%	0.19%	115.4
ShuffleNet-V2	17.27%	0.00%	0.00%	17.8
ViT-Small	9.58%	6.75%	1.35%	274.1
MobileNet-V3	7.64%	0.29%	0.00%	15.3
EfficientNet-B0	7.13%	0.05%	0.00%	9.1
RegNet-Y-400	1.97%	0.00%	0.00%	2.4
VGG-16-BN	0.79%	0.00%	0.00%	0.8
WRN-28-10	0.26%	0.00%	0.00%	0.4
ResNet-18	0.24%	0.00%	0.00%	0.2
ResNet-50	0.14%	0.00%	0.00%	0.1
DenseNet-121	0.05%	0.00%	0.00%	0.05
MLP-Mixer	0.03%	0.00%	0.00%	0.03
ConvNeXt-T*	0.00%	3.00%	0.00%	33.4
ResNet-18-W	0.00%	0.00%	0.00%	0.0

*Non-monotonic: 0% at step 100, recovers to 3% at step 1k.

Table 4: Spearman rank correlations between landscape metrics and forgetting resistance ($n = 14$ unless noted). Bold: $p < 0.05$.

Metric	ρ_{ret}	p	ρ_{AURC}	p	n
H_1 Pers.	**0.61**	**.021**	**0.65**	**.012**	14
H_0 Pers.	0.32	.263	**0.71**	**.005**	14
Fisher Tr.	−0.50	.072	**−0.75**	**.002**	14
Barrier (N)	0.54	.088	0.17	.612	11
Hessian Tr.	−0.12	.719	−0.49	.125	11
$\lambda_{\max}$	−0.14	.689	−0.50	.117	11
Barrier (raw)	−0.05	.864	−0.24	.418	14

(14/14 significant). The most influential architecture is DenseNet-121 for H_1 correlations—it has relatively high H_1 (0.26) but near-zero retention, pulling ρ down. Removing it *strengthens* ρ to 0.69, confirming that no single architecture drives the overall result.

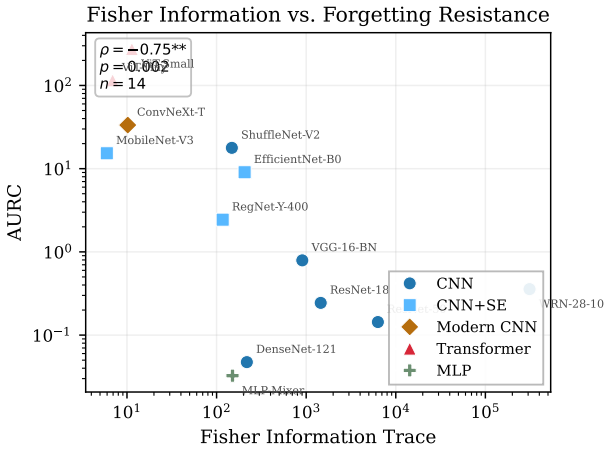

Figure 3: Fisher information trace vs. AURC (log–log scale). Spearman $\rho = -0.75$ ($p = 0.002$). Architectures with higher aggregate parameter sensitivity exhibit worse forgetting resistance.

Figure 2 shows the primary H_1–retention relationship and Figure 3 illustrates the Fisher–AURC anti-correlation.

5.4 The MLP-Mixer Counterexample

MLP-Mixer presents a challenge to any simple topology–retention hypothesis. Despite having the third-highest H_0 persistence (15,799) and moderate initial accuracy (61.5%), it exhibits zero retention (ret@100 = 0.03%). Critically, MLP-Mixer shows *zero H_1 features*, placing it among the architectures with the simplest higher-dimensional topology.

This supports the refined hypothesis: *H_1 structure, not H_0 alone, captures the topological property relevant to forgetting resistance.* H_0 measures how many isolated basins exist and their depth; H_1 measures the presence of closed ridges—topological barriers that may protect knowledge encoded in local minima from being overwritten by new task gradients.

5.5 Baseline Metric Comparison

Standard geometry metrics proved less reliable than topological features for cross-architecture comparison (Table 4). Three issues arose:

1. **Missing data.** Three architectures (ViT-Small, ViT-Tiny, MobileNet-V3) failed Hessian/eigenvalue computation due to memory constraints, reducing n from 14 to 11 for those metrics.

2. **Saddle points.** WRN-28-10 yielded a negative maximum Hessian eigenvalue ($\lambda_{\max} = -2{,}942$). This indicates the converged checkpoint lies at a saddle point rather than a local minimum—an inherent risk of stochastic optimization that invalidates sharpness as a basin descriptor for this architecture.

3. **Numerical overflow.** Raw loss barrier values for ResNet-18 and ResNet-50 overflowed to 10^{19}–10^{36}, likely due to filter-normalized perturbations pushing parameters into degenerate loss regions. These values are clamped at 10^6 in our analysis but illustrate the brittleness of barrier estimation for large models.

Persistent homology avoids all three failure modes: it operates on a 2D array of scalar loss values and requires no second-order gradients, no eigendecomposition, and no parameter-space perturbation beyond the landscape grid itself. A summary of all metric correlations is shown in Figure 4.

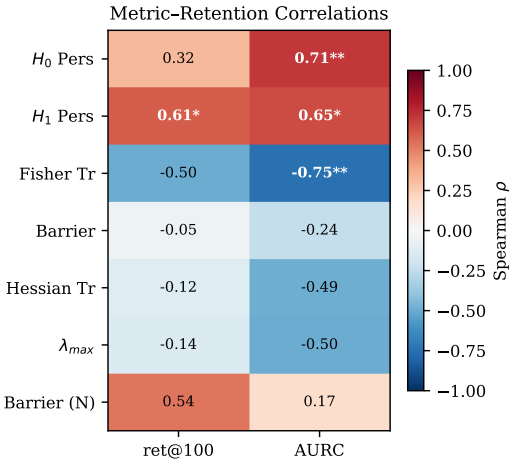

Figure 4: Spearman rank correlations between landscape metrics and retention measures. Stars indicate significance ($^*p < 0.05$, $^{**}p < 0.01$). H_1 persistence and Fisher trace show the strongest associations.

6 Discussion

H_1 as a Forgetting Correlate. The emergence of H_1 as the stronger correlate was unexpected. Initial hypotheses focused on H_0 (basin count and depth). However, H_1 features—loops in sublevel

sets—capture a qualitatively different property: the presence of closed ridges that separate regions of the loss surface. We hypothesize that these ridges act as topological barriers to gradient flow, slowing the migration of parameters away from Task A minima during Task B training. Architectures whose landscapes lack H_1 features (MLP-Mixer, VGG-16-BN, standard ResNets) have "smoother" topology that offers less resistance to parameter drift.

The Fisher–Forgetting Connection. The strong negative correlation between Fisher trace and AURC ($\rho = -0.75$) connects to elastic weight consolidation [Kirkpatrick et al., 2017], which uses Fisher information to identify "important" parameters. Our finding suggests an irony: architectures with higher aggregate Fisher information—where many parameters are individually important—are *more* vulnerable to forgetting, perhaps because task-specific knowledge is distributed across many sensitive parameters rather than concentrated in robust subspaces.

Transformer Advantage. Vision Transformers (ViT-Small, ViT-Tiny) show 40–90× better retention than standard CNNs. This aligns with emerging evidence that attention mechanisms create more modular representations [Dosovitskiy et al., 2021]. In our framework, this modularity manifests as richer H_1 topology in the loss landscape.

Limitations. The most important caveats concern unquantified variance:

1. **Single 2D slice per architecture.** Each topology is computed from one random 2D cross-section of a high-dimensional landscape. Different slices will yield different persistence diagrams. Until multi-slice runs confirm stability, the reported H_1 values carry unknown sampling variance.

2. **Single training seed.** All models are trained with seed 42. Variance across training seeds—which would change the converged minimum and thus the landscape—is not quantified. This is a confounder for any architecture-level claim.

3. **Single dataset.** All experiments use Split-CIFAR-100 with a fixed 50/50 class split. Generalization to other datasets, split ratios, and non-image domains is untested.

4. **Correlation, not causation.** Topology may correlate with forgetting through a shared confound (e.g., architectural inductive bias creating both flat landscapes and modular representations) rather than causally mediating it.

5. **Naive sequential baseline only.** Whether topology correlates with effectiveness of continual learning methods (EWC, replay) remains open.

6. **Grid resolution.** At 50×50, fine H_1 features may be under-resolved; 200×200 runs may reveal additional structure or change existing patterns.

Practical Implications. If the H_1–retention correlation generalizes beyond Split-CIFAR-100 and single-slice sampling, it could inform a practical diagnostic: after training on Task A, compute loss landscape topology to rank-order forgetting risk across candidate architectures *before* deploying in a continual learning setting. The computation requires only forward passes on a grid (no second-order gradients), making it feasible for large models. However, this application requires multi-dataset and multi-slice validation before deployment recommendations can be made.

7 Conclusion

We present evidence across 14 architectures that persistent homology features of neural network loss landscapes—specifically H_1 (loop) persistence—significantly correlate with resistance to catastrophic forgetting ($\rho = 0.61$, $p = 0.021$, stable under LOO). This outperforms H_0 persistence, Hessian-based sharpness, and loss barrier metrics. Fisher information trace provides a complementary negative association ($\rho = -0.75$, $p = 0.002$). These correlations are based on a single 2D landscape slice and single training seed per architecture; causal claims require further validation.

The MLP-Mixer counterexample demonstrates that the relationship is mediated by higher-dimensional topology (H_1) rather than simple basin count (H_0), and the transformer advantage in retention suggests that attention-based computation creates loss landscape topology conducive to knowledge preservation.

Future work will: (1) validate on multiple datasets (Split-ImageNet, CORe50, Permuted MNIST), (2) run multi-slice stability analysis with 3–5 random 2D sections per architecture, (3) increase grid resolution to 200×200, (4) run multiple training seeds per architecture, (5) compare against established CL baselines, and (6) investigate causal mechanisms linking

H_1 structure to gradient dynamics during sequential training.

All code, configurations, and results are available at `https://github.com/axiondeep/axiondeep-research`.

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.

Ulrich Bauer. Ripser: Efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5:391–423, 2021.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 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. In *International Conference on Learning Representations*, 2021.

Herbert Edelsbrunner and John L. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3 (4):128–135, 1999.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics—Simulation and Computation*, 18(3):1059–1076, 1989.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. In *Journal of Machine Learning Research*, volume 21, pages 1–40, 2020.

Karthikeyan Natesan Ramamurthy, Kush R. Varshney, and Krishnamurthy Mody. Topological data analysis of decision boundaries with application to model selection. In *International Conference on Machine Learning*, pages 5351–5360, 2019.

Sylvestre-Alvise Rebuffi, Alexander Kober, Hakan Bilen, and Andrea Vedaldi. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

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. In *International Conference on Learning Representations*, 2019.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. In *arXiv preprint arXiv:1606.04671*, 2016.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in Neural Information Processing Systems*, 30, 2017.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995, 2017.