TRANSFERABILITY ESTIMATION VIA PERSISTENT HOMOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting a pre-trained model for a downstream task is central to transfer learning, yet estimating transferability without fine-tuning remains challenging. We argue that this challenge reflects a deeper architectural question: whether the topology of a representation remains informative under domain shift. We study this question using 0-dimensional persistent homology. From this representation, we derive the Persistence Ratio (PR), a topology-based score that captures the balance between inter-class separation and intra-class variability in neural embeddings without the need for labels. Across ImageNet-pretrained CNNs and Vision Transformers evaluated on diverse downstream datasets, PR strongly correlates with fine-tuning performance, but only when topology is measured in the architecture-appropriate domain. Target-domain topology provides the most direct signal when it remains coherent. However, for Convolutional Neural Networks (CNNs), target embeddings often fragment before adaptation, making source-domain topology a more stable proxy for representation quality. For Vision Transformers (ViTs), by contrast, target embeddings preserve coherent cluster structure under domain shift, allowing topology measured directly on target data prior to fine-tuning to provide the most reliable estimates. This contrast shows that target-space transferability estimation is not universally reliable; its effectiveness depends on whether the architecture preserves meaningful target-domain topology before fine-tuning. These results reveal an architectural divide in representation geometry and position persistent homology as both a practical estimator for model selection and a diagnostic tool for understanding transferable representations.

1 INTRODUCTION

Selecting an optimal pre-trained model for a new task has become a central challenge in computer vision. As collections of pretrained vision backbones expand, fine-tuning every candidate model becomes computationally prohibitive. Transferability estimation (You et al., 2021; Tran et al., 2019; Nguyen et al., 2020) addresses this problem by predicting downstream performance prior to fine-tuning, enabling efficient model selection and reuse. However, transferability estimation also raises a more basic question: when does the geometry of a pretrained representation remain informative under domain shift?

Recent work has established a connection between the geometric structure of learned representations and their downstream performance (Ding et al., 2023; Wang et al., 2023; Suresh et al., 2023), in particular, with the phenomenon of neural collapse (Papyan et al., 2020), which shows that during the late stages of training, features of the same class concentrate around their class means, while features of different classes become increasingly separated in the embedding space. These observations suggest that transferable representations often exhibit compact intra-class structure and strong inter-class separation, motivating transferability estimators based on geometric properties of neural embeddings. In principle, measuring this structure on the target dataset is a direct way to assess transferability, since it probes the representation of the downstream data before fine-tuning. However, this signal is useful only if target embeddings preserve meaningful organisation under domain shift, and this reliability may depend strongly on the model architecture.

Building on this geometric perspective, our experiments reveal architectural differences in how the structure of representation behaves under domain shift. Convolutional networks (CNNs) tend to pro-

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

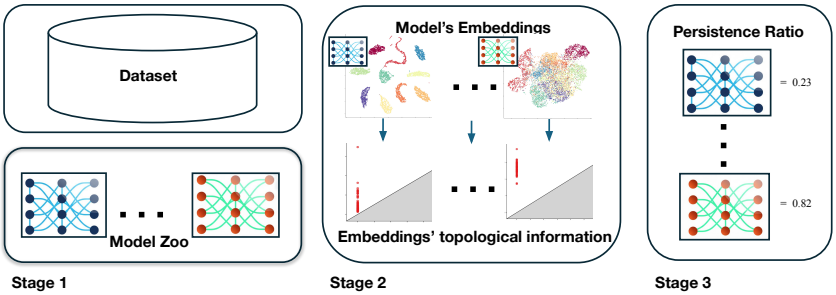

Figure 1: Overview of the proposed approach. Starting from a model zoo of pretrained networks, we compute embeddings by forwarding data through each backbone up to the penultimate layer. A Vietoris–Rips filtration is applied to the embeddings to obtain a persistence diagram, from which the persistence ratio is computed. Sorting models by this score yields a ranking that predicts downstream fine-tuning performance.

duce embeddings that become topologically fragmented on target datasets prior to fine-tuning, while Vision Transformers (ViTs) preserve coherent cluster structure on unseen datasets. This difference has important implications for transferability estimation, as it determines whether representation geometry can be reliably measured in the target domain or whether source-domain geometry provides a more stable proxy for the quality of the learned representation. In this sense, source topology is not treated as intrinsically preferable. Rather, it becomes useful when target topology is degraded by domain shift. Beyond improving transferability estimation, our results reveal architectural differences in how representation geometry behaves under domain shift, suggesting that the topology of embedding spaces encodes meaningful information about the robustness of learned features.

Currently, most existing approaches characterise the representation geometry using local statistics or second-order measures in the feature space. Such descriptors capture separation and variance, but do not explicitly describe the global connectivity structure of the embedding space. This distinction is relevant under domain shift, where class-wise statistics may not fully reveal whether target embeddings preserve coherent cluster connectivity, become fragmented, or develop spurious connections between clusters. Understanding how clusters connect, merge, or fragment across scales requires a more global perspective.

Topological data analysis (TDA) (Edelsbrunner et al., 2008; Edelsbrunner et al., 2002) provides tools to study this type of structure. In particular, persistent homology (PH) offers a multiscale description of the topology of point clouds by tracking how connected components appear and merge as the distance threshold increases. Applied to neural embeddings, persistent homology reveals how class clusters organize and interact across scales, distinguishing stable large-scale structures from transient local noise. This makes PH a natural tool for analysing whether representation topology remains informative when a pretrained model is evaluated outside its source domain.

In this work, we leverage PH to analyse the topology of neural embeddings and introduce a topology-based score for model selection. Given the embeddings produced by a pre-trained model, we compute a Vietoris–Rips filtration and extract the corresponding persistent diagram. Using this, we derive the *Persistence Ratio*, which compares the persistence of cluster separation with that of intra-cluster variability. Intuitively, representations in which classes form tight and well-separated clusters yield favourable persistence ratios.

We evaluate this measure on a diverse set of pretrained CNNs and Vision Transformers and compare it against several established transferability estimators. Across these benchmarks, the proposed topological score achieves the strongest mean correlation with downstream fine-tuning performance when evaluated in the architecture-appropriate domain.

Furthermore, the domain in which topology should be measured depends on the architecture. Although target-domain topology is the most direct signal when it remains coherent, for CNNs, embeddings of the target dataset prior to fine-tuning often become topologically fragmented under domain shift, making their geometry unreliable for transferability estimation. In this case, topology computed on the source dataset provides a more stable proxy for representation quality. In contrast,

ViTs preserve coherent topological structure even on unseen datasets, allowing topology computed directly on target embeddings to accurately predict downstream performance.

Finally, we investigate the relationship between topology and semantic learning. By training models on randomly permuted labels, we show that superficially favourable topological structures can arise in synthetic memorisation regimes but fail to transfer. This experiment highlights that topology captures meaningful representation structure only when it emerges from genuine learning. It also clarifies a limitation of source-domain topology: a favourable source organisation is predictive only when it reflects semantic feature learning rather than memorisation or source-specific shortcuts. Our major contributions are the following:

- We uncover an architectural divide in representation topology under domain shift: while CNN target embeddings become fragmented prior to adaptation, ViTs preserve coherent target-domain structure that remains predictive of transferability.
- We introduce the Persistence Ratio, a simple topology-based score for analysing the global structure of neural embeddings through 0-dimensional persistent homology.
- We show that PR achieves a strong correlation with fine-tuning performance when measured in the architecture-appropriate domain: source embeddings for CNNs and target embeddings for ViTs.
- We demonstrate that the predictive signal captured by topology reflects meaningful learned structure and disappears in synthetic memorisation regimes, highlighting the interpretability and limitations of the proposed measure.

2 Related Work

Transferability Estimation: Transferability estimation aims to rank pre-trained models based on their expected downstream performance before fine-tuning. Existing methods have been categorised by their required source access and by the form of the score, including probabilistic, information-theoretic, linear-evidence, matrix/spectral, feature-structure, and dynamic remapping approaches (Ding et al., 2024). Representative estimators include LEEP (Nguyen et al., 2020), NCE (Tran et al., 2019), SFDA (Shao et al., 2022), and LogME (You et al., 2021).

Within feature-structure approaches, neural-collapse-based methods (Papyan et al., 2020) connect transferability to the geometry of embeddings: features of the same class contract around their class mean while different class means become separated. Several scores (Ding et al., 2023; Wang et al., 2023; Suresh et al., 2023) exploit this behaviour through class-separation or compactness statistics. However, these measures typically rely on label-wise or second-order summaries, capturing only limited aspects of the embedding space's global organisation.

Our work belongs to this embedding-structure perspective but adopts a global, multiscale view: rather than evaluating separability only through class-wise statistics, we use persistent homology to track how embedding components merge across scales, yielding a broader characterisation of representation structure.

Topological Analysis of Neural Networks: Topological Data Analysis (TDA) offers tools for characterising high-dimensional data and model representations through topological invariants (Edelsbrunner et al., 2002; Hensel et al., 2021). It has been used to study neural-network capacity (Guss & Salakhutdinov, 2018), dataset and task complexity (Narayanan & Mitter, 2010; Ali et al., 2023), decision-boundary topology (Ramamurthy et al., 2019), architectural expressiveness (Yang et al., 2021; Carrière et al., 2020), topology regularization (Weeks & Rivera, 2021), and layer-wise representation changes (Naitzat et al., 2020; Papillon et al., 2023). These works primarily use topology as an analysis or diagnostic tool with a focus on understanding representation dynamics rather than estimating transferability across models.

The closest prior work to ours is the Topological Transferability Measure (TTP) (Suresh et al., 2024). TTP uses the layer-wise cubical homology (Piotr, 2006) of target embeddings and fits a polynomial to the decay of Betti numbers across layers. Although effective, TTP requires topological computation across multiple layers and introduces design choices through the polynomial fitting procedure.

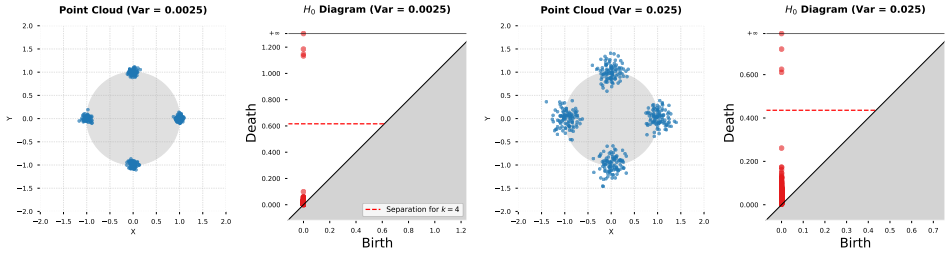

Figure 2: Persistent diagram of 4 equidistant classes sampled with different variance in a circle. A representation with k classes with high inter-cluster distance and low within-cluster variability should have $k-1$ homology modules with high persistence, and near zero on the remaining ones.

While prior geometric and statistical methods (Nguyen et al., 2020; Tran et al., 2019; You et al., 2021; Suresh et al., 2023) capture local properties of neural embeddings, and topological approaches such as Suresh et al. (2024) require layer-wise analysis, our approach focuses on the connectivity and characterises the global structure of the embedding space using persistent homology computed once on the representations. This yields a fast, versatile, and interpretable measure of representation topology that correlates strongly with downstream transfer performance.

3 Methodology

3.1 Overview and Motivation

Transferability is closely related to the organisation of a model's representation space. For a k-class dataset, compact and well-separated embedding clusters are expected to support easier separation and adaptation, consistent with neural-collapse geometry (Papyan et al., 2020). We quantify this structure with persistent homology, yielding global, multiscale summaries of embedding connectivity.

We focus on 0-dimensional persistent homology because it characterises cluster structure connectivity: large persistences correspond to inter-class separation, and small persistences reflect intra-class tightness. Aggregating this, we obtain metrics that summarise how a representation organises data.

3.2 Persistent Homology of Embeddings

Given a point cloud of embeddings $X = \{x_1, \ldots, x_n\}$ extracted from a pre-trained model, we construct a Vietoris-Rips filtration. For an increasing sequence of radii $R \geq 0$, the Vietoris-Rips complex $\mathrm{VR}(X, R)$ is formed by adding an edge between any pair of points $x_u, x_v \in X$ if their distance is $d(x_u, x_v) \leq R$. As the radius increases, isolated points connect and merge into larger connected components. 0-dimensional persistent homology (H_0) tracks the evolution of these connected components by recording their *birth* and *death* radii.

Specifically, in H_0, every point $x \in X$ appears as an independent connected component at radius $R = 0$, so the birth radius for all components is $b_i = 0$. As R increases, when two previously disjoint components become connected by an edge at a radius R_{merge}, the components merge. When components merge, one component continues to exist while the other is considered to die at $d_i = R_{\text{merge}}$, leaving at the end of the process just one connected component that dies at $R_{\text{merge}} = inf$.

The collection of birth-death pairs $\{(b_i, d_i)\}_{i=1}^{n}$ forms the H_0 persistence diagram. The persistence of each feature is defined as $p_i = d_i - b_i$. Since $b_i = 0$ for H_0, persistence corresponds directly to the merge radius d_i.

In well-structured representations, we expect a characteristic pattern: $k-1$ finite homology modules with large persistences corresponding to merges between class clusters, followed by small persistences capturing within-class variability (fig. 2). While persistent homology itself is agnostic to class labels, modern deep representations typically organise samples of the same class into compact

Algorithm 1 Computation of Topology-Based Measures

Require: Pre-trained model f, dataset S, number of classes k
1: Extract embeddings $X \leftarrow \{f(x) : x \in S\}$
2: Uniformally sample a subset $X' \subset X$ of points
3: Compute the H_0 persistence diagram D_0 on X'
4: Sort finite persistences $\{p_i\}_{i=1}^{n-1}$ in descending order
5: Compute $\mathcal{H}_{\text{sep}}$, $\mathcal{H}_{\text{tight}}$, $\mathcal{H}_{\text{ratio}}$ using eqs. (1) to (3)
6: **return** $(\mathcal{H}_{\text{sep}}, \mathcal{H}_{\text{tight}}, \mathcal{H}_{\text{ratio}})$

regions of the embedding space, making these large merges align closely with inter-class structure in practice.

3.3 Topology-Based Measures

Let $\{p_i\}_{i=1}^{n-1}$ denote the sorted persistences ($p_1 \geq \cdots \geq p_{n-1}$) of finite components. For a k-class problem, we define:

- Inter-class separation: the top $k-1$ persistences quantify the radii at which different components merge:

$$\mathcal{H}_{sep} = \frac{1}{k-1} \sum_{i=1}^{k-1} p_i \tag{1}$$

- Intra-class tightness: the remaining persistences reflect merges within components:

$$\mathcal{H}_{tight} = \frac{1}{n-k} \sum_{i=k}^{n-1} p_i \tag{2}$$

- Finally, the ratio of these measures summarises the relation between global connectivity and local connectivity:

$$\mathcal{H}_{ratio} = \frac{\mathcal{H}_{tight}}{\mathcal{H}_{sep}} \tag{3}$$

Representations that generalise well typically exhibit large separation, small within-class spread, and therefore a small ratio.

3.4 Computational Pipeline

We compute persistent homology once per model on a fixed-size subset of embeddings, making the procedure scalable and parameter-free. The full pipeline is summarised in Algorithm 1.

This procedure is simple and efficient while capturing the global topology of the embedding space in a compact form that summarises its cluster structure.

3.5 Source and Target Topology Under Domain Shift

Target-domain topology is the most direct signal for transferability when it remains coherent, because it measures how downstream samples are organised before fine-tuning. However, this signal can be degraded by domain shift: target embeddings may become fragmented, weakly connected, or organised around non-semantic structure. Source-domain topology, by contrast, measures the global organisation learned during pre-training and can serve as a more stable proxy for representation quality when target topology is unreliable.

This suggests that the informative domain for topological transferability estimation need not be universal. If an architecture preserves meaningful target-domain topology, then target measurements should provide the most direct estimate of transferability. If the target topology is disrupted by a

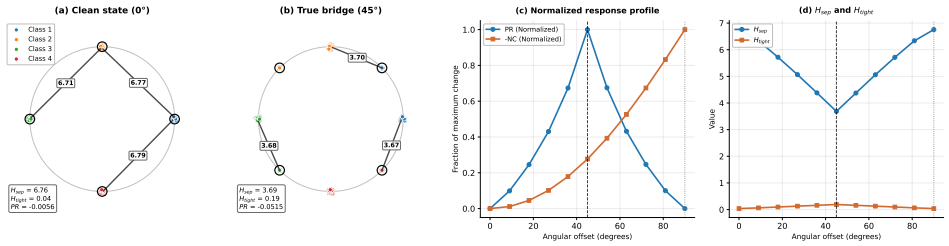

Figure 3: Sensitivity to low-density bridges between clusters. We simulate a four-class collapsed representation and rotate one point per class. At $45°$, the points create bridges between adjacent classes, strongly perturbing global connectivity. Unlike neural-collapse (NC) statistics, PR is most sensitive to these connectivity shortcuts rather than to points embedded entirely within the wrong cluster.

domain shift, source measurements may better reflect the semantic organisation learned during pre-training. Our experiments in Section 4 show that this distinction is architecture-dependent: target topology is most reliable for ViTs, whereas source topology is more stable for CNNs.

While $\mathcal{H}_{ratio}$ captures aspects of the neural collapse process, it is not limited to label-conditioned second-order statistics. As training organises the embedding space, class clusters are expected to become more compact and more clearly separated, leading to larger $\mathcal{H}_{sep}$, smaller $\mathcal{H}_{tight}$, and a lower $\mathcal{H}_{ratio}$. These dynamics provide an intuitive interpretation of the quantities measured by our method.

However, $\mathcal{H}_{ratio}$ differs from standard neural collapse proxies. Existing NC-based scores primarily summarise label-conditioned second-order statistics, such as class means and within-class variances. In contrast, $\mathcal{H}_{ratio}$ captures the multiscale connectivity of the embedding space through H_0 persistent homology, tracking the radii at which disconnected regions become connected during the filtration process.

As the filtration radius increases, connected components merge at scales that reflect spatial connectivity rather than label-conditioned statistics. Consequently, these merged scales are sensitive to low-density structures that create shortcuts between clusters. As illustrated in Figure 3, such bridge points can strongly alter global connectivity while having a more indirect effect on class means and variances. Conversely, samples embedded entirely within the wrong cluster may strongly affect neural-collapse statistics while leaving global connectivity comparatively unchanged. This distinction highlights a key property of $\mathcal{H}_{ratio}$: it exposes connectivity vulnerabilities in the embedding space.

In practical embedding spaces, such bridges may arise from ambiguous features, shared low-level patterns, background textures, or dataset-specific correlations. Similar phenomena have been discussed in *shortcut learning*, where models rely on non-semantic cues present in the training data (Geirhos et al., 2020; Ilyas et al., 2019).

By quantifying both inter-class separation, $\mathcal{H}_{sep}$, and intra-class tightness, $\mathcal{H}_{tight}$, the ratio penalises representations with weak separation, diffuse clusters, or shortcut connections between clusters. This sensitivity to global connectivity is precisely what makes topology useful for diagnosing whether source or target embeddings provide a reliable transferability signal under domain shift.

4 Experiments and Results

We evaluate the proposed topology-based transferability measure as a predictor of downstream transfer performance. Our experiments address three questions: (i) how accurately the metric predicts fine-tuning performance compared with established transferability estimators, (ii) how its behaviour differs across major architectural families (CNNs vs Vision Transformers), and (iii) whether favourable topology alone is sufficient for transferability, or whether its predictive power depends on the presence of semantically meaningful features.

4.1 PREDICTING FINE-TUNING PERFORMANCE

We begin by evaluating if the proposed measure can reliably rank pre-trained models according to their downstream fine-tuning performance. Because transferability estimators aim to predict the *relative* suitability of models, the standard evaluation protocol compares the ranking induced by each heuristic with the ranking induced by fine-tuned accuracies. The agreement between the predicted and ground-truth rankings is measured using Kendall's τ (Kendall, 1938). We use this evaluation not only to compare PR against existing estimators, but also to test whether source-domain or target-domain topology provides the more reliable signal for each architectural family. Additional correlation analysis, including Weighted Kendall in Appendix D

For CNNs, we evaluate a diverse set of ImageNet-1k pre-trained models from PyTorch (Paszke et al., 2019) model zoo, following the benchmark from You et al. (2021). The target datasets include Aircraft (Maji et al., 2013), Birdsnap (Berg et al., 2014), Caltech-101 (Fei-Fei et al., 2006), Stanford Cars (Krause et al., 2013), CIFAR-10/100 (Krizhevsky et al., 2009), DTD (Cimpoi et al., 2014), Oxford Pets (Parkhi et al., 2012), and SUN397 (Xiao et al., 2010). For Vision Transformers (ViTs), we follow the benchmark of Shao et al. (2022), which evaluates models trained using the Timm library (Wightman, 2019) across a broader set of datasets, including Oxford Flowers 102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), and Pascal VOC 2007 (Everingham et al.).

For ranking purposes, we define the Persistence Ratio score as: $\mathrm{PR} = -\mathcal{H}_{\text{ratio}}$, so that higher values correspond to better transferability. Because smaller values of $\mathcal{H}_{\text{ratio}}$ indicate tighter intra-class clusters and stronger inter-class separation, this transformation aligns the direction of our metric with existing transferability estimators.

Table 1: Kendall's τ results on evaluated benchmarks. $\mathrm{PR_s}$ and $\mathrm{PR_t}$ denote PR computed on source and target embeddings. PR values are averaged over ten uniform subsamples. Mean uncertainty is reported as a 95% percentile bootstrap confidence interval over downstream datasets. Higher is better. Bold indicates the best and underline the second-best result within each dataset.

(a) CNNs

Dataset	LEEP	NCE	LogME	SFDA	NCTI	$\mathrm{PR_t}$	$\mathrm{PR_s}$
Aircraft	0.289	0.511	0.360	<u>0.644</u>	0.333	**0.649**	0.556
Birdsnap	0.333	0.644	0.467	-0.067	0.644	**0.684**	<u>0.683</u>
Caltech101	0.422	0.733	0.584	0.644	**0.778**	<u>0.756</u>	0.689
StanfordCars	0.424	0.566	0.519	0.424	<u>0.660</u>	0.292	**0.666**
CIFAR10	0.477	0.159	<u>0.705</u>	**0.750**	<u>0.705</u>	0.377	0.615
CIFAR100	0.422	0.111	0.629	<u>0.733</u>	**0.778**	-0.053	0.646
DTD	0.022	-0.289	<u>0.494</u>	0.490	0.111	-0.556	**0.701**
Pets	0.467	**0.733**	0.511	0.422	<u>0.644</u>	0.316	0.559
SUN397	0.449	0.764	0.719	0.360	**0.854**	0.382	<u>0.811</u>
Mean	0.367	0.437	0.554	0.489	<u>0.612</u>	0.316	**0.658**
95% CI	[0.267, 0.439]	[0.197, 0.640]	[0.483, 0.626]	[0.321, 0.628]	[0.452, 0.742]	[0.041, 0.542]	[0.612, 0.708]

(b) ViTs

Dataset	LogME	SFDA	NCTI	$\mathrm{PR_t}$	$\mathrm{PR_s}$
Aircraft	0.422	0.156	0.289	<u>0.427</u>	**0.560**
Caltech101	<u>0.467</u>	0.422	<u>0.467</u>	**0.569**	0.462
StanfordCars	0.556	0.378	0.511	<u>0.667</u>	**0.693**
CIFAR10	**0.523**	<u>0.477</u>	<u>0.477</u>	0.400	0.209
CIFAR100	**0.689**	0.556	0.511	<u>0.627</u>	0.382
DTD	0.556	<u>0.600</u>	0.289	<u>0.600</u>	**0.649**
Flowers102	0.156	0.552	0.156	<u>0.689</u>	**0.738**
Food101	0.467	0.511	0.289	**0.538**	<u>0.516</u>
Pets	<u>0.778</u>	**0.822**	**0.822**	0.569	0.604
SUN397	<u>0.600</u>	0.422	0.289	**0.618**	0.560
VOC2007	<u>0.764</u>	**0.899**	0.719	0.665	0.589
Mean	<u>0.543</u>	0.527	0.438	**0.579**	0.542
95% CI	[0.440, 0.637]	[0.414, 0.645]	[0.329, 0.556]	[0.523, 0.628]	[0.453, 0.619]

To investigate where topology should be measured, we compute the Persistence Ratio both on the source dataset, ImageNet-1k, denoted $\mathrm{PR_s}$, and on each target dataset before fine-tuning, denoted $\mathrm{PR_t}$, as described in Section 3. This comparison directly tests the central question of whether target topology remains informative under domain shift, or whether source topology provides a more stable proxy for representation quality.

Ground-truth fine-tuning accuracies are obtained from the widely used benchmarks of You et al. (2021) for CNNs and Shao et al. (2022) for Vision Transformers. For each heuristic ($\mathrm{PR_s}$, $\mathrm{PR_t}$, LEEP (Nguyen et al., 2020), NCE (Tran et al., 2019), LogME (You et al., 2021), SFDA (Shao et al., 2022), and NCTI (Wang et al., 2023)), we compute Kendall's τ between the predicted ranking and the ground-truth ranking for each dataset. Table 1 presents the results. LEEP and NCE are not computed for Transformers, as not all models were pretrained in a supervised manner.

Our results reveal an architectural divide in which the topological signal is most informative. For ViTs, $\mathrm{PR_t}$ emerges as the strongest transferability estimator among all evaluated methods, with a mean Kendall's τ of 0.579. This supports the view that the target-domain topology is the most direct signal when the architecture preserves coherent target embeddings prior to fine-tuning.

In contrast, CNNs exhibit a different pattern. Computing PR on the source dataset, $\mathrm{PR_s}$, provides the most reliable transferability estimate, achieving the highest mean Kendall's τ (0.658), across downstream datasets. Conversely, applying the target-based variant, $\mathrm{PR_t}$, to CNNs yields volatile results: it performs well on some fine-grained datasets, such as Aircraft and Birdsnap, but degrades severely on others, such as DTD. Thus, for CNNs, source topology should be interpreted as a more

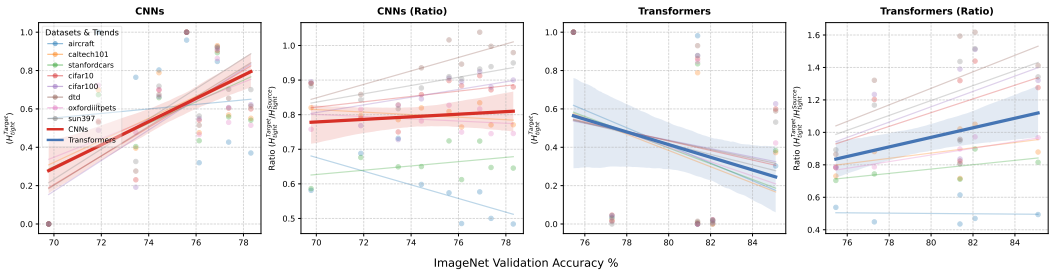

Figure 4: Topological divergence under domain shift. Target tightness $\mathcal{H}_{\text{tight}}^{\text{Target}}$ and its target-to-source ratio $\mathcal{H}_{\text{tight}}^{\text{Target}}/\mathcal{H}_{\text{tight}}^{\text{Source}}$ are plotted against ImageNet validation accuracy. CNN target clusters become looser as accuracy improves, whereas ViTs preserve or improve target compactness. This indicates that target-domain topology is reliable only when it remains coherent before fine-tuning.

stable proxy when target topology is degraded, rather than as an intrinsically superior measurement domain. This instability motivates the analysis in Section 4.2.

Overall, PR achieves the strongest mean correlation among the evaluated estimators across both CNN and Transformer benchmarks when evaluated in the architecture-appropriate domain. More importantly, these results show that the reliability of target-space transferability estimation is itself architecture-dependent.

4.2 Analysis on the Differences Between CNN and Transformers

The contrast in the predictive power of $\mathrm{PR_t}$ between CNNs and ViTs raises a fundamental question about how different architectures represent out-of-distribution data. As shown in Section 4.1, target-based topology fails for CNNs but succeeds for ViTs. This suggests that target-domain transferability estimation is reliable only when the architecture preserves coherent representation topology before fine-tuning.

To understand this discrepancy, we analyse the absolute target intra-class variance, $\mathcal{H}_{\text{tight}}$, across both model families as a function of ImageNet pre-training accuracy. The results, shown in Figure 4, reveal a divergent geometric behaviour. For CNNs, higher source accuracy correlates with looser and more fragmented clusters in the target domain. In contrast, for ViTs, higher source accuracy correlates with significantly tighter clusters under domain shift.

We also analyse how $\mathcal{H}_{\text{tight}}$ evolves between the source and target datasets. For Transformers, target clusters remain compact and often tighten as models improve. For CNNs, clusters expand substantially when evaluated on out-of-distribution data, weakening global cluster structure. Thus, the poor performance of $\mathrm{PR_t}$ for CNNs should not be interpreted as evidence against target-space transferability estimation, but as evidence that CNN target topology is often degraded before adaptation.

These distinct behaviours help explain the contrasting performance of $\mathrm{PR_t}$ across architectures. Because ViTs preserve cohesive cluster structures even on unseen datasets, target embeddings retain meaningful topology, allowing $\mathrm{PR_t}$ to accurately predict transferability. CNN embeddings, in contrast, become unstable under domain shift, making target-based topology unreliable. In this setting, source-domain topology serves as a more stable proxy for the quality of the representation learned during pre-training, rather than as a universally preferable domain for measuring transferability.

We hypothesise that these dynamics stem from the different inductive biases of the architectures. Prior work has shown that Vision Transformers exhibit stronger and more symmetric neural collapse behaviour than CNNs (Ammar et al., 2024). Moreover, their global self-attention mechanism has been associated with representations that remain coherent under domain shift (Naseer et al., 2021). These properties are consistent with the preservation of target-domain topology observed in Fig. 4.

CNNs, which rely more heavily on localised feature extraction, appear more vulnerable to domain-specific variations. As a result, their target embeddings can lose global cluster integrity, explaining why $\mathrm{PR_t}$ becomes unreliable while $\mathrm{PR_s}$ remains predictive.

Table 2: **Effect of pre-training regime on topology and downstream performance.** We compare three CIFAR-100 pre-training regimes and their resulting STL-10 fine-tuning accuracy: *Standard*, *Overtrained*, and *Memorized*. We report $\mathcal{H}_{\text{sep}}$, $\mathcal{H}_{\text{tight}}$, $\mathcal{H}_{\text{ratio}}$, and fine-tuned accuracy.

Pre-training Regime	**Fine-tune Acc.** (STL-10)	$\mathcal{H}_{ratio}$	$\mathcal{H}_{sep}$	$\mathcal{H}_{tight}$
Standard (Best Val Loss)	0.7113	0.7174	7.9705	5.7184
Overtrained (Best Train Loss)	**0.7345**	**0.2272**	10.2992	2.3397
Memorized (Permuted Labels)	0.4103	**0.2630**	10.3123	2.7124

4.3 Analysis of Training Regimes and Topology

A natural question raised by our results is whether a favourable topological structure (low $\mathcal{H}_{\text{ratio}}$) is sufficient to achieve good transferability, regardless of whether the learned features are semantically meaningful. To investigate this, we exploit the ability of neural networks to fit their training sets even when labels are randomly permuted (Zhang et al., 2017). Training under such conditions forces the network to memorise the dataset without learning meaningful semantic features.

We train ResNet-18 on CIFAR-100 and evaluate transfer to STL-10 under three pre-training regimes: (i) training run where the model is selected based on validation loss, (ii) an overtrained variant from the same run selected based on the best training loss, and (iii) a memorisation regime with permuted labels. All runs train for 200 epochs with learning-rate reductions on plateau.

For fine-tuning on STL-10, we perform a grid search over learning rates $\{0.01, 0.001\}$ and weight decays $\{0.001, 0.0001, 0.00001\}$, training until validation loss plateaus and reporting the best test accuracy obtained for each pre-trained model. Results are summarised in Table 2.

The memorised model achieves a low $\mathcal{H}_{\text{ratio}}$ (0.2630), close to the overtrained model (0.2272), yet its transfer performance collapses (41% accuracy). This regime is intentionally synthetic and designed to destroy semantic structure. The experiment, therefore, isolates a boundary case: favourable topology alone is not sufficient for transferability if the learned features are not semantically meaningful.

In contrast, under natural training regimes, the expected relationship holds. The model selected by validation loss exhibits a higher $\mathcal{H}_{\text{ratio}}$ (0.7174) and achieves lower transfer performance than the overtrained model. The overtrained model develops a more mature embedding topology, which achieves the highest fine-tuning accuracy (0.7345). This suggests that continued training beyond the validation optimum could further refine the embedding space in ways that benefit transferability. Crucially, unlike the memorisation regime, these topological improvements arise from meaningful feature learning and consistently align with downstream performance.

5 Conclusion

We showed that transferability estimation depends on how different architectures preserve representation topology under domain shift. Using H_0 persistent homology, we introduced the Persistence Ratio, which quantifies the balance between inter-class separation and intra-class compactness. Across ImageNet-pretrained CNNs and ViTs, PR achieves the highest mean correlation with fine-tuned accuracy when evaluated in the architecture-appropriate domain: source embeddings for CNNs and target embeddings for ViTs.

This architectural divide shows that target-domain topology is the most direct transferability signal when it remains coherent, but that this condition is not guaranteed under domain shift. CNN target embeddings often fragment before adaptation, making source topology a more stable proxy for representation quality, whereas ViTs preserve coherent target structure that remains predictive of downstream performance. Finally, memorisation experiments show that favourable topology is not sufficient for transferability unless it reflects semantic feature learning. Overall, persistent homology provides both a practical tool for model selection and a lens for understanding when representation geometry remains meaningful under domain shift.

AI USE STATEMENT

We used ChatGPT and Gemini to assist with drafting and editing parts of the manuscript. We used ChatGPT to assist with the implementation, debugging, and modification of experimental and evaluation code. All AI-assisted text and code were reviewed by the authors, experimental outputs were checked for correctness and consistency, and all reported results and claims were verified by the authors. The authors take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

All models and datasets used are publicly available; when downloaded, they are from PyTorch when available, their official repository, and Huggingface when unavailable on either of the previous. Models were downloaded from Pytorch and TiMM. Models' fine-tuning results were extracted from the respective papers, and the ones used are also shown in the appendix E. All relevant implementation details can be found in section 3. Code will be uploaded after publication.

REFERENCES

Dashti Ali, Aras Asaad, Maria-Jose Jimenez, Vidit Nanda, Eduardo Paluzo-Hidalgo, and Manuel Soriano-Trigueros. A survey of vectorization methods in topological data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. ISSN 0162-8828.

Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection, 2024.

Nieves Atienza, Rocio Gonzalez-Diaz, and Matteo Rucco. Separating topological noise from features using persistent entropy. In Paolo Milazzo, Dániel Varró, and Manuel Wimmer (eds.), *Software Technologies: Applications and Foundations*, pp. 3–12, Cham, 2016. Springer International Publishing. ISBN 978-3-319-50230-4.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2018, 2014.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 446–461, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 2786–2796. PMLR, 2020. ISBN 2640-3498.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Yuhe Ding, Bo Jiang, Lijun Sheng, Aihua Zheng, and Jian Liang. Unleashing the power of neural collapse for transferability estimation, 2023. URL https://arxiv.org/abs/2310.05754.

Yuhe Ding, Bo Jiang, Aijing Yu, Aihua Zheng, and Jian Liang. Which model to transfer? a survey on transferability estimation. *ArXiv*, abs/2402.15231, 2024. URL https://api.semanticscholar.org/CorpusID:267897613.

Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, Nov 2002. ISSN 1432-0444. doi: 10.1007/s00454-002-2885-2. URL https://doi.org/10.1007/s00454-002-2885-2.

Herbert Edelsbrunner, John Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453(26):257–282, 2008.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006. doi: 10.1109/TPAMI.2006.79.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

William H Guss and Ruslan Salakhutdinov. On characterizing the capacity of neural networks using algebraic topology. *arXiv preprint arXiv:1802.04443*, 2018.

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

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81.

Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars, 2013. URL https://ai.stanford.edu/jkrause/papers/fgvc13.pdf.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft, 2013. URL https://arxiv.org/pdf/1306.5151.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020. URL http://jmlr.org/papers/v21/20-345.html.

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. *Advances in neural information processing systems*, 23, 2010.

Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23296–23308. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c404a5adbf90e09631678b13b05d9d7a-Paper.pdf.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/nguyen20b.html.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Mathilde Papillon, Sophia Sanborn, Mustafa Hajij, and Nina Miolane. Architectures of topological deep learning: A survey on topological neural networks. *arXiv preprint arXiv:2304.10031*, 2023.

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.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Kot Piotr. Homology calculation of cubical complexes in rn. *CMST*, 12(2):115–121, 2006.

Karthikeyan Natesan Ramamurthy, Kush Varshney, and Krishnan Mody. Topological data analysis of decision boundaries with application to model selection. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5351–5360. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ramamurthy19a.html.

Wenqi Shao, Xun Zhao, Yixiao Ge, Zhaoyang Zhang, Lei Yang, Xiaogang Wang, Ying Shan, and Ping Luo. Not all models are equal: Predicting model transferability in a self-challenging fisher space. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 286–302, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19830-4.

S. Suresh, B. Das, V. Abrol, and S. Dutta Roy. On characterizing the evolution of embedding space of neural networks using algebraic topology. *Pattern Recognition Letters*, 179:165–171, 2024. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2024.02.003. URL https://www.sciencedirect.com/science/article/pii/S0167865524000369.

Suryaka Suresh, Vinayak Abrol, and Anshul Thakur. Pitfalls in measuring neural transferability. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023. URL https://openreview.net/forum?id=KWUA0n6Dpv.

Anh T. Tran, Cuong V. Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5549–5558, October 2023.

Deborah Weeks and Samuel Rivera. Domain adaptation by topology regularization. In *Automatic Target Recognition XXXI*, volume 11729, pp. 72–89. SPIE, 2021.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, June 2010. doi: 10.1109/CVPR.2010.5539970.

Ji Yang, Lu Sang, and Daniel Cremers. Dive into layers: Neural network capacity bounding using algebraic geometry. *arXiv preprint arXiv:2109.01461*, 2021.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/you21b.html.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017. URL https://arxiv.org/abs/1611.03530.

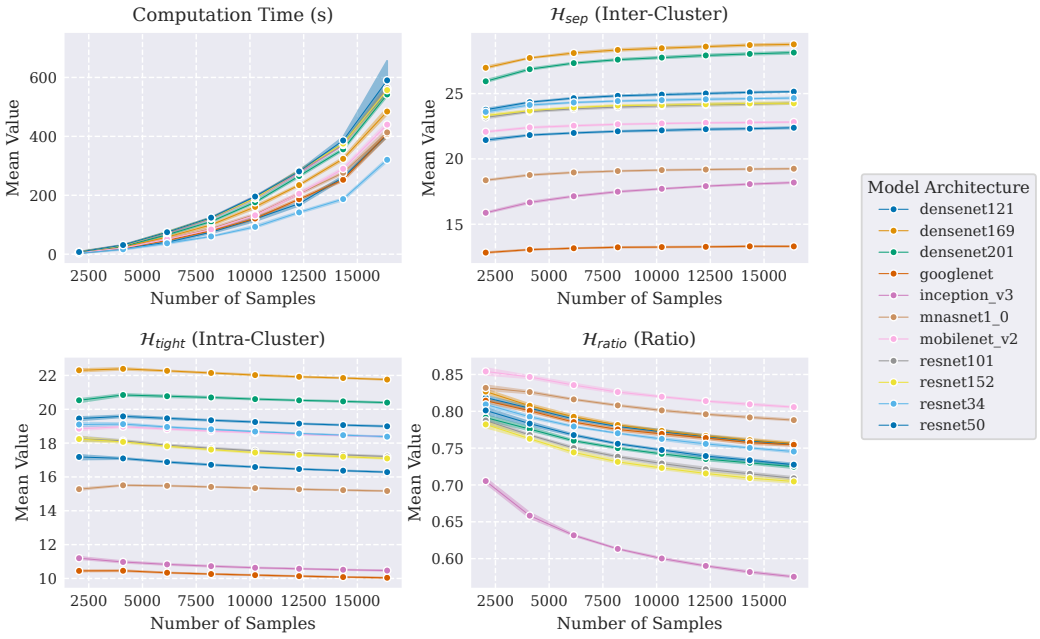

Figure 5: **Scalability of the proposed topological measures.** Computation time (top-left) increases with the number of sampled embeddings, consistent with the polynomial complexity of Vietoris–Rips persistent homology. The remaining panels show $\mathcal{H}_{\text{sep}}$, $\mathcal{H}_{\text{tight}}$, and $\mathcal{H}_{\text{ratio}}$ for different sampling sizes across multiple ImageNet-pretrained models. All measures exhibit highly similar trends across sampling sizes, indicating that a few thousand samples (e.g., 2,048 points) are sufficient to obtain stable and meaningful topology estimates. This makes our method computationally practical even for large models.

A Analysis of Scalability and Stability

The computational cost of our measure primarily stems from computing the persistence diagram via Vietoris–Rips filtration, whose complexity scales poorly with the number of points N. Using the full set of embeddings is often infeasible for large datasets.

To make our approach practical, we rely on random subsampling of embeddings from the *source* dataset when computing PR. Here, we analyze the impact of this strategy on both computational cost and the stability of the resulting scores. We also note that obtaining a representative sample of the embedding space typically requires accounting for the number of classes, which can be large for common pretraining datasets such as ImageNet or Places (1,000+ classes).

We evaluate the evolution of $\mathcal{H}_{\text{sep}}$, $\mathcal{H}_{\text{tight}}$, and $\mathcal{H}_{\text{ratio}}$ under different random sample sizes. We use stratified random sampling on the ImageNet embeddings of each model and compute persistence diagrams on subsamples of increasing size (from 2,048 to 16,384 in steps of 2,048). For each model and sampling size, we repeat the procedure 5 times and average $\mathcal{H}_{\text{sep}}$, $\mathcal{H}_{\text{tight}}$, $\mathcal{H}_{\text{ratio}}$, and computation time.

As shown in Figure 5, computation times scale as expected, while the topological metrics themselves follow very similar trajectories across models and sampling sizes. Moreover, the relative ordering between models is largely preserved, with only minor changes between closely related models. Since the relative ranking of models at low sampling sizes is nearly identical to those at high sampling sizes, the overall conclusions of our benchmark would remain unchanged even under aggressive subsampling. This is confirmed in Figure 6, where Kendall's τ remains high across sampling sizes, suggesting that PR can be computed with high confidence using a small, computationally feasible subset of a large-scale source dataset.

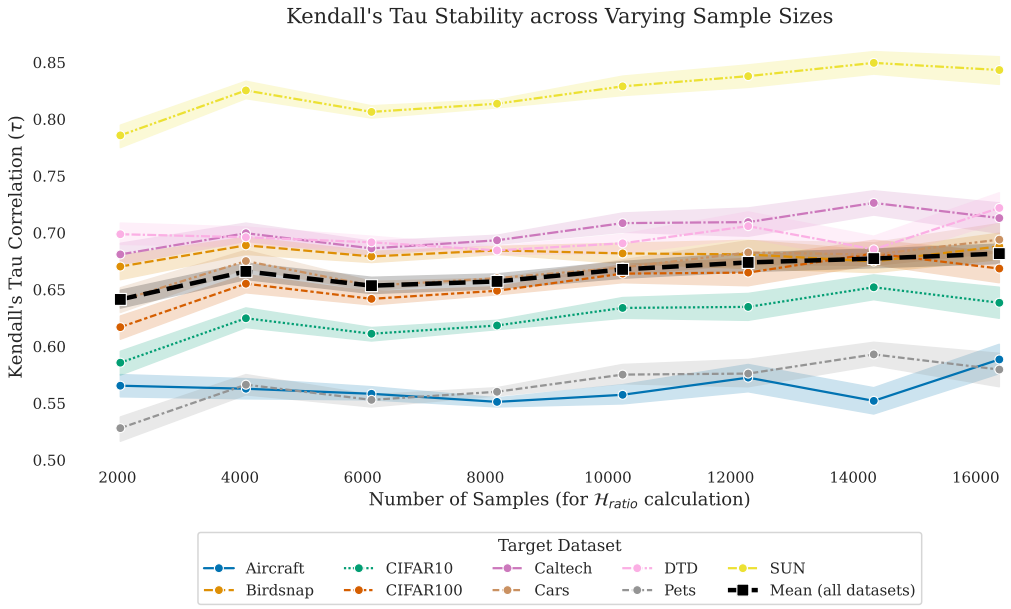

Figure 6: **Stability of transferability rankings across sampling sizes.** Using different subsample sizes of the source embeddings, we compute PR for each ImageNet-pretrained model, derive a ranking, and measure Kendall's τ relative to the ranking induced by fine-tuned accuracy. Even very small subsamples yield scores orderings that are nearly identical to those obtained with larger samples, demonstrating the robustness of the proposed topology-based measure. This stability further supports the practicality of computing persistent homology on modest-sized subsets of embeddings.

Table 3: Effect of replacing the class-count split $k-1$ with the automatic persistent-entropy split. Results are reported in the empirically relevant domain for each model family.

Model zoo / domain	Score	Kendall τ	Weighted Kendall τ_w
CNN / source	PR_k	0.658	0.722
CNN / source	$\mathrm{PR}_{\mathrm{auto}}$	0.647	0.665
ViT / target	PR_k	0.579	0.688
ViT / target	$\mathrm{PR}_{\mathrm{auto}}$	0.566	0.689

B Removing the Class-Count Requirement

The definition of PR uses the task class count k to split the finite H_0 persistence intervals. We investigate whether this dependence on k can be avoided.

We use an automatic persistence-based split and introduce $\mathrm{PR}_{\mathrm{auto}}$. Instead of fixing the split at $k-1$, we apply the persistent-entropy procedure of Atienza et al. (2016) to the ordered persistence lengths. Their method separates persistent intervals into topological features and topological noise by measuring the contribution of each interval to the entropy of the barcode.

We use the threshold filter between features and noise obtained from this method and use it to replace k/ As shown in Table 3, the automatically selected split retains a strong ranking signal without requiring the class count. For CNNs, ordinary Kendall correlation changes only slightly. The larger change in τ_w indicates that the class-count split remains useful for recovering the highest-ranked CNN models, although the automatic variant remains positively correlated with downstream performance across the evaluated datasets.

For ViTs, removing k has essentially no effect. On target representations. Hence, the target-domain connectivity signal observed for ViTs does not rely on knowing the semantic class count. Together, these results show that the $k-1$ split provides useful task structure but is not necessary for obtaining a reliable transferability ranking from the persistence spectrum. Moreover, it hints the possibility to extend the analysis of persistence features to self supervised settings, althought is expected that another formulation would be needed and higher homology groups, since the connetivity pattern studied in this work is related to the neural collapse expected to be induced in supervised training settings.

C A Label-Aware Extension of PR

The main formulation of PR deliberately avoids semantic labels. In supervised pre-training settings, however, source class annotations may already be available as part of the pre-training data. We therefore consider whether this additional information can be used to enrich the geometric signal captured by PR without requiring any target annotations.

We construct a simple label-aware extension, denoted $\mathrm{PR}_{\mathrm{LA}}$. In addition to PR, we compute the Silhouette coefficient on the source representation using the ground-truth source class assignments. We denote this supervised source statistic by Sil_s. Rather than combining their raw values, which have unrelated scales, we combine their model rankings with an equal-weight rank average:

$$\mathrm{PR}_{\mathrm{LA}} = \frac{1}{2}\left[\mathrm{rank}(\mathrm{PR}_{d^\star}) + \mathrm{rank}(\mathrm{Sil}_s)\right], \tag{4}$$

where $d^\star$ denotes the representation domain found to be informative in the main experiments: $d^\star = s$ for CNNs and $d^\star = t$ for ViTs.

This construction has a deliberately different information regime from the main method. $\mathrm{PR}_{\mathrm{LA}}$ uses semantic *source* labels through Sil_s, and should therefore be viewed as a supervised extension rather than as a label-free transferability estimator. Importantly, it still requires no target labels.

Table 4 shows that source semantic structure can be incorporated into PR when such supervision is available. The behaviour of the extension also follows the asymmetry between model families observed in the main experiments. For CNNs, both terms are computed from the source representation

Table 4: Label-aware enrichment of PR with supervised source class structure. $\mathrm{PR}_{\mathrm{LA}}$ combines source-label Silhouette with PR from the representation domain used by the corresponding model family. Source semantic labels are required by this extension; target labels are never used.

Model zoo	Label-aware score	Kendall τ	Weighted Kendall τ_w
CNN	$\mathrm{PR}_{\mathrm{LA}} = \mathrm{RF}(\mathrm{PR}_s, \mathrm{Sil}_s)$	0.734	0.765
ViT	$\mathrm{PR}_{\mathrm{LA}} = \mathrm{RF}(\mathrm{PR}_t, \mathrm{Sil}_s)$	0.690	0.757

with modest improvement. The relatively limited additional benefit is consistent with source PR already capturing much of the representation structure associated with transferability in this model family.

For ViTs, the two terms instead provide information from different representations: Sil_s summarizes supervised semantic organization of the source representation, while PR_t summarizes persistent connectivity of the unlabeled target representation. Their combination reaches $\tau = 0.690$ and $\tau_w = 0.757$, indicating that source class structure and target topological organization provide complementary information in this model family.

D Additional Correlation Analyses

To provide a more comprehensive evaluation of the transferability metrics presented in the main text, this appendix expands our analysis using widely adopted correlation measures. We evaluate these metrics for the ViTs case and the CNNs case.

D.1 Definitions of Correlation Measures

To aid interpretation, we define the statistical properties of the measures used in this work:

Kendall's Rank Correlation Coefficient (τ): The primary metric used in transferability literature. It measures the ordinal association between two measured quantities. It is defined as the difference between the number of concordant and discordant pairs divided by the total number of pairs. A value of 1 implies identical ranking, 0 implies no correlation, and -1 implies a perfectly inverse ranking.

Pairwise Accuracy (P_c): Also known as Concordant Probability, this intuitive metric represents the probability that a transferability estimator will correctly rank a randomly selected pair of models. A value of 0.5 indicates random chance, while 1.0 indicates perfect ordering. It is linearly related to Kendall's τ via the formula $P_c = (\tau + 1)/2$.

Weighted Kendall's τ_w: Standard correlation metrics treat all ranking errors equally. In transfer learning, users may be primarily interested in identifying the top-performing models. Weighted Kendall's τ_w utilizes a weighting scheme to assign higher importance to the top of the list, severely penalizing errors among the best models while being more forgiving at the bottom.

D.2 Analysis of Results

The results for ViTs are presented in Tables 7 and 8, while the results for CNNs are in Tables 5 and 6.

As shown in Tables 6 and 8, PR computed on target embeddings for transformers, and on source embeddings for CNNs outperform the compared baselines, including LogME, SFDA, and NCTI, Under weighted Kendall's τ_w, the highest mean correlation is obtained by source-domain PR for CNNs and target-domain PR for ViTs. This domain asymmetry is consistent with the ordinary Kendall and pairwise-accuracy results, although the dataset-bootstrap intervals indicate uncertainty in the relative ordering of methods across downstream tasks.

For PR, correlations are first averaged over ten uniform subsampling seeds within each downstream dataset. We report the mean across downstream datasets together with a 95% percentile bootstrap confidence interval obtained by resampling downstream datasets. The remaining transferability es-

Table 5: Pairwise Accuracy (ground truth from You et al. (2021)). PR values are averaged over ten uniform subsampling seeds. Mean uncertainty is a 95% percentile bootstrap confidence interval over downstream datasets. Highest value is in **bold** and second-highest is underlined. Higher is better.

Dataset	LEEP	NCE	LogME	SFDA	NCTI	$\mathbf{PR_t}$	$\mathbf{PR_s}$
Aircraft	0.644	0.756	0.680	0.822	0.644	**0.824**	0.778
Birdsnap	0.667	0.822	0.733	0.467	0.822	**0.842**	0.841
Caltech101	0.711	0.867	0.792	0.822	**0.911**	0.878	0.844
StanfordCars	0.712	0.783	0.759	0.712	**0.854**	0.646	0.833
CIFAR10	0.739	0.580	0.852	**0.875**	0.852	0.689	0.807
CIFAR100	0.711	0.556	0.815	**0.867**	**0.867**	0.473	0.823
DTD	0.511	0.356	0.747	0.745	0.556	0.222	**0.850**
Pets	0.733	**0.867**	0.756	0.711	0.822	0.658	0.780
SUN397	0.725	0.882	0.860	0.680	**0.927**	0.691	0.905
Mean	0.684	0.719	0.777	0.745	0.806	0.658	**0.829**
95% CI	[0.635,0.720]	[0.600,0.821]	[0.741,0.813]	[0.660,0.814]	[0.725,0.874]	[0.522,0.772]	[0.806,0.854]

Table 6: Weighted Kendall's τ_w correlation (ground truth from You et al. (2021)). PR values are averaged over ten uniform subsampling seeds. Mean uncertainty is a 95% percentile bootstrap confidence interval over downstream datasets. Highest value is in **bold** and second-highest is underlined. Higher is better.

Dataset	LEEP	NCE	LogME	SFDA	NCTI	$\mathbf{PR_t}$	$\mathbf{PR_s}$
Aircraft	0.110	0.395	0.530	0.664	0.437	**0.735**	0.676
Birdsnap	0.269	0.741	0.672	-0.169	0.759	0.728	**0.766**
Caltech101	0.270	0.648	0.691	0.603	**0.871**	0.787	0.787
StanfordCars	0.406	0.352	0.654	0.209	**0.842**	0.498	0.835
CIFAR10	0.679	0.471	0.785	**0.832**	0.715	0.476	0.650
CIFAR100	0.615	0.435	0.783	**0.818**	0.812	-0.095	0.666
DTD	-0.085	-0.383	0.468	0.564	-0.057	-0.537	**0.819**
Pets	0.648	**0.839**	0.580	0.506	0.658	0.373	0.609
SUN397	0.581	0.770	**0.863**	0.297	0.828	0.418	0.800
Mean	0.388	0.474	0.670	0.480	0.652	0.376	**0.734**
95% CI	[0.218,0.543]	[0.226,0.667]	[0.590,0.749]	[0.270,0.663]	[0.448,0.801]	[0.087,0.613]	[0.682,0.785]

timators are deterministic given the extracted representations and are bootstrapped across datasets using the same procedure.

In addition Tables 7 and 5 show that PR strong pairwise accuracy in the corresponding beneficial domain for each model family. This implies that for any random pair of pre-trained models, PR correctly predicts the superior model, providing a highly actionable signal for practitioners compared to the other measures we evaluated. Given the results obtained using Weighted Kendall's τ_w, we know that the results are not biased toward detecting low-performing models, but rather toward correctly identifying the high-performing ones (e.g., the best or second-best model). In this regard, in the specific case of PR computed on source embedding, for the Transformers case, we observe very strong results in concordance probability, but this is actually due to high performance on models that are not regarded as the best. As a result, the source PR can be considered an effective heuristic for selecting competent models, but it might not pick the best models when comparing transformers.

Table 7: Pairwise Accuracy (ground truth from Shao et al. (2022)) on transformer models. PR values are averaged over ten uniform subsampling seeds. Mean uncertainty is a 95% percentile bootstrap confidence interval over downstream datasets. Highest value is in **bold** and second-highest is underlined. Higher is better.

Dataset	LogME	SFDA	NCTI	$\mathbf{PR_t}$	$\mathbf{PR_s}$
Aircraft	0.711	0.578	0.644	0.713	**0.780**
Caltech101	0.733	0.711	0.733	**0.784**	0.731
StanfordCars	0.778	0.689	0.756	0.833	**0.847**
CIFAR10	**0.761**	0.739	0.739	0.700	0.605
CIFAR100	**0.844**	0.778	0.756	0.813	0.691
DTD	0.778	0.800	0.644	0.800	**0.824**
Flowers102	0.578	0.776	0.578	0.844	**0.869**
Food101	0.733	0.756	0.667	**0.769**	0.758
Pets	0.889	**0.911**	**0.911**	0.784	0.802
SUN397	0.800	0.711	0.644	**0.809**	0.780
VOC2007	0.882	**0.949**	0.860	0.833	0.794
Mean	0.772	0.763	0.721	**0.789**	0.771
95% CI	[0.720,0.818]	[0.707,0.822]	[0.667,0.779]	[0.761,0.814]	[0.727,0.810]

Table 8: Weighted Kendall's τ_w correlation (ground truth from Shao et al. (2022)) on transformer models. PR values are averaged over ten uniform subsampling seeds. Mean uncertainty is a 95% percentile bootstrap confidence interval over downstream datasets. Highest value is in **bold** and second-highest is underlined. Higher is better.

Dataset	LogME	SFDA	NCTI	$\mathbf{PR_t}$	$\mathbf{PR_s}$
Aircraft	0.491	0.250	0.402	0.661	**0.758**
Caltech101	0.455	0.429	0.470	**0.589**	0.399
StanfordCars	0.585	0.427	0.558	**0.808**	0.769
CIFAR10	**0.640**	0.587	0.621	0.599	0.267
CIFAR100	**0.771**	0.665	0.666	0.769	0.420
DTD	0.596	0.711	0.438	**0.785**	0.586
Flowers102	0.331	0.555	0.331	**0.713**	0.663
Food101	0.578	0.578	0.465	**0.583**	0.511
Pets	0.794	**0.870**	0.814	0.532	0.507
SUN397	0.624	0.516	0.426	**0.731**	0.472
VOC2007	0.762	**0.943**	0.738	0.796	0.498
Mean	0.602	0.594	0.539	**0.688**	0.532
95% CI	[0.521,0.680]	[0.483,0.707]	[0.456,0.627]	[0.632,0.742]	[0.447,0.619]

E GROUND TRUTH RESULTS

This appendix section presents the accuracy reported in previous fine-tuning studies; these values served as the ground truth for our experiments. We utilize two distinct sets of fine-tuning results. For our transformer results, we used the one reported in Shao et al. (2022), and for the CNN, we used You et al. (2021).

Table 9: Ground truth accuracy obtained from You et al. (2021).

Task	ResNet-50	ResNet-101	ResNet-152	DenseNet-121	DenseNet-169	DenseNet-201	Inception v1	Inception v3	MobileNet v2	NASNet-A
Aircraft	86.6	85.6	85.3	85.4	84.5	84.6	82.7	88.8	82.8	72.8
Birdsnap	74.7	73.8	74.3	73.2	71.4	72.6	73.0	77.2	69.3	68.3
Caltech	91.8	93.1	93.2	91.9	92.5	93.4	91.7	94.3	89.1	91.5
Cars	91.7	91.7	92.0	91.5	91.5	91.0	91.0	92.3	91.0	88.5
CIFAR10	96.8	97.7	97.9	97.2	97.4	97.4	96.2	97.5	95.7	96.8
CIFAR100	84.5	87.0	87.6	84.8	85.0	86.0	83.2	86.6	80.8	83.9
DTD	75.2	76.2	75.4	74.9	74.8	74.5	73.6	77.2	72.9	72.8
Pets	92.5	94.0	94.5	92.9	93.1	92.8	91.9	93.5	90.5	89.4
SUN	64.7	64.8	66.0	62.3	63.0	64.7	62.0	65.7	60.5	60.7

Table 10: Ground truth accuracy for Transformer models obtained from Shao et al. (2022)

Task	ViT-T	ViT-S	ViT-B	PVTv2-B2	PVT-T	PVT-S	PVT-M	Swin-T	MoCov3-S	DINO-S
Aircraft	71.26	73.12	78.39	84.14	69.76	75.20	76.70	81.90	76.04	72.18
Caltech	89.39	92.70	93.47	93.13	90.04	93.02	93.75	91.90	89.84	86.76
Cars	82.09	86.72	89.26	90.60	84.10	87.61	87.66	88.93	82.18	79.81
CF-10	96.52	97.69	98.56	97.96	94.87	97.34	97.93	97.34	97.92	97.96
CF-100	81.58	86.62	89.96	88.24	75.26	86.20	87.36	85.97	85.84	85.66
DTD	71.86	75.08	77.66	77.16	72.92	75.77	77.10	77.04	71.88	75.96
Flowers	95.50	96.79	97.98	97.89	95.80	97.32	97.36	97.40	93.89	95.96
Food	81.96	86.26	88.96	88.67	83.78	86.98	85.56	86.67	82.84	85.69
Pets	91.44	94.02	94.61	93.86	91.48	94.13	94.48	94.50	90.44	92.59
SUN	58.40	64.76	68.62	66.44	61.86	65.78	67.22	65.51	60.60	64.14
VOC	83.10	86.62	87.88	86.44	84.60	86.62	87.36	87.54	81.84	84.80