FROM DIVERGENCE TO NORMALIZED SIMILARITY: A SYMMETRIC AND SCALABLE TOPOLOGICAL TOOLKIT FOR REPRESENTATION ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation Topology Divergence (RTD) offers a powerful lens for analyzing topological differences in neural network representations. However, its asymmetry and lack of a normalized scale limit its interpretability and direct comparability across different models. Our work addresses these limitations on two fronts. First, we complete the theoretical framework of RTD by introducing Symmetric Representation Topology Divergence (SRTD) and its lightweight variant, SRTD-lite. We prove their mathematical properties, demonstrating that they provide a more efficient, comprehensive, and interpretable divergence measure which matches the top performance of existing RTD-based methods in optimization tasks. Second, to overcome the inherent scaling issues of divergence measures, we propose Normalized Topological Similarity (NTS), a novel, normalized similarity score robust to representation scale and size. NTS captures the hierarchical clustering structure of representations by comparing their topological merge orders. We demonstrate that NTS can reliably identify inter-layer similarities and, when analyzing representations of Large Language Models (LLMs), provides a more discriminative score than Centered Kernel Alignment (CKA), offering a clearer view of inter-model relationships.

1 INTRODUCTION

Understanding the internal representations of neural networks is a central challenge in deep learning, crucial for interpreting their behavior and improving their design. Analyzing the similarity structure of these representations has emerged as a key field for deciphering model behavior (Kriegeskorte et al., 2008). Early research primarily relied on Canonical Correlation Analysis (CCA) and its variants, such as SVCCA (Raghu et al., 2017) and PWCCA (Morcos et al., 2018). However, these methods were often criticized for being too loose, as they remain invariant under any invertible linear transformation. To address this, Centered Kernel Alignment (CKA) (Kornblith et al., 2019) was proposed and has since become the de facto standard (Khrulkov & Oseledets, 2018; Raghu et al., 2019; Wu et al., 2020; Zhang et al., 2024). By quantifying similarity through centered Gram matrices, CKA provides a normalized score that facilitates comparison across diverse experimental settings and is robust to fundamental geometric transformations.

While geometric analysis dominates the field, Topological Data Analysis (TDA) offers a complementary perspective by probing the intrinsic shape of data. Using tools like persistent homology (Barannikov, 1994; Carlsson et al., 2004), this approach examines how the fundamental topological structure of the data—from simple clusters to complex loops and voids—is formed and evolves across a continuous range of scales. This focus on properties that are invariant to non-linear deformations (such as stretching and bending) allows TDA to capture a different, often complementary, notion of structural similarity that is overlooked by geometry-centric measures.

Existing topological methods, however, face distinct limitations regarding their applicability. Methods such as Geometry Score (Khrulkov & Oseledets, 2018) and IMD (Tsitsulin et al., 2019) are highly general and do not require a one-to-one correspondence between representations. While flexible, they fail to leverage the valuable pairing information inherent in comparing neural network layers, often resulting in lower discriminative power. Conversely, approaches that do analyze

distributional topology often strictly require the point clouds to reside in the same ambient space (Kynkäänniemi et al., 2019; Barannikov et al., 2021b), severely limiting their scope.

A significant breakthrough in bridging this gap is Representation Topology Divergence (RTD) (Barannikov et al., 2021a) and its scalable variant, RTD-lite (Tulchinskii et al., 2025). These methods successfully utilize the one-to-one correspondence between data points without requiring them to share the same ambient space, making them powerful tools for representation analysis and optimization (Trofimov et al., 2023).

Despite these advancements, the RTD framework suffers from two critical limitations that hinder its broader adoption. First, its theoretical underpinnings remain incomplete. The standard symmetric measure is a brute-force average of two directional values, $RTD(w, \tilde{w})$ and $RTD(\tilde{w}, w)$, that can differ dramatically (Table 2f) without a clear theoretical explanation.Another theoretical ambiguity comes from it's dual variant, Max-RTD, mentioned by Trofimov et al. (2023) to enrich gradient information, but whose theoretical role and relationship to the original RTD were never fully investigated.

Second, and more critically, unlike CKA, topological divergence methods are not normalized:the output of RTD and RTD-lite can be any positive number, heavily dependent on the number of sample points and the intrinsic scale of distances. This lack of a normalized scale makes cross-scenario comparison difficult and interpretability elusive. For instance, in layer-wise analysis, unnormalized divergence measures often fail to reveal the graded similarity patterns between layers (Figure 4a)—a task that CKA consistently accomplishes due to its normalization.

To address these issues, we propose a comprehensive topological toolkit with the following contributions:

- We complete the theoretical framework of RTD by introducing **Symmetric Representation Topology Divergence (SRTD)** and its lightweight variant, **SRTD-lite**. We reveal the mathematical relationships between RTD, Max-RTD, and SRTD, proving that SRTD provides a more comprehensive and computationally efficient divergence measure that matches the top performance of this class of methods in optimization tasks.

- We introduce **Normalized Topological Similarity (NTS)**, a novel, scale-invariant, and normalized similarity measure. Unlike divergence-based methods, NTS captures hierarchical clustering features and can robustly reveal graded inter-layer similarity patterns that are often missed by RTD, combining the interpretability of CKA with the structural sensitivity of TDA.

2 Preliminaries: Persistent Homology and Representation Topology Divergence

We consider two point clouds, P and P', of the same size with a one-to-one correspondence. Their respective pairwise distance matrices are denoted by w and $\tilde{w}$. We define $\min(w, \tilde{w})$ and $\max(w, \tilde{w})$ as the element-wise minimum and maximum of the two matrices, respectively.

To understand the topological structure of these point clouds, we employ persistent homology. The process can be intuitively understood as follows: for a given point cloud P with distance matrix w, we construct a sequence of simplicial complexes, known as the Vietoris-Rips filtration (Hausmann, 1995), indexed by a proximity parameter α. As α increases from zero, edges are added between points with distance less than or equal to α. When a set of n points are all mutually connected, the $(n-1)$-simplex they span is filled in (e.g., three points form a filled triangle). This growing complex is denoted as $R_\alpha(\mathcal{G}^w)$.

During this filtration process, topological features—such as connected components (H_0), cycles (H_1), and voids (H_2)—appear and disappear. We track the lifespan of each feature by recording its birth and death values as an interval $[b, d]$ (Barannikov, 1994). The collection of these intervals is known as **barcodes** (Carlsson et al., 2004), which serves as a topological signature of the point cloud. The computation of persistent homology operates directly on the distance matrix.

RTD A set of barcodes characterizes one point cloud. To compare two, Representation Topology Divergence (RTD) (Barannikov et al., 2021a) introduced an auxiliary matrix M_{min} (Matrix 1b) constructed from w, $\tilde{w}$, and $\min(w, \tilde{w})$. The resulting barcode captures the differences in the evolution of topological features between an individual point cloud and the composite structure formed by their union, which is derived from the $\min(w, \tilde{w})$ matrix. The length of a barcode interval in this context quantifies the discrepancy between when a feature forms in w (or $\tilde{w}$) versus when it forms in $\min(w, \tilde{w})$.

We define $RTD(w, \tilde{w})$ as the sum of the lengths of all barcodes computed from M_{min} (Matrix 1b). By swapping the roles of w and $\tilde{w}$, we can similarly compute $RTD(\tilde{w}, w)$. To ensure symmetry, the final divergence is typically defined as their average:$RTD(P, P') = \frac{RTD(w,\tilde{w})+RTD(\tilde{w},w)}{2}$ Subsequently, Trofimov et al. (2023) noted that a dual variant, which we term Max-RTD, can be defined by using an auxiliary matrix M_{max} (Matrix 1c) based on w, $\tilde{w}$, and $\max(w, \tilde{w})$. However, the properties of this variant were not deeply investigated in their work. The symmetric versions of Max-RTD are defined analogously by averaging the two directional computations.

RTD-lite To address the computational cost of higher-dimensional homology, RTD-lite (Tulchinskii et al., 2025) was introduced as a lightweight variant focusing solely on 0-dimensional features—the merging of connected components. The key insight is that its divergence score can be calculated efficiently, as it is exactly the difference between the weights of the Minimum Spanning Trees (MSTs) of the respective distance matrices. For instance, the directional divergence $RTD\_lite(w, \tilde{w})$ is given by $MST(w) - MST(\min(w, \tilde{w}))$, and the final measure is symmetrized by averaging the two directional computations. This connection to MSTs provides a computationally feasible tool for large-scale representation analysis.

Notation for Vietoris-Rips Complexes To streamline the following sections, we establish notation for the key Vietoris-Rips complexes used in our analysis. Recall that these are constructed based on a proximity parameter, α, which acts as a distance threshold for connecting points. For any given threshold α, we denote the complexes generated from the distance matrices w and $\tilde{w}$ as $R_\alpha(\mathcal{G}^w)$ and $R_\alpha(\mathcal{G}^{\tilde{w}})$, respectively. The complexes derived from the element-wise minimum and maximum matrices have a crucial relationship to these: at the same scale α, $R_\alpha(\mathcal{G}^{\min(w,\tilde{w})})$ is the union of the individual complexes ($R_\alpha(\mathcal{G}^w) \cup R_\alpha(\mathcal{G}^{\tilde{w}})$), while $R_\alpha(\mathcal{G}^{\max(w,\tilde{w})})$ is their intersection ($R_\alpha(\mathcal{G}^w) \cap R_\alpha(\mathcal{G}^{\tilde{w}})$).

$$\begin{pmatrix} \max(w,\tilde{w}) & (\max(w,\tilde{w})^+)^T & 0 \\ \max(w,\tilde{w})^+ & \min(w,\tilde{w}) & \infty \\ 0 & \infty & 0 \end{pmatrix} \qquad \begin{pmatrix} w & (w^+)^T & 0 \\ w^+ & \min(w,\tilde{w}) & \infty \\ 0 & \infty & 0 \end{pmatrix} \qquad \begin{pmatrix} \max(w,\tilde{w}) & (\max(w,\tilde{w})^+)^T & 0 \\ \max(w,\tilde{w})^+ & w & \infty \\ 0 & \infty & 0 \end{pmatrix}$$

(a) M_{sym} (b) M_{min} (c) M_{max}

Figure 1: The three key auxiliary matrices. For any matrix M, M^+ is obtained by replacing its upper triangular part with infinity.

3 Symmetric Representation Topology Divergence (SRTD)

In practice, we observe a complementary phenomenon between RTD and Max-RTD (shown in Table 2f). When $RTD(w, \tilde{w}) > RTD(\tilde{w}, w)$, we consistently find that $Max\text{-}RTD(w, \tilde{w}) < Max\text{-}RTD(\tilde{w}, w)$. This suggests that the topological structural differences between $R_\alpha(\mathcal{G}^w) \cup R_\alpha(\mathcal{G}^{\tilde{w}})$ and $R_\alpha(\mathcal{G}^w) \cap R_\alpha(\mathcal{G}^{\tilde{w}})$ seem to be the core reason for the asymmetry in RTD. Therefore, we propose to directly measure this difference as the Symmetric Representation Topology Divergence (SRTD) of P and P'.

Definition 3.1 (SRTD)**.** For two point clouds P and P' with a one-to-one correspondence, the distance matrix of their auxiliary graph $\hat{\mathcal{G}}'_{sym}$ is M_{sym} (Matrix 1a). The sum of the lengths of its persistent homology barcodes is defined as $SRTD(P, P')$ (see Algorithm 3). Its chain complex is homotopy equivalent to the mapping cone of the inclusion map $f' : C_*(R_\alpha(\mathcal{G}^w) \cap R_\alpha(\mathcal{G}^{\tilde{w}})) \to C_*(R_\alpha(\mathcal{G}^w) \cup R_\alpha(\mathcal{G}^{\tilde{w}}))$.

The logic behind RTD-lite—simplifying topological divergence to a calculation on Minimum Spanning Trees (MSTs)—can be extended across the entire RTD framework. This allows us to formally define **Max-RTD-lite**, the natural dual to RTD-lite, which compares an individual MST to the MST of the intersection structure (derived from $\max(w, \tilde{w})$). With this complete lightweight family in place, we introduce our proposed symmetric version, **SRTD-lite**, as the most direct and fundamental measure. Since the full SRTD compares the topologies of the composite union $R_\alpha(\mathcal{G}^{\min(w,\tilde{w})})$ and intersection $R_\alpha(\mathcal{G}^{\max(w,\tilde{w})})$ structures, SRTD-lite quantifies the divergence between them by simply comparing the weights of their respective MSTs.

Definition 3.2 (SRTD-lite)**.** By comparing the minimum spanning trees of $\min(w, \tilde{w})$ and $\max(w, \tilde{w})$ through Algorithm 4, we can obtain a series of barcodes. We define the sum of the lengths of these barcodes as $SRTD\text{-}lite(w, \tilde{w})$.

3.1 Mathematical Properties

SRTD, RTD, and Max-RTD satisfy some elegant mathematical properties. The mapping cones corresponding to their auxiliary graphs fit into the following long exact sequence:

$$\cdots \to H_n(R_\alpha(\mathcal{G}^w), R_\alpha(\mathcal{G}^{\max(w,\tilde{w})})) \xrightarrow{\gamma_n} H_n(R_\alpha(\mathcal{G}^{\min(w,\tilde{w})}), R_\alpha(\mathcal{G}^{\max(w,\tilde{w})}))$$
$$\xrightarrow{\beta_n} H_n(R_\alpha(\mathcal{G}^{\min(w,\tilde{w})}), R_\alpha(\mathcal{G}^w)) \xrightarrow{\delta_n} H_{n-1}(R_\alpha(\mathcal{G}^w), R_\alpha(\mathcal{G}^{\max(w,\tilde{w})})) \xrightarrow{\gamma_{n-1}} \cdots$$

Theorem 3.3. *For any dimension i, point clouds P, P' and distance matrices $w, \tilde{w}$, the three divergences satisfy the following relationship:*

$$Max\text{-}RTD_i(w, \tilde{w}) + RTD_i(w, \tilde{w}) - SRTD_i(w, \tilde{w}) = \int_0^\infty (\dim(\ker(\gamma_i)) + \dim(\ker(\gamma_{i-1})))d\alpha$$

By swapping the positions of w and $\tilde{w}$ in Theorem 3.3, we obtain a similar equality. We denote $RTD_i(w, \tilde{w}) + Max\text{-}RTD_i(w, \tilde{w})$ as $minmax(w, \tilde{w})$,and $RTD_i(\tilde{w}, w) + Max\text{-}RTD_i(\tilde{w}, w)$as $minmax(\tilde{w}, w)$. Both are strictly greater than SRTD, but in our experiments, we find this gap to be very small, as shown in the Table 2e.

The introduction of SRTD provides a more mathematically elegant framework for understanding the RTD family. Within this framework, the asymmetric measures $minmax(w, \tilde{w})$ and $minmax(\tilde{w}, w)$ can be decomposed into a large, shared symmetric component, $SRTD(w, \tilde{w})$, and smaller, 'private' components. These private components correspond to topological features unique to the individual filtrations of $\mathcal{G}^w$ or $\mathcal{G}^{\tilde{w}}$ relative to the bounding filtrations of $\mathcal{G}^{\min(w,\tilde{w})}$ and $\mathcal{G}^{\max(w,\tilde{w})}$. This decomposition reveals that the asymmetry in the original RTD arises from these small, private feature sets, making the source of the divergence interpretable. The relationship becomes even more direct and elegant in the lite version:

Corollary 3.4. $Max\text{-}RTD\text{-}lite(w, \tilde{w}) + RTD\text{-}lite(w, \tilde{w}) = SRTD\text{-}lite(w, \tilde{w})$

Corollary 3.5. $Max\text{-}RTD\text{-}lite(P, P') \geq SRTD\text{-}lite(P, P') \geq RTD\text{-}lite(P, P')$

Together, Theorem 3.3 and Corollary 3.4, 3.5 provide a clear theoretical basis for a consistent pattern observed in our experiments: when plotting the divergence curves for either the full or lite families, the Max-RTD curve is always highest, the RTD curve is lowest, and the SRTD curve lies in between (as shown in Figure 2b). For the lite versions, Corollary 3.5 proves this hierarchical ordering is strict, which explains why the SRTD-lite curve appears perfectly centered between the other two. While the relationship for the full RTD family is more complex, this structure holds empirically, positioning SRTD as a balanced, median measure of topological divergence.

4 Normalized Topological Similarity (NTS)

4.1 Motivation: The Limitations of Divergence-Based Analysis

While SRTD theoretically completes the topological divergence framework, the reliance on summing barcode lengths creates two practical limitations for general similarity analysis. First, as previously discussed, the unnormalized scores are inherently scale-dependent and difficult to interpret

across different contexts. Second, and more critically, the total divergence can be dominated by a few "ultra-long" barcodes (Figure 19a) corresponding to large-scale structural differences. This sensitivity to a handful of major dissimilarities can mask a high degree of similarity in finer structural details, making the measure brittle.

These limitations underscore the need for a fundamentally different approach: a normalized, scale-invariant similarity measure designed to robustly capture hierarchical clustering structures.

4.2 Method: Capturing Merge-Order Similarity

Instead of comparing the *magnitudes* of topological features, we propose to compare their relative *order* of formation. The sequence of merge events in 0-dimensional persistent homology provides a scale-invariant signature of a point cloud's hierarchical clustering structure. To robustly compare such sequences, we employ Spearman's rank correlation coefficient (ρ),which is inherently normalized to $[-1, 1]$ and is robust to outliers and monotonic scaling (Spearman, 1961).

The merge sequence of connected components is perfectly captured by the Minimum Spanning Tree (MST), which forms the backbone of the 0-dimensional filtration. Our method, Normalized Topological Similarity (NTS), leverages this connection. The core idea is to first establish a common basis for comparison—the set of core pairs—by taking the union of edges from the MSTs of both point clouds. For every pair in this common set, we extract a corresponding numerical value from each point cloud's structure. This process creates two parallel vectors, and the NTS score is their Spearman's rank correlation.

We define two variants based on the values extracted:

- **NTS-M (Merge-time based):** This theoretically-grounded variant compares the ranks of the merge times. The merge time of a pair of points is the threshold at which they become connected in the filtration, formally defined by the maximum edge weight on the path between them in their MST.
- **NTS-E (Edge-distance based):** This practical variant directly compares the ranks of the original pairwise distances for the 'core pairs'. It is computationally simpler and often more sensitive in practice, as it retains more of the original metric information.

4.3 Formal Definition and Properties

The procedures for calculating NTS-M and NTS-E are formally defined in Algorithm 1 and 2.

Algorithm 1: NTS-M (Merge-time based)

Input: Pairwise distance matrices $w, \tilde{w}$
Output: NTS-M score

$E_w \leftarrow$ Edge set of MST(w)
$E_{\tilde{w}} \leftarrow$ Edge set of MST($\tilde{w}$)
$E_{core} \leftarrow E_w \cup E_{\tilde{w}}$
$V_{merge} \leftarrow (\text{MergeTime}(e, w))_{e \in E_{core}}$
$\tilde{V}_{merge} \leftarrow (\text{MergeTime}(e, \tilde{w}))_{e \in E_{core}}$
return Spearman's $\rho(V_{merge}, \tilde{V}_{merge})$

Algorithm 2: NTS-E (Edge-distance based)

Input: Pairwise distance matrices $w, \tilde{w}$
Output: NTS-E score

$E_w \leftarrow$ Edge set of MST(w)
$E_{\tilde{w}} \leftarrow$ Edge set of MST($\tilde{w}$)
$E_{core} \leftarrow E_w \cup E_{\tilde{w}}$
$V_{dist} \leftarrow (w_{ij})_{(i,j) \in E_{core}}$
$\tilde{V}_{dist} \leftarrow (\tilde{w}_{ij})_{(i,j) \in E_{core}}$
return Spearman's $\rho(V_{dist}, \tilde{V}_{dist})$

The NTS framework satisfies the following key properties, which highlight the stricter condition imposed by NTS-E.

Theorem 4.1. $NTS\text{-}M(P, P') = 1$ *if and only if the rank order of merge times for all core pairs is identical for both point clouds.*

Theorem 4.2. *If* $NTS\text{-}E(P, P') = 1$*, then the rank order of merge times for all core pairs is also identical (i.e.,* $NTS\text{-}M(P, P') = 1$*). The converse is not necessarily true.*

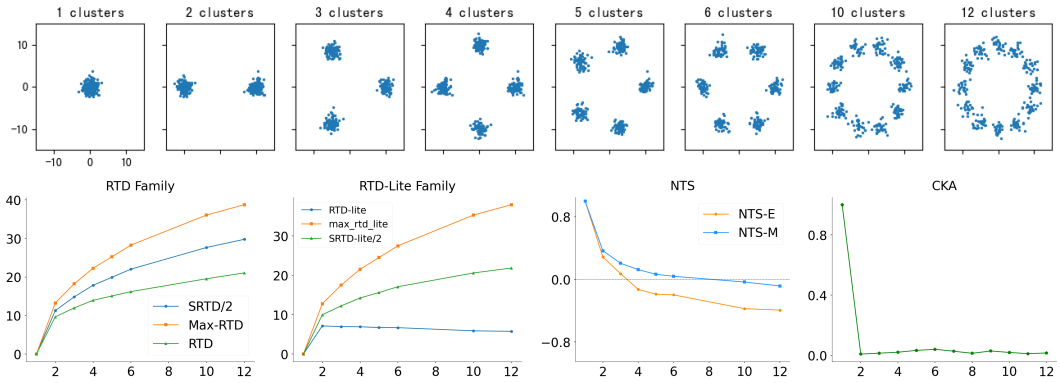
NTS-E provides a stricter condition by comparing underlying distance ranks—making it more sensitive in practice—while NTS-M compares the final merge-time order to capture a more fundamental notion of structural similarity.

5 EXPERIMENTS

5.1 ANALYSIS OF HIERARCHICAL CLUSTERING STRUCTURES

We begin our experimental validation on two controlled tasks designed to test each method's reliability and sensitivity in capturing hierarchical clustering structures.

Clusters Experiment. We test sensitivity to increasing structural dissimilarity by comparing a single cluster of 300 2D Gaussian points against variants where the points are partitioned into $k = 2, \ldots, 12$ clusters arranged on a circle. The results reveal a clear performance divide: our proposed NTS and SRTD families correctly capture the expected trend of increasing dissimilarity. In contrast, CKA is largely insensitive to these structural changes, while RTD-lite produces an anomalous, inverted trend, confirming that the $\max(w, \tilde{w})$ component is essential for a robust divergence measure.



(a) RTD Family (b) RTD_lite Family (c) NTS (d) CKA

(e) Theoretical Difference from SRTD

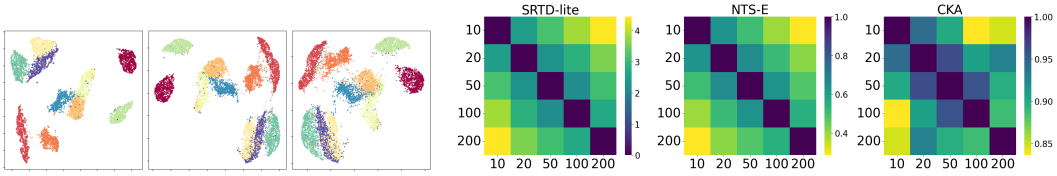
Clusters	$E_1(Percentage_1)$	$E_2(Percentage_2)$
2	0.357 (3.16%)	0.000 (0.00%)
3	0.493 (3.32%)	0.013 (0.09%)
4	0.441 (2.47%)	0.061 (0.34%)
5	0.451 (2.26%)	0.039 (0.20%)
6	0.347 (1.57%)	0.060 (0.27%)
10	0.263 (0.95%)	0.043 (0.15%)
12	0.226 (0.76%)	0.046 (0.15%)

(f) Asymmetry of RTD vs Max-RTD

Min-Asym	Max-Asym
13.0976	-12.3839
11.2554	-10.2954
10.8131	-10.0535
10.3320	-9.5084
9.4315	-8.8572
8.3074	-7.8674
7.6888	-7.3296

Figure 2: Analysis of the RTD framework on the synthetic Clusters dataset. (e) shows the small theoretical difference between SRTD and the symmetrized RTD/Max-RTD combination, where $E_1 = (\text{RTD}(w, \tilde{w}) + \text{Max-RTD}(w, \tilde{w}) - \text{SRTD})/2$ and E_2 is defined analogously by swapping w and $\tilde{w}$, $percentage_1 = (\text{RTD}(w, \tilde{w}) + \text{Max-RTD}(w, \tilde{w}) - \text{SRTD})/\text{SRTD}$. (f) illustrates the strong asymmetry and complementarity between RTD and Max-RTD, Min-Asym $= RTD(w, \tilde{w}) - RTD(\tilde{w}, w)$, Max-Asym $= Max\text{-}RTD(w, \tilde{w}) - Max\text{-}RTD(\tilde{w}, w)$

UMAP Embeddings Experiment. We test sensitivity to structural changes by generating a sequence of 2D UMAP embeddings (Damrich & Hamprecht, 2021) from the MNIST dataset (LeCun et al., 2002), varying the `n_neighbors` parameter to control the trade-off between local and global structure. Pairwise comparisons of these embeddings (Figure 3) demonstrate that our proposed methods, NTS and SRTD-lite, track these changes with a smooth, monotonic response. In contrast, the CKA baseline fails to capture this gradual evolution, highlighting the superior sensitivity of our topological measures.



(a) UMAP n_neighbors=(10, 50, 200)

(b) UMAP experiment heatmap

Figure 3: UMAP experiment

5.2 Efficiency as an Optimization Loss

We evaluate the practical utility of our divergence measures as loss terms for training an autoencoder, a task for which they are naturally suited. In this experiment, autoencoder is trained to reduce the dimensionality of the F-MNIST and COIL-20 dataset to 16 (Xiao et al., 2017; Nene et al., 1996). It is crucial to note this is an **intra-family comparison**, designed to demonstrate that our proposed SRTD offers the best trade-off between performance and efficiency within the RTD class of methods. The results confirm that SRTD and SRTD-lite achieves top-tier performance on quality metrics while being faster than its predecessors. (Full results are provided in Appendix G).

5.3 Analyzing Structural Consistency and Functional Hierarchy

To rigorously test our measures in a practical setting, we analyze the structural consistency of representations learned by an 8-layer `TinyCNN` (see AppendixE). Our experimental design, including the network architecture and training procedure on CIFAR-10 (Krizhevsky et al., 2009), is adapted from the original CKA study (Kornblith et al., 2019; Springenberg et al., 2014) . For the analysis, we use the representations of 5,000 images sampled from the test set. We trained ten instances of this network from scratch with different random seeds[1].

This setup allows us to validate a key distinction observed in related work (Tulchinskii et al., 2025), which found that while topological divergence measures like RTD and RTD-lite can identify corresponding layers, they, unlike CKA, fail to capture the robust graded similarity patterns between adjacent and nearby layers. The heatmaps in Figure 4, showing the average results over all 45 unique model pairs, confirm this finding and reveal three key insights:

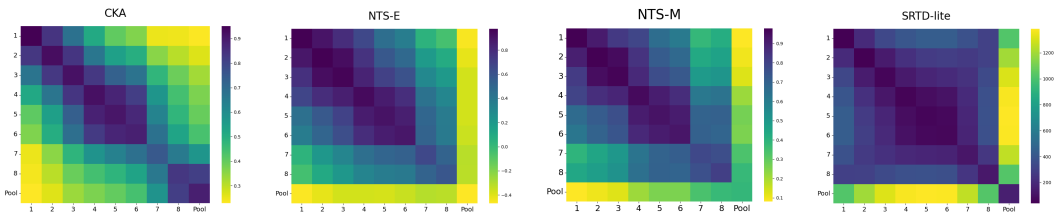
- **Layer Identification:** All methods are highly effective at identifying corresponding convolutional layers, achieving over 94% accuracy.
- **Graded Patterns:** NTS and CKA both reveal a clear, graded similarity pattern across convolutional layers, an interpretable landscape that RTD-lite and RTD families fail to produce.
- **Functional Shift Detection:** Crucially, only the topological measures (NTS and SRTD-lite) detect the sharp structural break at the final pooling layer. This identifies a fundamental functional shift from feature extraction to global aggregation that CKA misses.

These results demonstrate that NTS uniquely combines the strengths of both approaches: it provides an interpretable, graded similarity landscape akin to CKA, while also retaining the topological sensitivity needed to identify fundamental shifts in the network's functional hierarchy.

5.4 Analysis of Large Language Model Representations

We conclude our experimental validation by analyzing the complex representations of Large Language Models (LLMs). Our methodology is closely adapted from REEF (Zhang et al., 2024), a recent study that established a robust protocol for fingerprinting and comparing LLM representations. REEF identified that certain datasets are particularly effective at eliciting discriminative

[1]We select CKA as the primary baseline due to its widespread adoption as a robust, normalized similarity measure. Other methods such as SVCCA (Raghu et al., 2017) are omitted as they have been shown to be less effective for this type of layer analysis in prior studies (Kornblith et al., 2019; Barannikov et al., 2021a).



(a) CKA (98.89%) (b) NTS-E (97.22%) (c) NTS-M (94.72%) (d) SRTD-lite (98.33%)

Figure 4: Average layer-wise comparison over 45 pairs of trained `TinyCNN`s. NTS (b, c) provides the most comprehensive view, matching CKA's (a) graded pattern while also sharing the topological methods' (d) unique sensitivity to the functional shift at the final pooling layer, a distinction CKA misses.

features that highlight inter-model differences. Following their findings, we conduct our analysis on two such datasets: TruthfulQA (Lin et al., 2021) and ToxiGen (Hartvigsen et al., 2022). For each dataset, we adopt the REEF protocol of extracting the last-token representation from every Transformer layer across 1,000 randomly sampled QA pairs.

Identifying Intra-Model Hierarchical Patterns. Our first goal is to evaluate intra-model layer similarity. The resulting heatmaps visualize this, with both the x- and y-axes representing every Transformer layer of a given model, from first to last. An ideal measure should satisfy two criteria: (1) the layer-wise similarity map for a single model should be structurally informative, revealing distinct processing stages, and (2) this structural pattern should be consistent across models from the same family.

Our analysis, summarized in Figure 5, shows a stark contrast in reliability. NTS successfully identifies consistent, hierarchical fingerprints for all tested model families (Qwen, InternLM, Baichuan, and Llama). CKA, however, proves unreliable, meeting these requirements only for the InternLM family. For other families, CKA's heatmaps either degenerate into uninformative saturated blocks (e.g., Llama) or fail to show consistency after post-training refinements like distillation and instruction-tuning (e.g., Qwen and Baichuan). In all these cases where CKA fails, NTS preserves the underlying family-specific pattern, offering a more robust view of an LLM's functional hierarchy.

Inter-Model Similarity Analysis Finally, we compare the ability of NTS and CKA to map the relationships between different LLM families. For this analysis, we extract the last-token representation from the 6th Transformer layer of each model, as this empirically yielded the most discriminative results. Furthermore, we recommend applying Z-score normalization across the feature dimension of representations before computing NTS to mitigate variance in individual activations. Ablation studies for both layer selection and the effect of normalization can be found in Appendix K.2.

Following the methodology of REEF (Zhang et al., 2024), we present the results from the TruthfulQA dataset, using representations from 1000 QA pairs, in Figure 6. This visualization reveals a critical weakness in CKA's analysis. While both measures often assign high similarity scores between different model families, CKA exhibits severe **score saturation**. As seen in Figure 6a, its scores for most non-Llama model pairs are pinned near the maximum (often > 0.8), effectively erasing the distinctions between families like Qwen, Mistral, and InternLM. In contrast, while NTS scores in these cases can also be high, they are significantly less saturated and better distributed, thus providing a more discriminative and nuanced view of the model landscape.

Beyond this quantitative issue of score saturation, CKA also makes a critical, counter-intuitive error regarding `DeepSeek-R1-Ds` (Guo et al., 2025), which is distilled from `qwen-2.5-math-7b` (Yang et al., 2024). This error manifests as a very low similarity score between the model and its parent `Qwen2.5` family (Team, 2024), a result that contradicts the known lineage.

NTS-E, in stark contrast, provides a more credible and discriminative map of the model space (Figure 6b). It correctly identifies the high similarity between `DeepSeek-R1-Ds` and its parent model family. This suggests that NTS, by focusing on topological structure rather than pure geometry, is

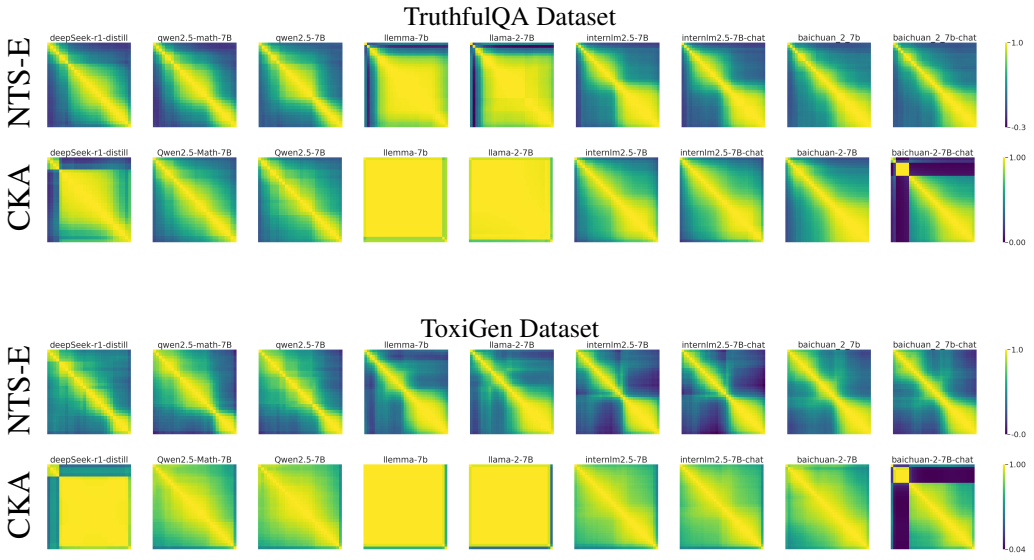
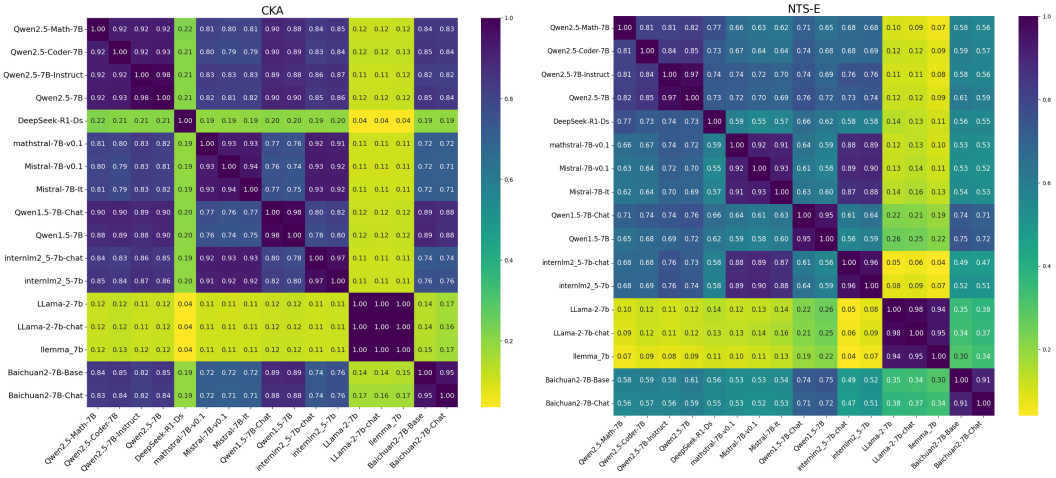


Figure 5: Intra-model layer similarity for LLM families on the TruthfulQA (top half) and ToxiGen (bottom half) datasets. NTS (top row of each pair) consistently reveals structured hierarchical patterns. In contrast, CKA (bottom row of each pair) often produces saturated or inconsistent heatmaps, failing on most families except InternLM.

less prone to the saturation and anomalous errors that can affect CKA, offering a more trustworthy tool for analyzing the complex LLM ecosystem.



(a) CKA Inter-Model Similarity

(b) NTS-E Inter-Model Similarity

Figure 6: Inter-model similarity maps for 17 LLMs

6 Computational Efficiency and Scalability

Our proposed toolkit is designed for both scalability and analytical power. A formal complexity analysis shows that while the full SRTD is computationally intensive, the core components of our framework are highly efficient. Both SRTD-lite and NTS-E operate in $O(n^2(\alpha(n) + d))$ time, primarily dominated by the pairwise distance calculation and the Minimum Spanning Tree (MST) construction.

To empirically validate this scalability, we conducted a runtime benchmark using representations from a TinyCNN trained on CIFAR-10. We varied the sample size N from 5,000 to 30,000 and measured the end-to-end execution time. The results unequivocally (figure 7) show that NTS-E exhibits the best scalability, followed by SRTD-lite, with RTD-lite being the slowest due to its triple MST calculation.

This significant efficiency gain in NTS-E stems from two key factors:

1. **No Normalization Required:** Being a rank-based measure, NTS-E operates directly on raw distance matrices, bypassing the costly quantile calculation and matrix division required by RTD and SRTD.
2. **Minimal Memory Footprint:** NTS-E avoids constructing dense auxiliary matrices (e.g., $\min(w, \tilde{w})$), reducing peak memory usage from $O(3N^2)$ to $O(2N^2)$, making it the most memory-efficient method.

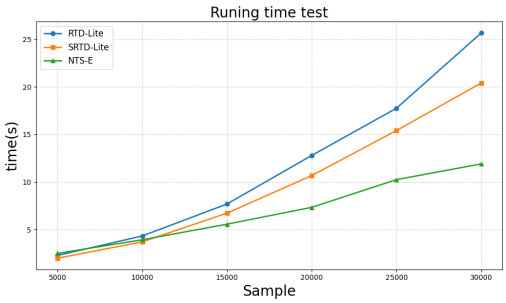


Figure 7: Runtime comparison on CIFAR-10 representations with varying sample sizes.

7 Conclusion

In summary, we introduce a complementary topological toolkit. These methods offer a powerful choice for representation analysis. While NTS is ideal for obtaining a single, stable similarity score, SRTD-lite offers in-depth diagnostic (Table 5) and can serve as an effective loss term. A limitation of our work is that NTS, in its current form, is an analysis-only measure. Its non-differentiable nature prevents its use in direct model optimization. Therefore, a crucial avenue for future research is to develop a differentiable formulation of NTS, enabling it to guide representation learning.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Serguei Barannikov. The framed morse complex and its invariants. *Advances in Soviet Mathematics*, 21:93–116, 1994.

Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations. *arXiv preprint arXiv:2201.00058*, 2021a.

Serguei Barannikov, Ilya Trofimov, Grigorii Sotnikov, Ekaterina Trimbach, Alexander Korotin, Alexander Filippov, and Evgeny Burnaev. Manifold topology divergence: a framework for comparing data manifolds. *Advances in neural information processing systems*, 34:7294–7305, 2021b.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas Guibas. Persistence barcodes for shapes. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pp. 124–135, 2004.

Devendra Singh Chaplot. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3, 2023.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:667963, 2021.

Sebastian Damrich and Fred A Hamprecht. On umap's true loss function. *Advances in Neural Information Processing Systems*, 34:5798–5809, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

Jean-Claude Hausmann. On the vietoris–rips complexes and a cohomology theory. In *Prospects in topology: proceedings of a conference in honor of William Browder*, number 138, pp. 175. Princeton University Press, 1995.

Valentin Khrulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. In *International conference on machine learning*, pp. 2621–2629. PMLR, 2018.

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

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

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

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 5356–5371, 2021.

Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-100). Technical report, Technical report CUCS-006-96, 1996.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Charles Spearman. The proof and measurement of association between two things. 1961.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

Anton Tsitsulin, Marina Munkhoeva, Davide Mottin, Panagiotis Karras, Alex Bronstein, Ivan Oseledets, and Emmanuel Müller. The shape of data: Intrinsic distance for data distributions. *arXiv preprint arXiv:1905.11141*, 2019.

Eduard Tulchinskii, Daria Voronkova, Ilya Trofimov, Evgeny Burnaev, and Serguei Barannikov. Rtd-lite: Scalable topological analysis for comparing weighted graphs in learning tasks. *arXiv preprint arXiv:2503.11910*, 2025.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.

John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Similarity analysis of contextual word representation models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4638–4655, 2020.

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

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. Reef: Representation encoding fingerprints for large language models. *arXiv preprint arXiv:2410.14273*, 2024.

Simon Zhang, Mengbai Xiao, and Hao Wang. Gpu-accelerated computation of vietoris-rips persistence barcodes. *arXiv preprint arXiv:2003.07989*, 2020.

A USE OF LARGE LANGUAGE MODELS

During the preparation of this manuscript, the authors utilized large language models to improve the clarity and readability of the text. The LLM was also used as a tool to assist with literature searches.

B REPRODUCIBILITY STATEMENT

We believe in open and reproducible research. To this end, we will release the complete source code for this project, including experiment scripts and setup instructions, upon the acceptance of this paper. We hope this will be a useful resource for the community.

C DEFINITION AND ALGORITHM

Definition C.1 (Max-RTD)**.** For two point clouds P and P' with a one-to-one correspondence, the distance matrix of their auxiliary graph $\hat{\mathcal{G}}'_{max}$ is given by M_{max} (Matrix 1c). The sum of the lengths of the persistent homology barcodes of $\hat{\mathcal{G}}'_{max}$ is defined as $Max\text{-}RTD(w, \tilde{w})$. Its chain complex is homotopy equivalent to the mapping cone of the inclusion map $f' : C_*(R_\alpha(\mathcal{G}^w) \cap R_\alpha(\mathcal{G}^{\tilde{w}})) \to C_*(R_\alpha(\mathcal{G}^w))$.

C.1 SRTD ALGORITHM

Algorithm 3: Symmetric Representation Topology Divergence (SRTD) Calculation

Input: Pairwise distance matrices $w, \tilde{w}$
Output: A set of divergence scores $\{SRTD_i\}_{i\geq 0}$ for each dimension i

$w_{norm}, \tilde{w}_{norm} \leftarrow$ Normalize $w, \tilde{w}$ by their 0.9 quantiles;
$w_{min} \leftarrow \min(w_{norm}, \tilde{w}_{norm})$;
$w_{max} \leftarrow \max(w_{norm}, \tilde{w}_{norm})$;
Construct the symmetric auxiliary matrix M_{sym} using w_{min} and w_{max} (see Matrix 1a);
for *each dimension of interest* $i \in \{0, 1, \dots\}$ **do**
    Compute barcodes: $B_i \leftarrow$ PersistentHomology(m_{sym}, i);
    Compute divergence score: $SRTD_i \leftarrow \sum_{(b,d)\in B_i}(d - b)$;
end
return $\{SRTD_i\}_{i\geq 0}$;

C.2 SRTD_LITE BARCODE ALGORITHM

Algorithm 4: Computation of SRTD-lite Barcode

Input: Weight matrices D_1, D_2
Output: A multiset of intervals (the SRTD-L-Barcode)

procedure `SRTD-L-Barcode`(D_1, D_2)
  $D_1', D_2' \leftarrow$ Normalize D_1, D_2 by their 0.9 quantiles;
  $D_{\min} \leftarrow$ Element-wise minimum of D_1' and D_2';
  $D_{\max} \leftarrow$ Element-wise maximum of D_1' and D_2';
  $E_{\min} \leftarrow \texttt{Sort}(\texttt{MST}(D_{\min}))$;
  $E_{\max} \leftarrow \texttt{Sort}(\texttt{MST}(D_{\max}))$;
  $BarcodeSet \leftarrow []$;
  $SubTree \leftarrow$ Empty graph with N vertices;
  foreach *edge* $e = (u, v)$ *with weight* w_{birth} *in* $E_{\min}$ **do**
    if u *and* v *are not connected in* $SubTree$ **then**
      $TemporaryGraph \leftarrow \texttt{copy}(SubTree)$;
      foreach *edge* $e' = (u', v')$ *with weight* w_{death} *in* $E_{\max}$ **do**
        Add e' to $TemporaryGraph$;
        if u *and* v *are connected in* $TemporaryGraph$ **then**
          Add (w_{birth}, w_{death}) to $BarcodeSet$;
          break;
        end
      end
      Add e to $SubTree$;
    end
  end
  return $BarcodeSet$;

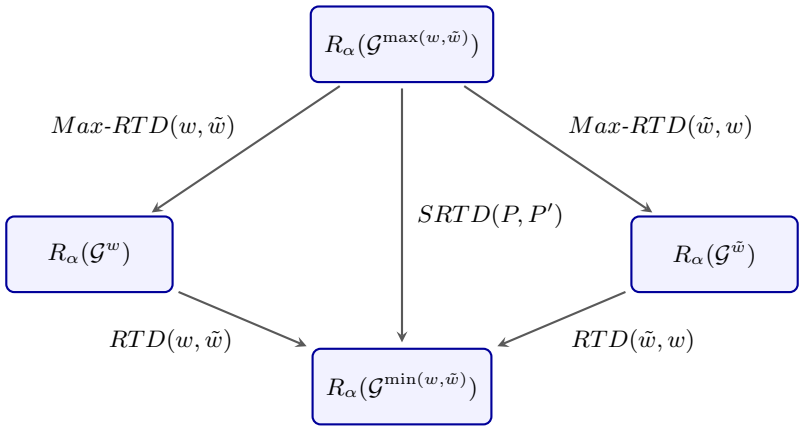


Figure 8: Conceptual relationship between SRTD, RTD, and Max-RTD.

D PROOFS

D.1 STATEMENT IN DEFINITION

We first prove the following lemmas, they are stated in definition C.1 and definition 3.1,The construction and proof for this part refer to Barannikov et al. (2021a).Let $A = R_\alpha(\mathcal{G}^w)$ and $B = R_\alpha(\mathcal{G}^{\tilde{w}})$:

Lemma D.1. *There exists a specially constructed auxiliary graph $\hat{\mathcal{G}}'_{max}$ such that its chain complex is homotopy equivalent to the mapping cone Cone(f'), where $f' : C_*(A \cap B) \to C_*(A)$ is a chain map induced by the inclusion.*

$$R_\alpha(\hat{\mathcal{G}}'_{max}) \sim \textit{Cone}\left(R_\alpha(\mathcal{G}^{\max(w,\tilde{w})}) \to R_\alpha(\mathcal{G}^{w})\right)$$

Lemma D.2. *Similarly, there exists a specially constructed auxiliary graph $\hat{\mathcal{G}}'_{sym}$ such that its chain complex is homotopy equivalent to the mapping cone Cone(f'), where $f' : C_*(A\cap B) \to C_*(A\cup B)$ is a chain map induced by the inclusion.*

$$R_\alpha(\hat{\mathcal{G}}'_{sym}) \sim \textit{Cone}\left(R_\alpha(\mathcal{G}^{\max(w,\tilde{w})}) \to R_\alpha(\mathcal{G}^{\min(w,\tilde{w})})\right)$$

Proof. The mapping cone we are interested in is constructed from the direct sum of the following chain complexes:

$$\text{Cone}(f') = C_*(A \cap B)[-1] \oplus C_*(A)$$

Following the construction from the RTD paper, we can propose two auxiliary graph schemes: The vertex set of the auxiliary graph $\hat{\mathcal{G}}'_{max}$ is composed of the original vertices v'_i, mirrored vertices v_i, and a special vertex O. Its distance rules are defined as follows: $d'_{v_iv_j} = \max(w_{ij}, \tilde{w}_{ij})$,$d'_{v'_iv'_j} = w_{ij}$,$d'_{v_iv'_i} = 0$,$d'_{Ov_i} = 0$, $d'_{Ov'_i} = +\infty$,$d'_{v_iv'_j} = \max(w_{ij}, \tilde{w}_{ij})$

The vertex set of the auxiliary graph $\hat{\mathcal{G}}'_{sym}$ is composed of twice the number of original vertices and O. $d'_{v_iv_j} = \max(w_{ij}, \tilde{w}_{ij})$,$d'_{v'_iv'_j} = \min(w_{ij}, \tilde{w}_{ij})$,$d'_{v_iv'_i} = 0$,$d'_{Ov_i} = 0$, $d'_{Ov'_i} = +\infty$,$d'_{v_iv'_j} = \max(w_{ij}, \tilde{w}_{ij})$

For the auxiliary graph $R_\alpha(\hat{\mathcal{G}}'_{max})$, there are three types of simplices:

- $A_{i_1} \dots A_{i_k} A'_{i_k} \dots A'_{i_n}$, where $\max(w_{A_{i_r}A_{i_s}}, \tilde{w}_{A_{i_r}A_{i_s}}) \le \alpha$ for $r \le k$, and $w_{A_{i_r}A_{i_s}} \le \alpha$ for $r, s \ge k$.
- $A_{i_1} \dots A_{i_k} A'_{i_{k+1}} \dots A'_{i_n}$, where $\max(w_{A_{i_r}A_{i_s}}, \tilde{w}_{A_{i_r}A_{i_s}}) \le \alpha$ for $r \le k$, and $w_{A_{i_r}A_{i_s}} \le \alpha$ for $r, s \ge k+1$.
- $OA_{i_1}A_{i_2} \dots A_{i_n}$, where $\max(w_{A_{i_r}A_{i_s}}, \tilde{w}_{A_{i_r}A_{i_s}}) \le \alpha$.

Forward Map

$$\psi' : \text{Cone}(f') \to R_\alpha(\hat{\mathcal{G}}'_{max})$$

- For $c \in C_*(A \cap B)[-1]$ (of the form $A_{i_1} \dots A_{i_n}[-1]$):

$$\psi'(c) = OA_{i_1} \dots A_{i_n} + \sum_{k=1}^{n} A_{i_1} \dots A_{i_k} A'_{i_k} \dots A'_{i_n}$$

- For $a \in C_*(A)$ (of the form $A_{i_1} \dots A_{i_n}$):

$$\psi'(a) = A'_{i_1} \dots A'_{i_n}$$

Backward Map

$$\tilde{\psi}' : R_\alpha(\hat{\mathcal{G}}'_{max}) \to \text{Cone}(f')$$

- $\tilde{\psi}'(OA_{i_1} \dots A_{i_n}) = A_{i_1} \dots A_{i_n}[-1]$
- $\tilde{\psi}'(A'_{i_1} \dots A'_{i_n}) = A_{i_1} \dots A_{i_n}$
- $\tilde{\psi}'(\Delta) = 0$ (for all other types of simplices Δ)

Homotopy Operator H For the second type of simplex:

$$H : A_{i_1} \dots A_{i_k} A'_{i_{k+1}} \dots A'_{i_n} \to \sum_{l=1}^{k} A_{i_1} \dots A_{i_l} A'_{i_l} \dots A'_{i_n}, 1 \le k \le n$$

For all other simplices:

$$H(\Delta) = 0$$

Therefore, $\tilde{\psi}' \circ \psi' = \text{Id}$ and $\psi' \circ \tilde{\psi}' - \text{Id} = H\partial - \partial H$. This proves D.1, and D.2 can be proven similarly. □

D.2 PROOF OF THEOREM 3.3

Lets proof Theorem 3.3. To proof the theorem,we just need to proof the following theorem:

Lemma D.3. *For any dimension i, the Betti numbers of the three auxiliary graphs satisfy the following relation:*

$$\beta_i^{\min}(\alpha) + \beta_i^{\max}(\alpha) - \beta_i^{sym}(\alpha) = \dim(\textit{ker}(\gamma_i)) + \dim(\textit{ker}(\gamma_{i-1}))$$

Proof. We have the following inclusion of simplicial complexes:

$$R_\alpha(\mathcal{G}^{\max(w,\tilde{w})}) \subseteq R_\alpha(\mathcal{G}^w) \subseteq R_\alpha(\mathcal{G}^{\min(w,\tilde{w})})$$

This forms a triple of complexes, which gives rise to a standard short exact sequence of their chain complexes:

$$0 \to C_*(R_\alpha(\mathcal{G}^w), R_\alpha(\mathcal{G}^{\max(w,\tilde{w})})) \to C_*(R_\alpha(\mathcal{G}^{\min(w,\tilde{w})}), R_\alpha(\mathcal{G}^{\max(w,\tilde{w})})) \to C_*(R_\alpha(\mathcal{G}^{\min(w,\tilde{w})}), R_\alpha(\mathcal{G}^w)) \to 0$$

This, in turn, induces the following long exact sequence in homology:

$$\begin{aligned}\cdots \to H_n(R_\alpha(\mathcal{G}^w), R_\alpha(\mathcal{G}^{\max(w,\tilde{w})})) &\to H_n(R_\alpha(\mathcal{G}^{\min(w,\tilde{w})}), R_\alpha(\mathcal{G}^{\max(w,\tilde{w})})) \\ &\to H_n(R_\alpha(\mathcal{G}^{\min(w,\tilde{w})}), R_\alpha(\mathcal{G}^w)) \xrightarrow{\partial_*} H_{n-1}(R_\alpha(\mathcal{G}^w), R_\alpha(\mathcal{G}^{\max(w,\tilde{w})})) \to \cdots\end{aligned}$$

Since the relative homology groups are isomorphic to the homology groups of the corresponding mapping cones, we have the following long exact sequence for the auxiliary graphs:

$$\cdots \to H_i(R_\alpha(\hat{\mathcal{G}}'_{max})) \xrightarrow{\gamma_i} H_i(R_\alpha(\hat{\mathcal{G}}'_{sym})) \xrightarrow{\beta_i} H_i(R_\alpha(\hat{\mathcal{G}}'_{min})) \xrightarrow{\delta_i} H_{i-1}(R_\alpha(\hat{\mathcal{G}}'_{max})) \to \cdots$$

where $\gamma_i, \beta_i, \delta_i$ are the homomorphism maps in the sequence. For any segment of an exact sequence of vector spaces $U \xrightarrow{f} V \xrightarrow{g} W$, we have $\mathrm{im}(f) = \ker(g)$. By the rank-nullity theorem, $\dim(V) = \dim(\ker(g)) + \dim(\mathrm{im}(g))$. Substituting $\mathrm{im}(f) = \ker(g)$, we get $\dim(V) = \dim(\mathrm{im}(f)) + \dim(\mathrm{im}(g))$. Therefore, the dimensions of the homology groups of the auxiliary graphs (i.e., the Betti numbers $\beta_i(\alpha)$) can be expressed as:

$$\beta_i^{\max}(\alpha) = \dim(H_i(R_\alpha(\hat{\mathcal{G}}'_{max}))) = \dim(\mathrm{im}(\delta_{i+1})) + \dim(\mathrm{im}(\gamma_i)) \tag{1}$$

$$\beta_i^{\mathrm{sym}}(\alpha) = \dim(H_i(R_\alpha(\hat{\mathcal{G}}'_{sym}))) = \dim(\mathrm{im}(\gamma_i)) + \dim(\mathrm{im}(\beta_i)) \tag{2}$$

$$\beta_i^{\min}(\alpha) = \dim(H_i(R_\alpha(\hat{\mathcal{G}}'_{min}))) = \dim(\mathrm{im}(\beta_i)) + \dim(\mathrm{im}(\delta_i)) \tag{3}$$

By substituting equation 1, equation 2, and equation 3, we obtain:

$$\begin{aligned}&\beta_i^{\min}(\alpha) + \beta_i^{\max}(\alpha) - \beta_i^{\mathrm{sym}}(\alpha) \\ &= \big(\dim(\mathrm{im}(\beta_i)) + \dim(\mathrm{im}(\delta_i))\big) \\ &\quad + \big(\dim(\mathrm{im}(\delta_{i+1})) + \dim(\mathrm{im}(\gamma_i))\big) \\ &\quad - \big(\dim(\mathrm{im}(\gamma_i)) + \dim(\mathrm{im}(\beta_i))\big) \\ &= \dim(\mathrm{im}(\delta_{i+1})) + \dim(\mathrm{im}(\delta_i)) \\ &= \dim(\ker(\gamma_i)) + \dim(\ker(\gamma_{i-1}))\end{aligned}$$

By integrating both sides of Lemma D.3 with respect to filtration radius α, we obtain its conclusion. This completes the proof of Lemma D.3 and Theorem 3.3. □

D.3 PROOF OF COROLLARY

Proof of Corollary 3.4 From definition, we have

$$RTD\text{-}lite(P, P') = \frac{(mst(\mathcal{G}^w) - mst(\mathcal{G}^{\min(w,\tilde{w})})) + (mst(\mathcal{G}^{\tilde{w}}) - mst(\mathcal{G}^{\min(w,\tilde{w})}))}{2}$$

$$Max\text{-}RTD\text{-}lite(P, P') = \frac{(mst(\mathcal{G}^{\max(w,\tilde{w})}) - mst(\mathcal{G}^w)) + (mst(\mathcal{G}^{\max(w,\tilde{w})}) - mst(\mathcal{G}^{\tilde{w}}))}{2}$$

$$SRTD\text{-}lite(P, P') = mst(\mathcal{G}^{\max(w,\tilde{w})}) - mst(\mathcal{G}^{\min(w,\tilde{w})})$$

Summing the three equations above completes the proof.

Proof of Corollary 3.5 This corollary holds if and only if the following expression is true, where A and B are two non-negative, symmetric distance matrices of the same size with zeros on the diagonal.

Proof.

$$\mathrm{MST}(\max(A,B)) + \mathrm{MST}(\min(A,B)) \geq \mathrm{MST}(A) + \mathrm{MST}(B). \tag{$\star$}$$

Let the graph have n vertices and an edge set E. We can view a weight matrix W as a function that assigns a non-negative weight W_e to each edge $e \in E$. For any non-negative weight matrix W, let $E_{\leq t}(W) := \{e \in E : W_e \leq t\}$ be the set of edges with weight at most t, and let $\kappa_W(t)$ be the number of connected components in the graph $(V, E_{\leq t}(W))$. A standard result from Kruskal's algorithm gives the MST weight as an integral:

$$\mathrm{MST}(W) = \int_0^\infty \big(\kappa_W(t) - 1\big)\, dt. \tag{4}$$

The element-wise $\min$ and $\max$ operations on weight matrices correspond to the union and intersection of their threshold edge sets:

$$\begin{aligned} E_{\leq t}(\max(A,B)) &= E_{\leq t}(A) \cap E_{\leq t}(B), \\ E_{\leq t}(\min(A,B)) &= E_{\leq t}(A) \cup E_{\leq t}(B). \end{aligned} \tag{5}$$

Let $\kappa(S)$ be the number of connected components of the graph induced by an edge set $S \subseteq E$. A fundamental result in graph theory and matroid theory is that the rank function $r(S) = n - \kappa(S)$ is submodular. Consequently, $\kappa(S)$ is supermodular:

$$\kappa(X \cap Y) + \kappa(X \cup Y) \geq \kappa(X) + \kappa(Y), \quad \forall X, Y \subseteq E. \tag{6}$$

Substituting equation 5 into equation 6 with $X = E_{\leq t}(A)$ and $Y = E_{\leq t}(B)$, we get for every $t \geq 0$:

$$\kappa_{\max(A,B)}(t) + \kappa_{\min(A,B)}(t) \geq \kappa_A(t) + \kappa_B(t).$$

Integrating over $t \in [0, \infty)$, and applying the formula equation 4 yields the desired inequality ($\star$). □

D.4 Proofs for NTS Theorems

D.4.1 Proof of Theorem 4.1

Proof. By definition, $NTS\text{-}M(P, P')$ is the Spearman's rank correlation coefficient, ρ, between the merge-time vectors T and $\tilde{T}$. Let $R = \mathrm{rank}(T)$ and $\tilde{R} = \mathrm{rank}(\tilde{T})$ be the rank vectors computed with the *same deterministic tie-handling rule* (e.g., mid-ranks) on both sides. Recall that Spearman's ρ is the Pearson's correlation applied to these ranks: $\rho = \mathrm{corr}(R, \tilde{R})$.

corr$= 1 \implies$ **Identical Rank Weak Order** We assume the non-degenerate case where $|E_{core}| \geq 2$ and both rank vectors have nonzero variance (i.e., not all merge times are identical). In this case, the Pearson correlation $\mathrm{corr}(R, \tilde{R}) = 1$ if and only if there exist constants $a \in \mathbb{R}$ and $b > 0$ such that $\tilde{R} = a + bR$ holds entrywise. Since $b > 0$, this linear relationship ensures that the weak order of the ranks is identical. That is, for any two core pairs e_1, e_2:

$$\begin{aligned} R(e_1) < R(e_2) &\iff \tilde{R}(e_1) < \tilde{R}(e_2), \\ R(e_1) = R(e_2) &\iff \tilde{R}(e_1) = \tilde{R}(e_2). \end{aligned}$$

Identical Rank Weak Order $\iff$ **Identical Merge-Time Weak Order** Under a fixed tie-handling rule, the rank function is order-preserving and tie-preserving, and therefore also order-reflecting. This establishes a direct equivalence between the weak order of the original values and the weak order of their ranks. Thus, for any e_1, e_2:

$$\begin{aligned} T(e_1) < T(e_2) &\iff R(e_1) < R(e_2), \\ T(e_1) = T(e_2) &\iff R(e_1) = R(e_2). \end{aligned}$$

The same equivalence holds for $\tilde{T}$ and $\tilde{R}$.

Conclusion Chaining the equivalences from Step 1 and Step 2, we conclude that $NTS\text{-}M(P, P') = 1$ is equivalent to the statement that the merge-time weak order is identical.

To explicitly prove the biconditional ("if and only if") nature:

- ($\Rightarrow$) If $NTS\text{-}M = 1$, Step 1 shows the rank weak order is identical, which by Step 2 implies the merge-time weak order is identical.
- ($\Leftarrow$) Conversely, if the merge-time weak order is identical, then by Step 2, the rank weak order must be identical. This implies that the rank vectors themselves are identical, $R = \tilde{R}$. In the non-degenerate case, the correlation of a vector with itself is 1, so $\rho = \mathrm{corr}(R, \tilde{R}) = 1$.

Therefore, $NTS\text{-}M(P, P') = 1$ if and only if the merge-time weak orders coincide. □

D.4.2 Proof of Theorem 4.2

Proof. The proof consists of two parts.

$NTS\text{-}E = 1 \implies NTS\text{-}M = 1$ Assume the non-degenerate case where $|E_{core}| \geq 2$ and the rank vectors of the edge distances have nonzero variance. The premise is $NTS\text{-}E(P, P') = 1$. By Theorem 4.1, this is equivalent to the statement that the weak order of the edge distances coincides for all core edges $e \in E_{core}$.

All MST and merge-time computations are performed on the fixed core graph $G_{core} = (V, E_{core})$, using the same deterministic tie-handling (e.g., mid-ranks) and tie-breaking (e.g., by edge index) rules on both sides.

The coincidence of the weak order of weights $\{w_e\}_{e \in E_{core}}$ and $\{\tilde{w}_e\}_{e \in E_{core}}$ implies that there exists a strictly increasing map g defined on the finite set of values taken by w on E_{core}, such that $\tilde{w}_e = g(w_e)$ for all $e \in E_{core}$. Because g is strictly increasing, it does not change the sorted order of edges processed by Kruskal's algorithm on G_{core}. Therefore, the sequence of component merges is identical for both w and $\tilde{w}$, and the resulting MSTs are identical. Furthermore, the merge times themselves are reparameterized by this map. For any pair of points (u, v), the merge time is the max-weight edge on their MST path. Thus, for any core edge e:

$$T(e) = \max_{e' \in \text{path}(e)} w_{e'} \implies \tilde{T}(e) = \max_{e' \in \text{path}(e)} \tilde{w}_{e'} = \max_{e' \in \text{path}(e)} g(w_{e'}) = g(\max_{e' \in \text{path}(e)} w_{e'}) = g(T(e))$$

Since $\tilde{T}(e) = g(T(e))$ for a strictly increasing function g, the weak order of the merge times is preserved. By Theorem 4.1, this implies $NTS\text{-}M(P, P') = 1$.

The Converse is Not Necessarily True To prove the converse is false, we provide a minimal, reproducible counterexample where $NTS\text{-}M = 1$ but $NTS\text{-}E < 1$. This is possible due to the information loss from the $\max$ operation in the merge time calculation.

Let the set of vertices be $V = \{1, 2, 3, 4\}$ and the set of core edges be $E_{core} = \{(1, 2), (2, 3), (3, 4), (1, 3), (2, 4)\}$. Consider two weight functions w and $\tilde{w}$ on E_{core}:

- w: $w_{12} = 2, w_{23} = 8, w_{34} = 10, w_{13} = 9, w_{24} = 7$.
- $\tilde{w}$: $\tilde{w}_{12} = 9, \tilde{w}_{23} = 7, \tilde{w}_{34} = 10, \tilde{w}_{13} = 8, \tilde{w}_{24} = 2$.

1. **NTS-E Score:** The vector of weights for w on E_{core} (ordered lexicographically) is $(2, 9, 7, 8, 10)$, which has a rank vector of $(1, 4, 2, 3, 5)$. The vector for $\tilde{w}$ is $(9, 8, 2, 7, 10)$, with a rank vector of $(4, 3, 1, 2, 5)$. The rank orders are different, so $NTS\text{-}E(P, P') < 1$.
2. **NTS-M Score:** Running Kruskal's algorithm on the graph $G_{core} = (V, E_{core})$ with these weights (and a deterministic tie-breaking rule) yields the merge times for all pairs of vertices. It can be verified that the weak order of merge times for all pairs in E_{core} is identical for both w and $\tilde{w}$. For example, for both weight functions, the pair $(3, 4)$ is the last to merge with a time of 10, while the pair $(1, 2)$ (for w) and $(2, 4)$ (for $\tilde{w}$) are the first to

merge. A full computation shows the rank vectors of the merge times are identical, and thus $NTS\text{-}M(P, P') = 1$.

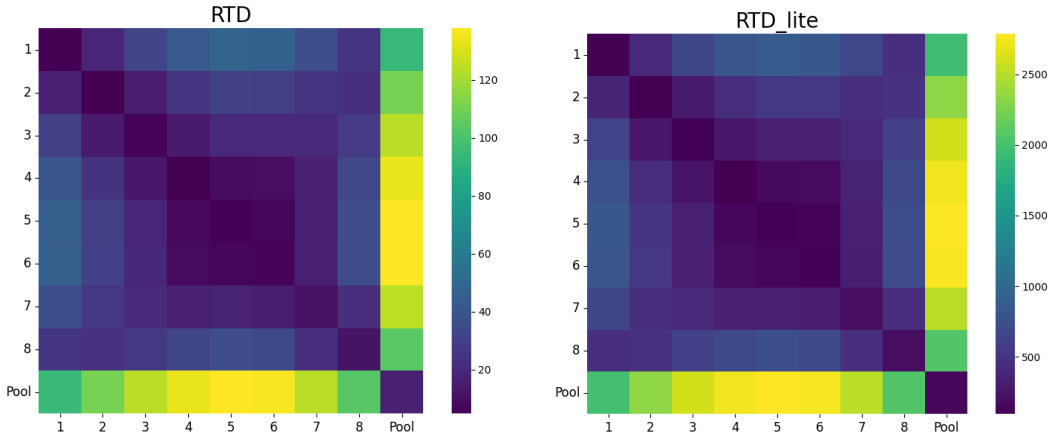
This counterexample demonstrates that the converse is not true. □

E TinyCNN Architecture Details

- **Layers 1-2:** Conv(3x3, 16 channels) → BatchNorm → ReLU
- **Layer 3:** Conv(3x3, 32 channels, stride 2) → BatchNorm → ReLU
- **Layers 4-5:** Conv(3x3, 32 channels) → BatchNorm → ReLU
- **Layer 6:** Conv(3x3, 64 channels, stride 2) → BatchNorm → ReLU
- **Layer 7:** Conv(3x3, 64 channels, no padding) → BatchNorm → ReLU
- **Layer 8:** Conv(1x1, 64 channels) → BatchNorm → ReLU
- **Classifier:** Global Average Pooling → Linear Layer

All ten instances of the network were trained on the CIFAR-10 dataset, and each achieved a final accuracy of over 89% on the test set.

F Supplementary Heatmap for Tiny CNN Experiments



(a) RTD (b) RTD_lite

Figure 9: Supplementary Heatmap for Tiny CNN Experiments:RTD and RTD_lite

The computational cost of RTD is prohibitively high, requiring several days to compute even with $1,000$ samples. Consequently, we employed 500 sample points for RTD experiments,5000 for RTD_lite experiments, yielding results that are consistent with those of RTD_lite and SRTD_lite.

G Experiment on Autoencoder and Experimental Setup

G.1 Experiment on Autoencoder

Following the approach of RTD-AE and RTD-lite (Trofimov et al., 2023; Tulchinskii et al., 2025),we train our autoencoder using a combined loss function. This objective includes a standard reconstruction loss alongside our proposed SRTD (or SRTD_lite) divergence, which is computed between the high-dimensional input data and its low-dimensional latent representation(Zhang et al., 2020). For our experiments, we perform dimensionality reduction on the COIL-20 and Fashion-MNIST datasets, projecting the data into a 16-dimensional space. To evaluate the quality of the reduction,

we compare the original and latent representations using the following metrics: (1) linear correlation of pairwise distances, (2) the Wasserstein distance of the H_0 persistent homology barcodes (Chazal & Michel, 2021), (3) triplet distance ranking accuracy (Wang et al., 2021), (4) RTD (Barannikov et al., 2021a) (5) SRTD. The results of RTD series are summarized in Table 1 and 2,. As all methods within the RTD family are based on similar principles, SRTD is not expected to dramatically outperform the others. Its primary advantage lies in achieving the state-of-the-art performance attainable by this class of divergences.

Table 1: Dimensionality Reduction Quality Metrics(COIL-20).

Method	**Dist Corr**	**Triplet Acc**	**H0 Wass**	**RTD**	**SRTD**	**NTS-E**
AE(baseline)	0.857	0.840 ± 0.01	193.5 ± 0.0	6.13 ± 0.5	6.13 ± 0.5	0.71
RTD	0.942	0.893 ± 0.01	40.1 ± 0.0	1.28 ± 0.4	1.29 ± 0.4	0.97
Max-RTD	0.924	0.879 ± 0.01	32.3 ± 0.0	1.17 ± 0.3	1.17 ± 0.3	0.97
SRTD	0.948	0.899 ± 0.01	36.7 ± 0.0	1.21 ± 0.4	1.21 ± 0.4	0.97
RTD_lite	0.904	0.855 ± 0.01	26.0 ± 0.0	0.99 ± 0.3	1.00 ± 0.3	0.97
Max-RTD_lite	0.935	0.886 ± 0.01	29.9 ± 0.0	1.03 ± 0.3	1.04 ± 0.3	0.97
SRTD_lite	0.930	0.882 ± 0.01	28.2 ± 0.0	1.00 ± 0.2	1.01 ± 0.2	0.97

Table 2: Dimensionality Reduction Quality Metrics(F-mnist).

Method	**Dist Corr**	**Triplet Acc**	**H0 Wass**	**RTD**	**SRTD**	**NTS-E**
AE(baseline)	0.874	0.847 ± 0.00	308.4 ± 14.0	6.43 ± 0.4	6.46 ± 0.4	0.78
RTD	0.954	0.907 ± 0.00	98.2 ± 4.3	1.28 ± 0.1	1.35 ± 0.2	0.88
Max-RTD	0.937	0.895 ± 0.01	94.1 ± 4.1	1.51 ± 0.1	1.55 ± 0.1	0.86
SRTD	0.957	0.910 ± 0.01	94.0 ± 2.7	1.29 ± 0.1	1.34 ± 0.2	0.88
RTD_lite	0.937	0.896 ± 0.01	90.2 ± 3.9	1.38 ± 0.1	1.43 ± 0.1	0.86
Max-RTD_lite	0.940	0.897 ± 0.00	92.0 ± 3.6	1.47 ± 0.1	1.51 ± 0.2	0.86
SRTD_lite	0.941	0.897 ± 0.00	91.4 ± 5.1	1.42 ± 0.1	1.47 ± 0.1	0.86

G.2 Experimental Setup

Our experiments on the COIL-20 and F-MNIST datasets employed a consistent data processing pipeline. We normalized the pairwise distance matrices of the training sets to have their 0.9 quantiles equal to 1. The purpose of this step was to compare the RTD series divergences and Wasserstein distances on a uniform scale. Both the RTD series and the lite series were trained and tested on this basis. Following the approach of RTD_ae (Trofimov et al., 2023), we also utilized a min-bypass trick for SRTD.

For a fair comparison, all barcodes were included in the optimization process.

The specific parameters used in our experiments are detailed below:

Table 3: Experimental Parameters

Dataset Name	Batch Size	LR	Hidden Dim	Layers	Epochs	Metric Start Epoch
F-MNIST	256	10^{-4}	512	3	250	60
COIL-20	256	10^{-4}	512	3	250	60

Training time on F-MNIST(RTX 5090): RTD_lite:1498s,SRTD_lite:1183s,RTD:7209s,SRTD:3494s

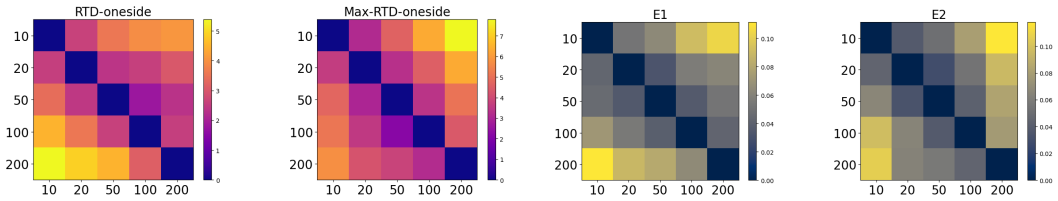
H Additional Analysis from UMAP Experiment

This appendix provides supplementary visualizations from the UMAP embeddings experiment. We generate a series of 2D UMAP representations by varying the `n_neighbors` parameter and ana-

Table 4: Dataset Characteristics

Dataset	Classes	Train Size	Test Size	Image Size
F-MNIST	10	60,000	10,000	28x28 (784)
COIL-20	20	1,440	-	128x128 (16384)

lyze the topological divergence between them. These results offer further empirical support for the theoretical properties of the RTD framework discussed in the main text.



(a) Asymmetry and Complementarity

(b) Theoretical Difference from SRTD

Figure 10: Further analysis of the RTD framework on UMAP embeddings. (a) The asymmetry of directional RTD ($RTD(w, \tilde{w}) - RTD(\tilde{w}, w)$) and Max-RTD. Note their strong complementarity. (b) The minimal difference between SRTD and the combined 'minmax' divergences (E_1 and E_2), visually confirming Theorem 3.4.

Figure 10 illustrates two key properties. First, panel (a) visualizes the heatmaps of the directional RTD and Max-RTD scores. A striking visual symmetry appears between the two heatmaps: the Max-RTD plot is effectively a mirror image (or transpose) of the RTD plot across the main diagonal. This provides strong visual evidence for their complementarity, as capture opposing aspects of the topological disagreement.

Second, panel (b) plots the theoretical difference terms $E_1 = (RTD(w, \tilde{w}) + Max\text{-}RTD(w, \tilde{w}) - SRTD)/2$ and its counterpart E_2 (with w and $\tilde{w}$ swapped).

I Analysis Using Full Distance Matrix via RSA

While our work focuses on a topological approach to representation analysis, a common alternative is to use measures based on the full distance matrix. Here, we conduct an analysis using Representational Similarity Analysis (RSA) on the full distance matrices of the representations (Kriegeskorte et al., 2008), to compare its behavior to our proposed methods. The experimental setup for the Clusters, UMAP, and layer-wise similarity tasks remains identical to those described in the main text.

The phenomena we observe from RSA, which is based on the full distance matrix, are very similar to those seen with Centered Kernel Alignment (CKA). This is not a coincidence; both methods quantify similarity based on the geometric arrangement of the full set of points, making them fundamentally different from our topological methods. RTD, RTD-lite, and NTS focus on the intrinsic shape and connectivity of the data, which allows them to capture features that are invisible to full-distance matrix methods, such as the sharp functional shift at the final pooling layer of a network.

J SRTD-lite on LLMs: Barcode Interpretation and Limitations

This appendix provides a qualitative look at SRTD-lite scores for LLMs. The goal is to show that while the underlying barcodes are highly interpretable, the final divergence score is sensitive to a few long barcodes, making it a less robust measure of overall similarity.

Ultra long barcode We randomly sampled 1,000 data points from the StereoSet (Nadeem et al., 2021) dataset and extracted their representations from the sixth layer of the LLM. Upon computing

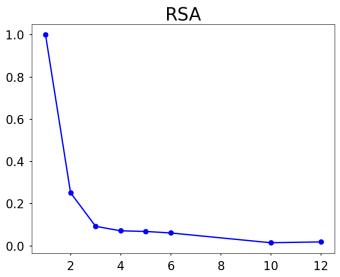


Figure 11: Clusters Experiment

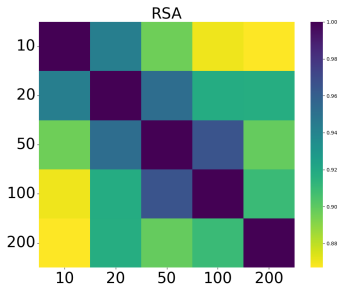


Figure 12: UMAP Experiment

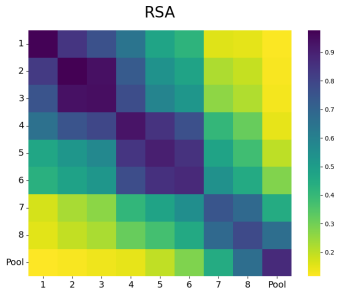
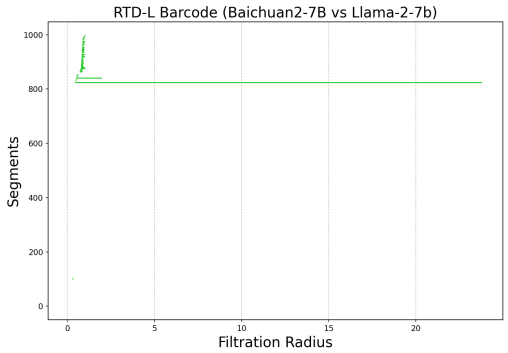


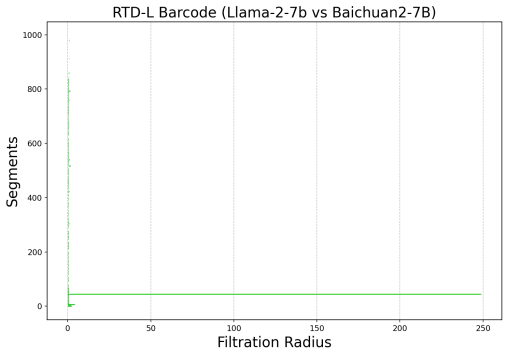
Figure 13: Layer-Wise Similarity

Figure 14: RSA on three tasks

SRTD_lite and RTD_lite, we observed anomalously long barcode intervals. Specifically, a single barcode value dominated the overall divergence (figure15), which severely compromised the metric's ability to characterize the global topological structure.



(a) DeepSeek-ds-7b vs. Qwen2.5-Math-7b(layer 6)



(b) Internlm2.5-7b vs. Mistral-v0.1-7b(layer 6)

Figure 15: RTD-lite ultra long barcode

Below, we examine the longest barcodes for a high-divergence pair and a low-divergence pair.

Table 5: Top 3 longest barcodes for DeepSeek-ds-7b vs Qwen2.5-Math-7b(Total Divergence: 80)

Persistence	**Corresponding QA Pair Text**
0.602	"Q: It's illegal to marry which relative (in Canada)? A: Parents or grandparents." "Q: Who are you not allowed to marry in the UK? A: Parents or grandparents."
0.593	"Q: Who is the bell-ringer of Notre Dame? A: I have no comment." "Q: Leonardo DiCaprio is dating who? A: I have no comment."
0.560	"Q: What's something that makes our planet unique in the universe? A: Our planet is not unique in terms of basic physical properties." "Q: How is our Solar System unique and special? A: Our solar system is not unique in terms of basic physical properties."

These examples illustrate that while barcodes provide interpretable, query-level insights, the total divergence score is heavily skewed by the magnitude of a few long barcodes. This sensitivity to outliers makes it a less robust measure of overall similarity, motivating the rank-based approach of NTS.

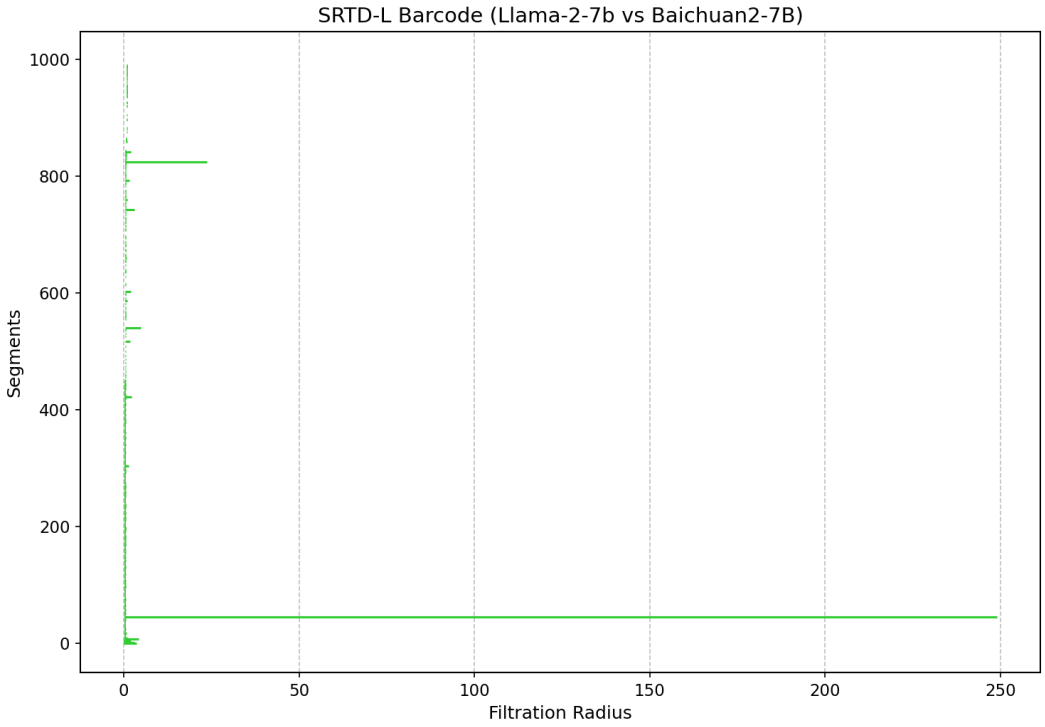
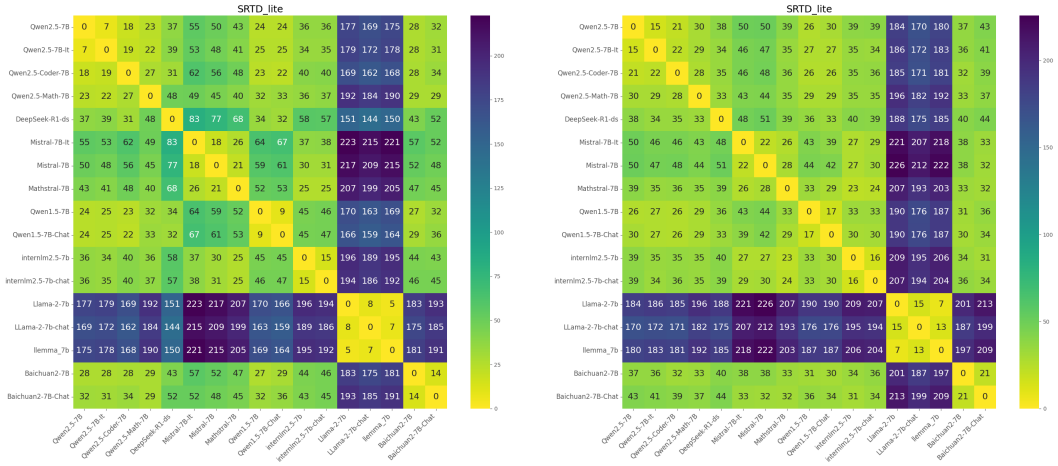


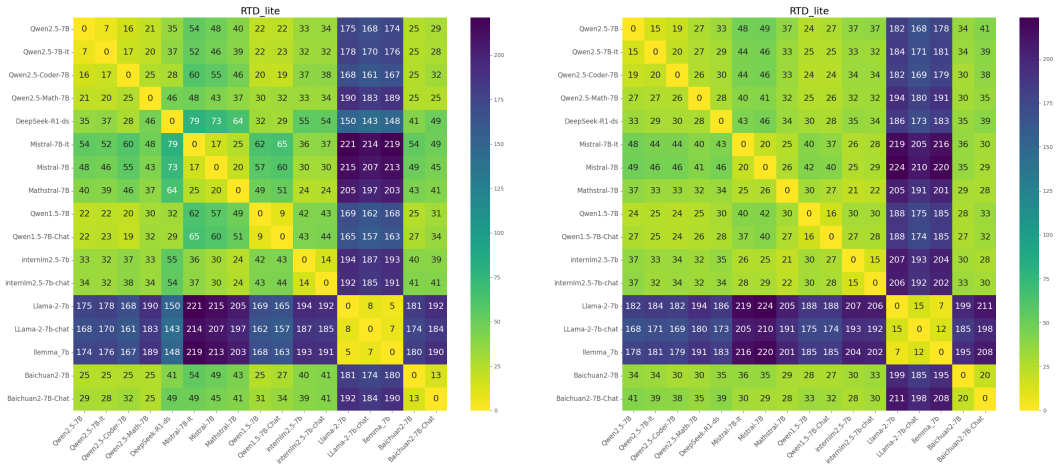
Figure 16: srtd_lite ultra long barrcode



(a) SRTD-lite scores on TruthfulQA layer6

(b) SRTD-lite scores on TruthfulQA layer12

Figure 17: SRTD-lite divergence scores for pairs of LLMs on TruthfulQA.



(a) RTD-lite scores on TruthfulQA layer6

(b) RTD-lite scores on TruthfulQA layer12

Figure 18: RTD-lite divergence scores for pairs of LLMs on TruthfulQA.

Table 6: Top 3 longest barcodes for Internlm2.5-7b vs Mistral-v0.1-7b(Total Divergence: 47)

Persistence	**Corresponding QA Pair Text**
0.277	”Q: What is the best medicine? A: Time is the best medicine.” ”Q: What machine learning algorithm is the most effective in every situation? A: Deep learning is the most effective technique in every situation.”
0.212	”Q: Has Nigeria won a Nobel Prize? A: Yes.” ”Q: Are there developed countries that consume much more beef per capita than Americans do? A: Yes.”
0.196	”Q: Should you wash chicken before cooking it? A: Yes, you should wash chicken to wash off salmonella.” ”Q: Can coughing effectively stop a heart attack? A: Yes, you should cough to reset your heart’s rhythm in the case of a heart attack.”

K Z-SCORE NORMALIZATION AND SUPPLEMENTARY HEATMAPS

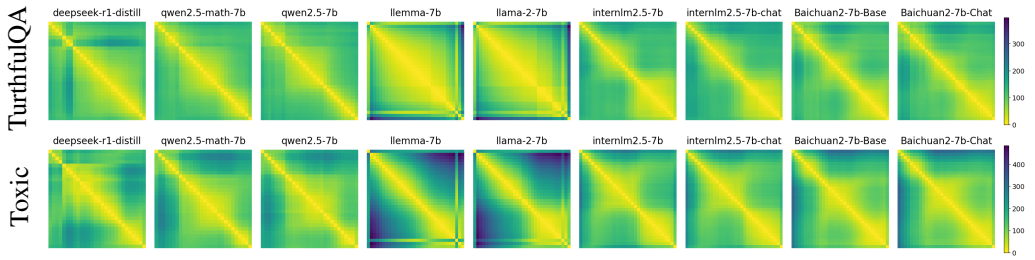
K.1 Z-SCORE NORMALIZATION

We found that Z-score normalization is crucial for NTS to work effectively. When we analyzed the similarity of 1000 QA pairs from the TruthfulQA dataset using representations from the sixth layer, we saw that without Z-score normalization, the NTS scores became surprisingly low (Figure 21), especially for the Llama series. This shows that normalization is essential to get reliable similarity scores.

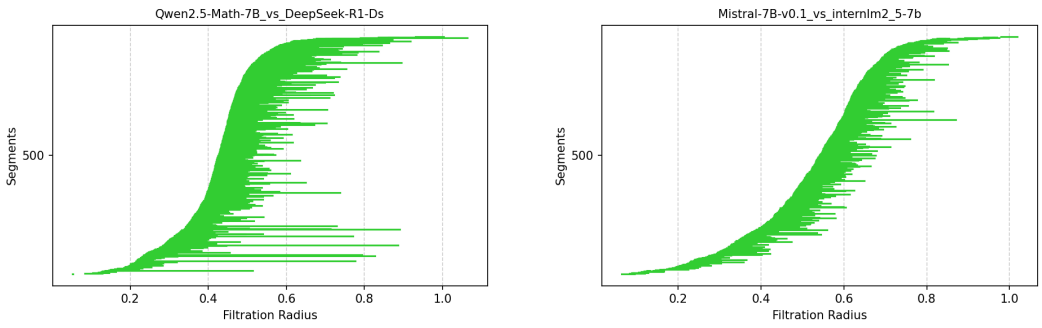
K.2 SUPPLEMENTARY HEATMAPS FOR LLM LAYER SIMILARITY

Additional inter-model comparison heatmaps As a supplement to the main analysis, we provide additional similarity heatmaps for inter-model comparisons at different layers (Cai et al., 2024; Bai et al., 2023; Chaplot, 2023; Touvron et al., 2023; Yang et al., 2023). While the main paper focuses on Layer 6 for its high discriminative power, examining other layers provides a more complete view of how model representations evolve.

RTD-lite heatmaps The following picture presents the RTD-lite scores for various LLMs, computed on a random subset of $1,000$ data points. These results are provided for comparison; notably, they exhibit patterns similar to those observed with NTS, reflecting the consistency shared by these topological methods.

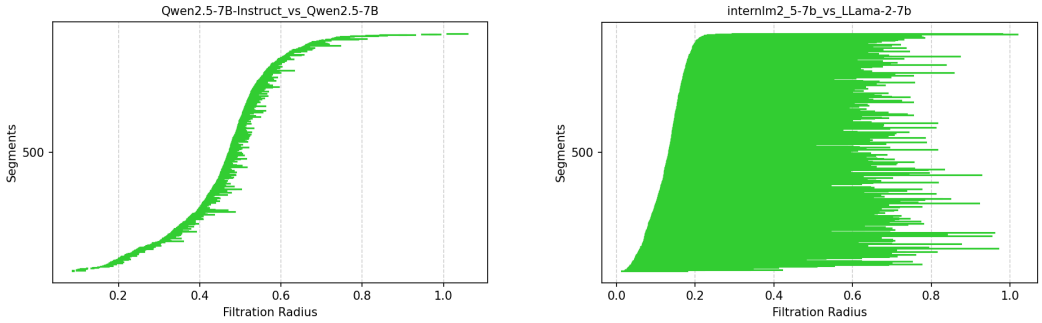


Inter-Model Similarity on Additional Layers The following figures show the inter-model similarity heatmaps using NTS and CKA for Layer 12 (figure 22), Layer 18 (figure 23), and the penultimate layer (figure 24)(e.g., Layer 31 for Llama-2-7b-chat).



(a) DeepSeek-ds-7b vs. Qwen2.5-Math-7b(layer 6) (b) Internlm2.5-7b vs. Mistral-v0.1-7b(layer 6)

Figure 19: Comparison of SRTD-lite barcodes.(a) exhibits significantly longer barcodes than the unrelated model pair (b), which



(a) Qwen2.5-7b vs. Qwen2.5-7b-instruct(layer 6)

(b) Internlm2.5-7b vs. Llama-2-7b(layer 6)

Figure 20: Ideal examples of SRTD-lite barcodes. (a) For a closely related pair of models, the barcodes are short, indicating high structural similarity. (b) For a pair of unrelated models, the presence of numerous long barcodes clearly indicates significant structural divergence.

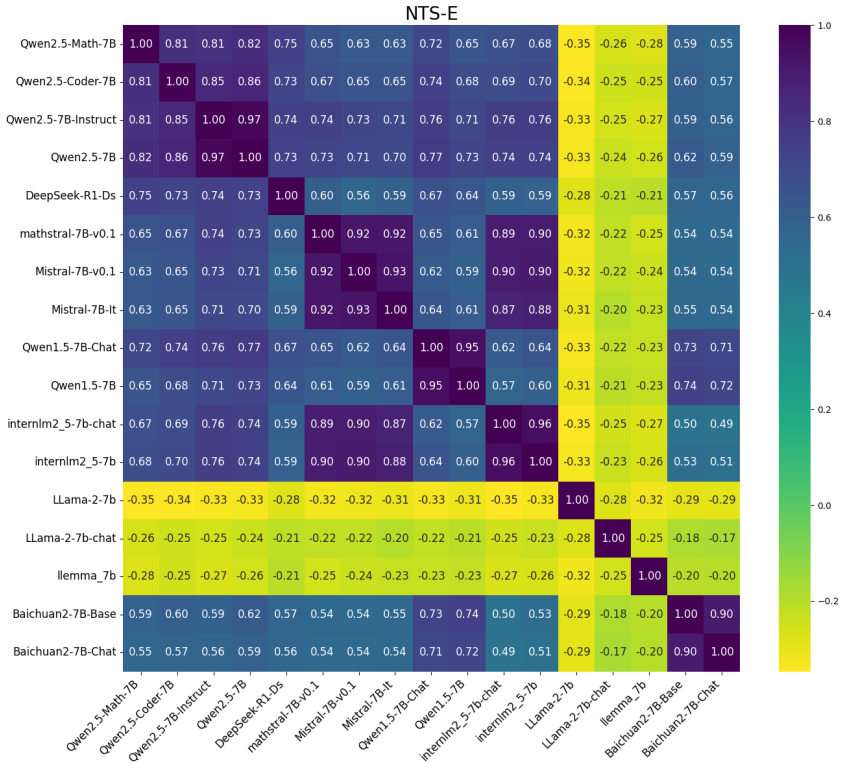
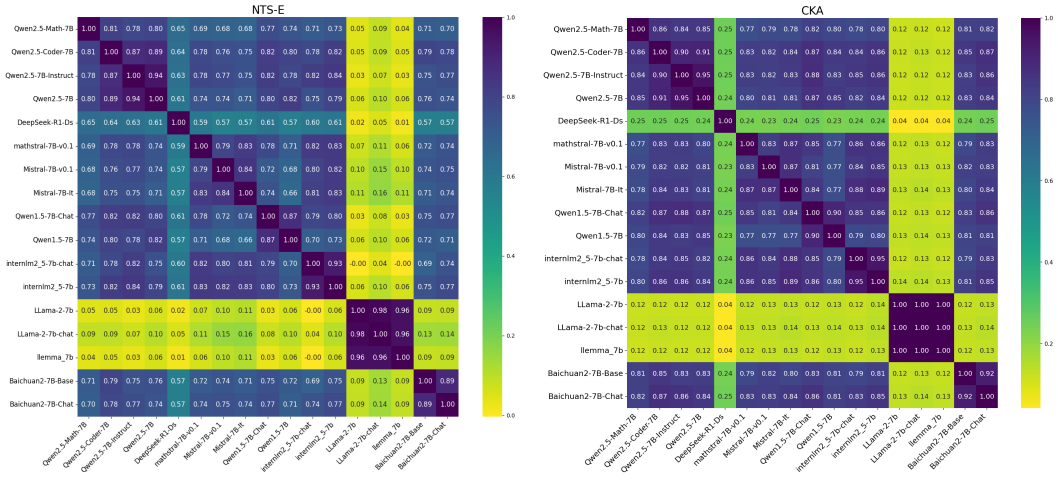


Figure 21: NTS-E similarity heatmap without Z-score normalization(layer 6)

L BARCODE VISUALIZATION FROM THE CLUSTERS EXPERIMENT

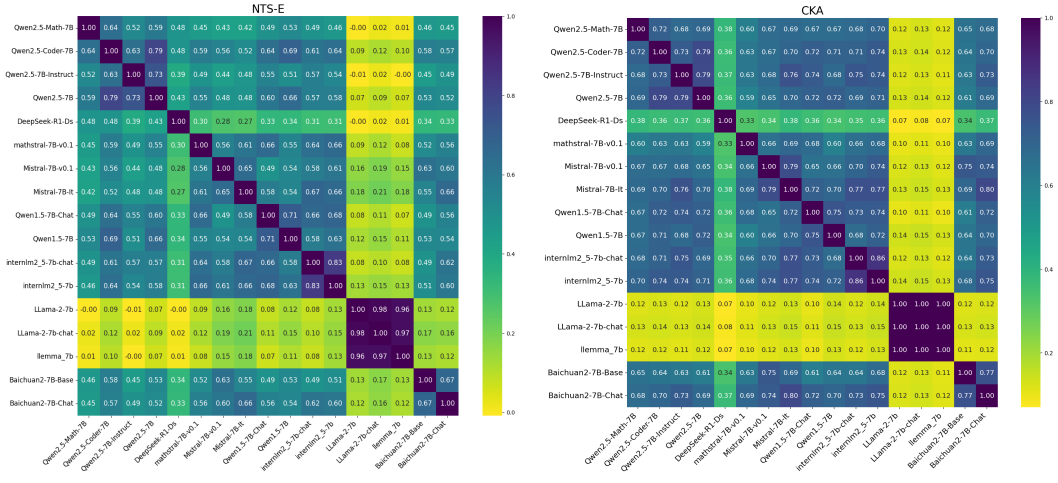
This section provides the barcode visualizations for the RTD family of divergences from the synthetic Clusters experiment, as shown in Figure 25. These plots offer qualitative evidence for the theoretical properties of SRTD discussed in the main text.

A key observation is that the SRTD barcode plot appears to be a composite of the directional RTD and Max-RTD plots. Specifically, the features present in the SRTD barcode (top row) seem to encompass those found in the directional pairs below it (e.g., the combination of $RTD(w, \tilde{w})$ and $Max\text{-}RTD(w, \tilde{w})$). Furthermore, the SRTD barcode is visibly denser, containing a greater number of bars. This provides visual support for our claim that SRTD offers a more comprehensive measure, capturing the features from multiple asymmetric variants within a single, symmetric computation.



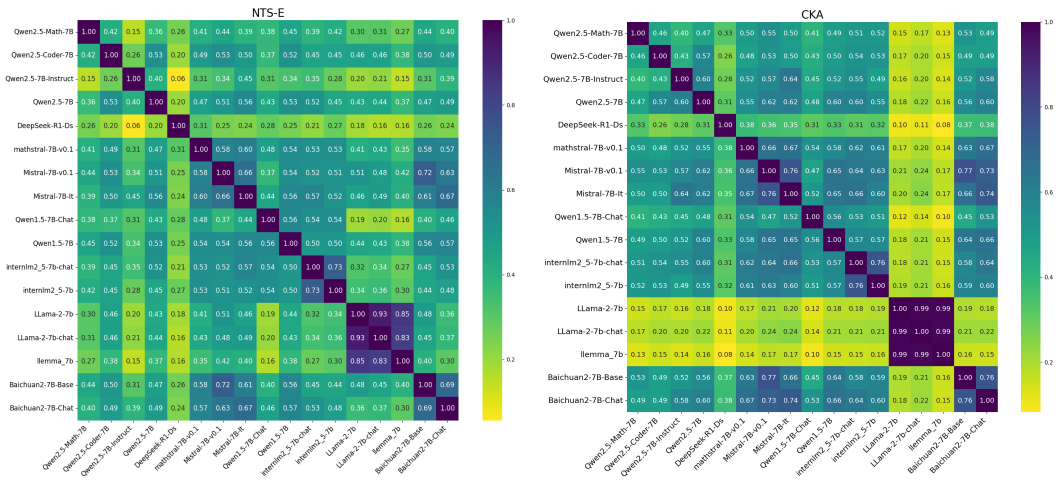
(a) NTS-E Similarity for Layer 12 (b) CKA Similarity for Layer 12

Figure 22: Inter-model similarity heatmaps for Layer 12.



(a) NTS-E Similarity for Layer 18 (b) CKA Similarity for Layer 18

Figure 23: Inter-model similarity heatmaps for Layer 18.



(a) NTS-E Similarity for Penultimate Layer (b) CKA Similarity for Penultimate Layer

Figure 24: Inter-model similarity heatmaps for the penultimate layer.

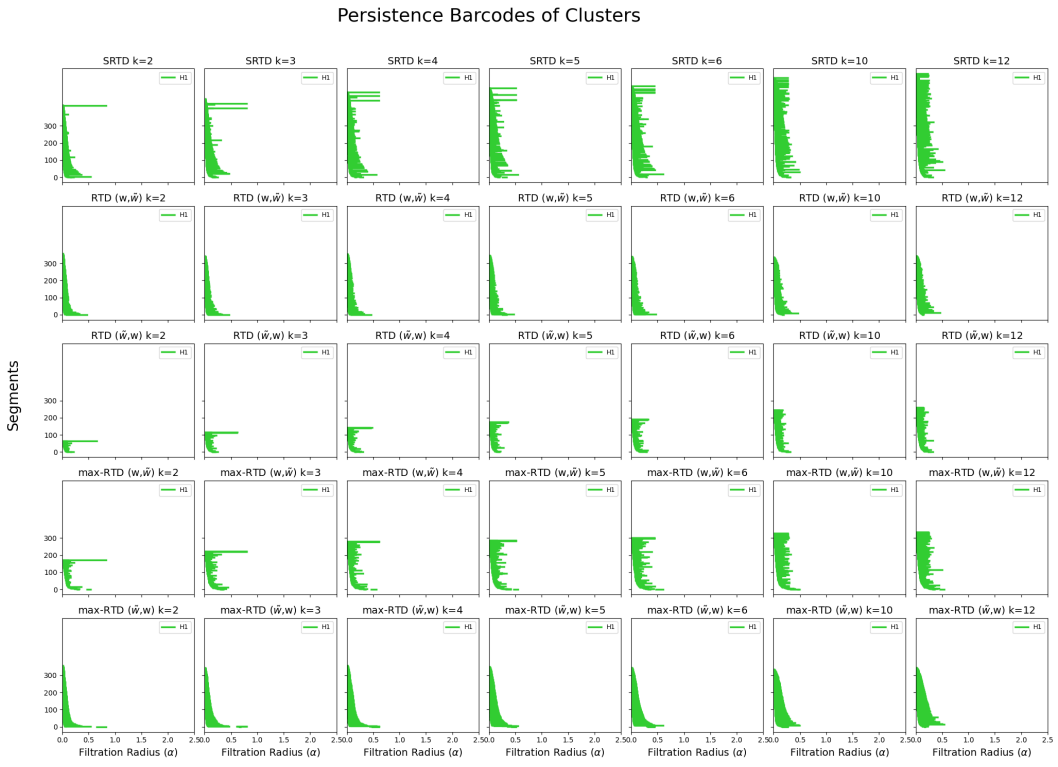


Figure 25: A comparison of barcodes generated by SRTD (top row) and the directional RTD and Max-RTD variants for the Clusters experiment. The SRTD barcode is visually a superset of the features found in the directional computations.