PERSISTENT CROSS ENTROPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent entropy is the Shannon entropy of a persistence-based probability measure defined on a persistence diagram. However, its cross-entropy version is not naturally defined because two persistence diagrams generally have different event spaces. To bridge these different event spaces, we utilize a similarity function with persistence weighting to introduce and define an induced probability. The induced probability reflects information from one diagram on the event space of the other diagram and assigns unexplained probability mass to the unexplained event. Using the induced probability, we introduce persistent cross entropy (PCE), a new extension of Shannon cross entropy to persistence diagrams. We establish the main properties of both the induced probability and PCE and prove stability theorems for the induced probability and the entropy excess. Through three numerical studies, we show that PCE distinguishes diagrams that have nearly identical persistent entropy, separates causal directions in a linear dynamical system without constructing a joint persistence diagram, and can be used as a directional topology loss for knowledge distillation.

1 INTRODUCTION

Persistent homology extracts multiscale topological features from data by tracking how connected components, cycles, and higher-dimensional holes appear and disappear across a filtration (Edelsbrunner et al., 2002). It represents their birth–death structure through persistence diagrams, which provide compact descriptors of the topological structure of complex and high-dimensional data (Carlsson, 2009). Persistent homology, as a primary tool in topological data analysis (TDA), has proven useful in a wide range of real-world applications. For example, persistent homology has been used to analyze protein domains and binding sites (Madsen et al., 2025) and relate spatial immune-cell patterns to clinical outcomes in cancer (Samorodnitsky et al., 2026).

Persistence diagrams are multisets of birth–death pairs in $\mathbb{R}^2$ that may contain different numbers of points and whose points have no canonical ordering, making them difficult to use directly in standard statistical and machine-learning pipelines. Consequently, various representations have been developed to encode persistence information as functions or finite-dimensional vectors. These include functional summaries such as Betti curves (Johnson & Jung, 2021), persistence landscapes (Bubenik, 2015), and persistence silhouettes (Chazal et al., 2015), as well as image-based summaries such as persistence images (Adams et al., 2017). Persistence-weighted kernel representations have also been developed to compare persistence diagrams (Kusano et al., 2016).

Among scalar summaries of persistence diagrams, persistent entropy (PE) normalizes feature persistences to a probability mass function and summarizes it by Shannon entropy (Chintakunta et al., 2015). Its stability and related entropy-based methods have also been studied (Atienza et al., 2020). Recent studies have also used PE to evaluate generative models (Suresh & Abrol, 2025) and relate rest–activity patterns to maternal and neonatal health outcomes (Vagus et al., 2026).

As PE extends Shannon entropy to persistence diagrams, can we do the same for cross entropy? Let p_X and p_Y denote the persistence-based probabilities on two different persistence diagrams D_X and D_Y, respectively. Shannon cross entropy compares two probability measures on a common event space, but p_X and p_Y are defined on different persistence diagrams. Thus the *naive* expression

$$H(p_X, p_Y) = -\sum_{u \in D_X} p_X(u) \log p_Y(u)$$

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.

cannot be applied directly. To deal with this issue, we introduce the *induced probability* p_X^Y, which incorporates information from D_Y and is defined on $D_X \cup \{\partial\}$, where ∂ is the *unexplained event*. We extend p_X to the same space $D_X \cup \{\partial\}$ by setting $p_X(\partial) = 0$. Using p_X^Y in place of p_Y in the above expression, we define *persistent cross entropy* (PCE)

$$H(p_X, p_X^Y) = -\sum_{u \in D_X} p_X(u) \log p_X^Y(u).$$

Section 2 provides the required background on persistence diagrams, and persistence-based probabilities and entropy. Section 3 constructs the induced probability and shows its properties. In this section, we prove that $p_X^X = p_X$ and that $p_X^Y(\partial)$ equals the total variation distance between p_X^Y and p_X and vanishes as the 1-Wasserstein distance between D_X and D_Y tends to zero. We also establish stability under changes in D_Y for fixed D_X. Section 4 provides the definitions of PCE. We show that the entropy excess, $H(p_X, p_X^Y) - H(p_X)$, equals $\mathrm{KL}(p_X \| p_X^Y)$. We further show that the entropy excess vanishes as the 1-Wasserstein distance between D_X and D_Y tends to zero and establish joint local stability when both diagrams vary. Finally, Section 5 provides numerical examples in three settings: distinguishing persistence diagrams with nearly identical PE, causality analysis in linear dynamical systems, and topology-aware knowledge distillation.

2 Background and Setup

2.1 Persistence Diagrams and Coordinates

Given a finite point cloud X in a metric space, persistent homology constructs a multiscale topological summary by building a filtration of simplicial complexes (Edelsbrunner et al., 2002; Carlsson, 2009). In a fixed homological dimension, each feature is recorded by its birth scale b and death scale d. The resulting finite off-diagonal multiset of points (b, d) forms the persistence diagram, and $d - b$ is the persistence of the feature.

Throughout the paper, we use birth–persistence coordinates $z = (b, \ell) \in \mathbb{R}^2$, where $\ell = d - b$. For any persistence point v, we write $v = (b_v, \ell_v)$. The death coordinate is recovered as $d = b + \ell$, so this change of coordinates loses no information. Separating birth from persistence makes explicit both when a feature appears and how long it lasts, while mapping the diagonal to the horizontal axis $\ell = 0$. This representation is also used in persistence images, template functions, and vectorized persistence blocks (Adams et al., 2017; Perea et al., 2023; Chan et al., 2022). It makes persistence weighting and matching to the diagonal simpler. We use $\| \cdot \|$ for the Euclidean norm in these coordinates.

For diagram distances, we follow the standard convention of adjoining the diagonal with infinite multiplicity (Cohen-Steiner et al., 2007). In our measure construction, however, D denotes only the finite off-diagonal multiset. The diagonal is used only for matching in Wasserstein or bottleneck distances. The precise definition of the 1-Wasserstein distance is given in Appendix A.4.

2.2 Persistence Probabilities and Entropy

We now assign a probability to the finite off-diagonal points of a persistence diagram. Let D be a nonempty finite persistence diagram, with all sums over D understood with multiplicity. We associate each $z = (b, \ell) \in D$ with the persistence probability

$$p_D(z) = \frac{\ell}{\sum_{v \in D} \ell_v}.$$

When a point is repeated, each occurrence is treated as a separate outcome. We omit labels for repeated points in the main text and explain how to distinguish them in Appendix A.3. For D_X, we write $p_X := p_{D_X}$. Every off-diagonal point has positive persistence, so $p_X(u) > 0$ for all $u \in D_X$. This probability assigns more mass to long-lived features, which are often interpreted as meaningful topological structure, and less to short-lived features, which may represent topological noise (Atienza et al., 2019).

Once we have constructed the persistence probability p_D from a persistence diagram D, we can naturally define PE as the Shannon entropy of p_D (Chintakunta et al., 2015):

$$H(p_D) = -\sum_{z \in D} p_D(z) \log p_D(z).$$

Throughout the paper, $\log$ denotes the natural logarithm. Note that, unlike an empirical distribution, p_D is determined by persistence rather than sampling frequency. Its entropy is small when persistence is concentrated on a few features and large when it is spread evenly. For a diagram with n points, the maximum $H(p_D) = \log n$ is attained when all persistence probabilities are equal (Cover & Thomas, 2006).

3 Induced Probabilities

Because D_X and D_Y generally have different event spaces, p_X and p_Y cannot be used directly in Shannon cross entropy. To connect the two different event spaces, we combine a similarity function with persistence weighting to measure how well each diagram explains the points of D_X. We then add the unexplained event ∂ to carry the remaining mass, which gives the probability on $D_X \cup \{\partial\}$ induced by D_Y. Section 3.1 introduces the similarity function, Section 3.2 constructs the induced probability, and Section 3.3 establishes its main properties and stability.

3.1 Similarity Between Persistence Points

We first define how similarity is measured between two persistence points.

Definition 1. A map $k : \mathbb{R}^2 \times \mathbb{R}^2 \to (0, 1]$ is called a **similarity function** if it satisfies:

(*i*) there exists a metric ρ on $\mathbb{R}^2$ and a strictly decreasing function $h : [0, \infty) \to (0, 1]$ with $h(0) = 1$ and $\lim_{r \to \infty} h(r) = 0$ such that

$$k(u, v) = h(\rho(u, v));$$

(*ii*) k is Lipschitz in its second argument: there exists $L_0 > 0$ such that

$$|k(u, v) - k(u, v')| \le L_0 \|v - v'\|$$

for all $u, v, v' \in \mathbb{R}^2$.

The birth and persistence coordinates can have different numerical ranges and roles. We therefore use fixed global scales $\sigma_b, \sigma_\ell > 0$, shared by all points and diagrams. For $u = (b_u, \ell_u)$ and $v = (b_v, \ell_v)$, a useful choice is the Gaussian similarity function

$$\kappa(u, v) = e^{-\frac{1}{2}\left\{\left(\frac{b_v - b_u}{\sigma_b}\right)^2 + \left(\frac{\ell_v - \ell_u}{\sigma_\ell}\right)^2\right\}},$$

which satisfies Definition 1. Moreover, it can also be obtained from the maximum-entropy density with prescribed coordinate-wise variances (Jaynes, 1957; Cover & Thomas, 2006). The derivation and Lipschitz bound are given in Appendix A.2.

3.2 Probability Construction

Using the similarity function in Section 3.1, we define a weight for the probability $p_X(u)$ of each point $u \in D_X$ using information from D_Y. The construction starts from the idea that no diagram should explain D_X better than D_X itself. We therefore use self-explanation by D_X as the reference.

Definition 2. Let D_X, D_Y be nonempty finite persistence diagrams in $\mathbb{R}^2$, and let k be a similarity function. Fix a global response scale $\tau > 0$ and a bounded increasing, globally Lipschitz function $g : [0, \infty) \to [0, \infty)$ with $g(0) = 0$. For $u \in D_X$, $v \in D_X \cup D_Y$, and $D \in \{D_X, D_Y\}$, we define the following:

(*i*) $\phi_u(v) = k(u, v) g(\ell_v)$, a score for how well v explains u.

(*ii*) $\alpha_u(D) = \sum_{v \in D} \phi_u(v)$, a score for how well D explains u.

(*iii*) $\delta_X^Y(u) = \alpha_u(D_Y) - \alpha_u(D_X)$, the difference in explanatory power at u.

(*iv*) $w_X^Y(u) = e^{-\frac{\delta_X^Y(u)^2}{2\tau^2}}$, a weight in $(0, 1]$ that increases as $|\delta_X^Y(u)|$ decreases.

The score $\alpha_u(D_X)$ is the reference for explaining u. Matching this score gives $w_X^Y(u) = 1$; either a larger or smaller score from D_Y reduces the weight.

The similarity function measures proximity, but persistence must also be weighted. For the fixed-scale Gaussian similarity, g must be bounded; otherwise, a diagram-independent stability bound does not hold (Proposition 3 in Appendix A.1). The function g allows points with greater persistence to receive larger weights while keeping their contributions bounded. For example, $g(t) = \frac{t}{1+t}$ satisfies these conditions. The regularity conditions on g are given in Appendix A.1.

The product $w_X^Y(u)p_X(u)$ is the mass at u explained by D_Y, but these masses generally sum to less than one. Equality holds only when $w_X^Y(u) = 1$ for every $u \in D_X$. We therefore introduce an unexplained event to carry the remaining mass.

Definition 3. Let $\partial \notin D_X$ denote the **unexplained event**. The **induced probability** on $D_X \cup \{\partial\}$ is

$$p_X^Y(u) = \begin{cases} w_X^Y(u)p_X(u), & u \in D_X, \\ \sum_{v \in D_X} \big(1 - w_X^Y(v)\big)p_X(v), & u = \partial. \end{cases}$$

One can regard p_X as a probability measure on the same space $D_X \cup \{\partial\}$ by setting $p_X(\partial) = 0$. Appendix A.3 verifies that p_X^Y is indeed a probability measure.

For each $u \in D_X$, the quantity $1 - w_X^Y(u)$ determines the fraction of the mass at u assigned to the unexplained event. Consequently, $(1 - w_X^Y(u))p_X(u)$ is the amount of unexplained probability mass from u. The value $p_X^Y(\partial)$, obtained by summing these amounts over D_X, is therefore the total unexplained mass.

3.3 Properties of Induced Probabilities

Since p_X^Y encodes information from D_Y on the event space of $D_X \cup \{\partial\}$, the self-induced probability p_X^X should be equal to p_X. We first verify this consistency and then study how the induced probability changes as the two diagrams become closer, using total variation and the 1-Wasserstein distance.

Proposition 1. *The induced probability satisfies the following properties.*

(i) Comparing D_X with itself recovers the original probability:

$$p_X^X = p_X.$$

(ii) The induced probability equals p_X exactly when the difference in explanatory power vanishes on D_X:

$$p_X^Y = p_X \quad \Longleftrightarrow \quad \delta_X^Y = 0 \quad \text{on} \quad D_X.$$

(iii) The total variation distance from p_X is exactly the unexplained mass:

$$d_{\mathrm{TV}}(p_X^Y, p_X) = p_X^Y(\partial).$$

The proof is given in Appendix A.5.

We next use the 1-Wasserstein distance to control the unexplained mass and changes in the induced probability.

Theorem 1. *Let D_X, D_Y be nonempty finite persistence diagrams. There exists a diagram-independent constant $C > 0$, depending only on the similarity function k and on g with τ fixed in Section 3.2, such that the induced probability satisfies the following properties.*

(i) The unexplained mass vanishes as D_Y approaches D_X:

$$p_X^Y(\partial) \leq C W_1(D_X, D_Y)^2.$$

(ii) For fixed D_X, the induced probability is stable under changes in D_Y; for any nonempty finite persistence diagram $D_{Y'}$,

$$d_{\mathrm{TV}}(p_X^Y, p_X^{Y'}) \leq C W_1(D_Y, D_{Y'}).$$

The proof is given in Appendix A.7.

4 Persistent Cross Entropy

The probabilities p_X and p_X^Y are defined on the same space $D_X \cup \{\partial\}$. Since $p_X(\partial) = 0$, the unexplained event does not contribute to their cross entropy.

Definition 4. The **persistent cross entropy** of D_X relative to D_Y is

$$H(p_X, p_X^Y) = -\sum_{u \in D_X} p_X(u) \log p_X^Y(u).$$

Its **entropy excess** is $\Delta H_X^Y = H(p_X, p_X^Y) - H(p_X)$.

The entropy excess measures differences in how D_X and D_Y explain the points of D_X under the chosen similarity function and persistence weighting. It is directional because the reverse comparison uses p_Y on $D_Y \cup \{\partial\}$; hence $H(p_X, p_X^Y) \neq H(p_Y, p_Y^X)$ in general.

Proposition 2. *For every pair of nonempty finite persistence diagrams,*

$$\Delta H_X^Y = H(p_X, p_X^Y) - H(p_X) = \mathrm{KL}(p_X \| p_X^Y) = \frac{1}{2\tau^2} \mathbb{E}_{u \sim p_X}\left[\delta_X^Y(u)^2\right] \geq 0.$$

The proof is given in Appendix A.8.

Together with Proposition 1, this identity shows that the entropy excess vanishes exactly when $p_X^Y = p_X$; in particular, $H(p_X, p_X^X) = H(p_X)$.

We next show that the entropy excess vanishes as D_Y approaches D_X in the 1-Wasserstein distance.

Theorem 2. *There exists a diagram-independent constant C[1] > 0 such that, for all nonempty finite persistence diagrams D_X, D_Y,*

$$0 \leq \Delta H_X^Y \leq C W_1(D_X, D_Y)^2.$$

The proof is given in Appendix A.9.

We finally establish joint local stability when both diagrams vary.

Theorem 3. *Fix nonempty finite persistence diagrams D_X, D_Y. The entropy excess is locally stable in both diagrams: there exists a finite constant $C_{X,Y}$ such that, whenever $D_{X'}$ and $D_{Y'}$ are sufficiently close to D_X and D_Y, respectively,*

$$|\Delta H_X^Y - \Delta H_{X'}^{Y'}| \leq C_{X,Y}\left(W_1(D_X, D_{X'}) + W_1(D_Y, D_{Y'})\right).$$

The same constant $C_{X,Y}$ applies throughout this neighborhood and does not depend on $D_{X'}$ or $D_{Y'}$. In particular, if $D_{X'} = D_X$, then

$$|\Delta H_X^Y - \Delta H_X^{Y'}| \leq C_{X,Y} W_1(D_Y, D_{Y'}).$$

The complete proof, including an explicit neighborhood and a choice of $C_{X,Y}$, is given in Appendix A.10.

[1] The constant C can be chosen to be the same as in Theorem 1.

5 EXPERIMENTS

We evaluate PCE through three experiments. Section 5.1 shows PCE can distinguish persistence diagrams that have nearly identical PE. Section 5.2 studies causality in a linear dynamical system and compares PCE with symmetric persistence-diagram methods that require a joint manifold reconstruction. Section 5.3 then applies PCE to knowledge distillation and evaluates it as a directional topology loss for transferring topological information. Full numerical settings are given in Appendix B.

5.1 TWO LOOPS WITH NEARLY EQUAL PE

We consider five point clouds. The point cloud X contains two loops of different sizes, along with several noise components. The point cloud Y_1 contains a large loop and a small loop without surrounding noise, Y_2 contains only a large loop, Y_3 contains only a small loop, and Y_4 contains only noise. Note that PE values of their H_1 persistence diagrams all equal 1.500 when rounded to three decimal places.

The unexplained mass and PCE can be written as

$$p_X^{Y_i}(\partial) = \sum_{u \in D_X} \big(p_X(u) - p_X^{Y_i}(u)\big),$$

$$H(p_X, p_X^{Y_i}) = H(p_X) + \sum_{u \in D_X} p_X(u) \log \frac{p_X(u)}{p_X^{Y_i}(u)}.$$

With $p_X(u)$ fixed, a smaller $p_X^{Y_i}(u)$ increases both $p_X(u) - p_X^{Y_i}(u)$ and $p_X(u) \log\big(\frac{p_X(u)}{p_X^{Y_i}(u)}\big)$. Figure 1 shows that $p_X^{Y_i}(\partial)$ and $H(p_X, p_X^{Y_i})$ increase from Y_1 to Y_4. Let $u_L, u_S \in D_X$ denote the points for the large and small loops. At both points, $p_X^{Y_1}(u)$ is largest and $p_X^{Y_4}(u)$ is smallest among the four comparisons. From Y_2 to Y_3, the mass at u_L decreases from approximately 0.325 to 0.068, while the mass at u_S is nearly unchanged. The decrease at u_L explains the larger unexplained mass and PCE for Y_3. For Y_4, the increases in the two loop terms exceed the decreases at noise points, giving the largest values for both quantities.

5.2 SPRING–MASS BENCHMARK

We study causal directions in a linear dynamical system using the spring–mass benchmark between the dynamical systems A for $x(t)$ and B for $y(t)$ (Bando et al., 2022),

$$\ddot{x} = -x + 0.7\alpha y, \qquad \ddot{y} = -0.7y + \beta x,$$

where α controls the influence $B \to A$ and β controls $A \to B$. The coupling grid contains independent, one-way, and bidirectional regimes. We turn each scalar time series into a point cloud of delay vectors, such as $X_A = \{(x_t, x_{t+q}, \ldots, x_{t+(E-1)q})\}_t$ and $X_B = \{(y_t, y_{t+q}, \ldots, y_{t+(E-1)q})\}_t$ (Takens, 1981), and denote the resulting persistence diagrams by D_A and D_B. In this benchmark, directional comparison with the symmetric baselines requires a joint reconstruction from both series (Cao et al., 1998; Bando et al., 2022). Its diagram is D_{AB}, and the direction is assessed by comparing $d(D_{AB}, D_A)$ and $d(D_{AB}, D_B)$. The benchmark associates $d(D_{AB}, D_B) < d(D_{AB}, D_A)$ with $A \to B$ and the reverse ordering with $B \to A$. Full settings are given in Appendix B.2.

We compare PCE with three groups of methods, shown in Figure 2. First, bottleneck and 2-Wasserstein distances compare diagrams through optimal matchings (Cohen-Steiner et al., 2007; Mileyko et al., 2011). Second, Betti curves (Johnson & Jung, 2021), persistence landscapes (Bubenik, 2015), and persistence silhouettes (Chazal et al., 2015) summarize diagrams as functions; we use uniform silhouette weights. Finally, the persistence scale-space kernel (Reininghaus et al., 2015) and persistence images (Adams et al., 2017) give smoothed representations.

Unlike these baselines, PCE directly produces the directed pair $(\Delta H_B^A, \Delta H_A^B)$ from D_A and D_B, without constructing D_{AB}. In this benchmark, $\Delta H_A^B < \Delta H_B^A$ for $A \to B$, and the inequality is reversed for $B \to A$. Figure 2 shows that all 16 one-way cases lie in the expected half-plane.

The unexplained mass $p_X^Y(\partial)$ gives a bounded view of the same asymmetry. The median pair $(p_B^A(\partial), p_A^B(\partial))$ is $(0.888, 0.024)$ for $A \to B$ and $(0.014, 0.883)$ for $B \to A$. The independent

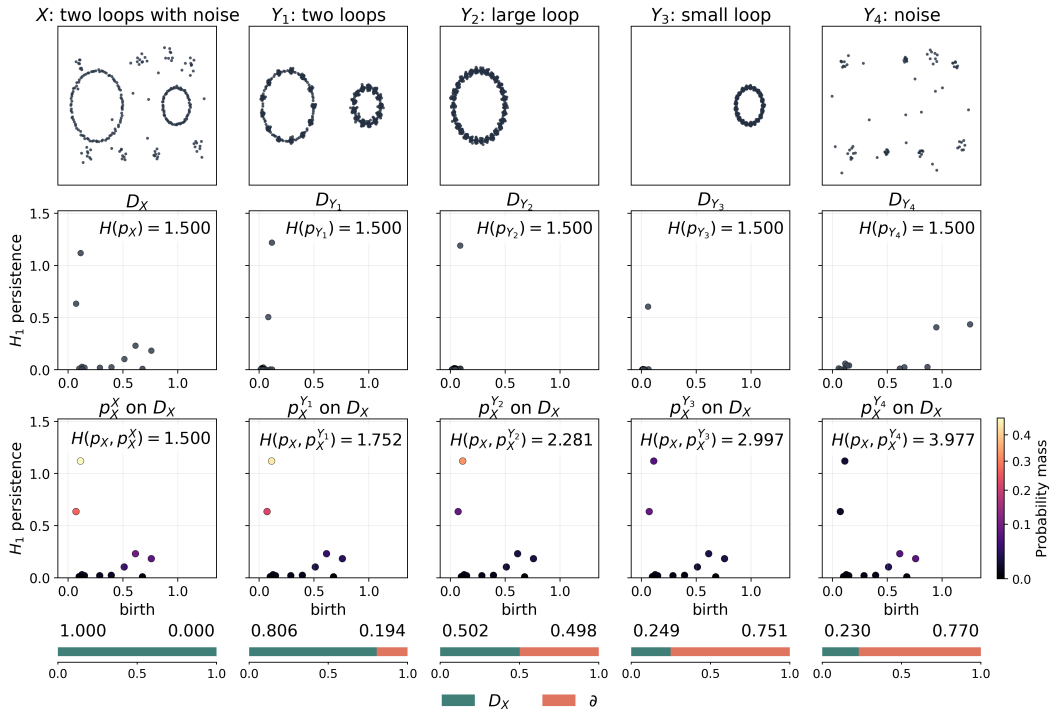

Figure 1: Two-loop experiment with nearly equal PE. The first row shows the point clouds $X, Y_1, \ldots, Y_4$, and the second row shows $D_X, D_{Y_1}, \ldots, D_{Y_4}$. Each diagram is annotated with its PE. The third row shows the actual augmented probability masses $p_X^X(u), p_X^{Y_1}(u), \ldots, p_X^{Y_4}(u)$ for $u \in D_X$ on a common color scale and reports the corresponding PCE. These masses are not renormalized within D_X. Each bar partitions unit mass between D_X and the unexplained event ∂, with the two masses shown above.

Table 1: Separation of the coupling regimes in the two-dimensional representations shown in Figure 2. Larger PERMANOVA R^2 indicates stronger separation. The best result in each evaluation is shown in bold.

Method	Four regimes R^2	Two directions R^2
Entropy excess	**0.560**	0.740
Unexplained mass	0.449	**0.993**
Bottleneck	0.180	0.733
Wasserstein	0.050	0.662
Betti curve	0.091	0.513
Persistence landscape	0.197	0.876
Persistence silhouette	0.082	0.616
Persistence scale-space kernel	0.082	0.678
Persistence image	0.107	0.508

case has almost equal values in both directions, so it shows no directional preference; their small size alone should not be read as evidence of coupling.

To compare the nine cases numerically, we treat the two plotted coordinates of each method as a two-dimensional representation and measure regime separation by PERMANOVA R^2 with Euclidean distance (Anderson, 2001). The four-regime evaluation uses all 81 settings with the independent, $A \to B$, $B \to A$, and bidirectional labels. The two-direction evaluation uses only the balanced set of 16 one-way cases. These scores measure between-group separation relative to total squared variation; they are not classification accuracies. In particular, the four-regime result is descriptive because the group sizes are $1, 8, 8$, and 64.

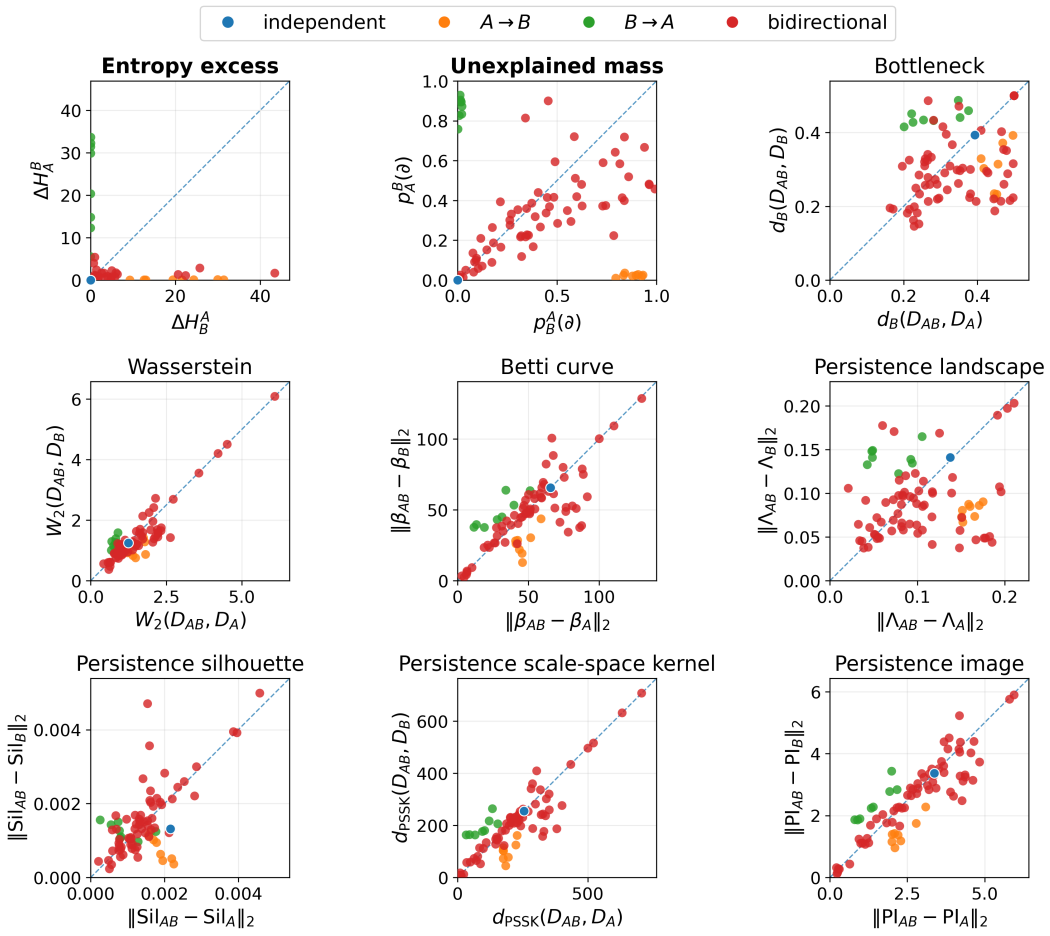

Figure 2: Spring–mass directional comparison. Each point represents one coupling pair, colored by regime. The dashed line is $y = x$; the lower-right half favors $A \to B$, the upper-left half favors $B \to A$, and points near the diagonal show little directional preference. The two PCE-based panels are computed only from D_A and D_B. They plot $(\Delta H_B^A, \Delta H_A^B)$ and $(p_B^A(\partial), p_A^B(\partial))$ for entropy excess and unexplained mass, respectively. Every remaining panel uses the joint diagram D_{AB} and plots $(d(D_{AB}, D_A), d(D_{AB}, D_B))$. Method details are given in Appendix B.2.

Entropy excess gives the strongest four-regime separation, followed by unexplained mass. For the balanced one-way comparison, unexplained mass reaches $R^2 = 0.993$, followed by the persistence landscape at 0.876.

The spring–mass experiment shows that PCE can identify causal directions in the one-way cases of this linear dynamical system. Our preliminary tests on nonlinear and chaotic systems produced mixed results for both PCE and methods based on a joint attractor. Further study is needed to understand when PCE works well in these systems.

5.3 Topology-aware knowledge distillation

Topology-aware knowledge distillation transfers the global structure of intermediate features in addition to class probabilities. TopKD approximates the persistence images of teacher and student feature clouds with a frozen RipsNet (de Surrel et al., 2022) and minimizes their mean squared difference (Kim et al., 2024). We use PCE as a directional topology loss.

For each mini-batch, we ℓ_2-normalize the teacher and student features and compute their exact finite H_0 Vietoris–Rips persistence diagrams, D_T and D_S. The student diagram D_S induces p_T^S on

$D_T \cup \{\partial\}$, with $p_T(\partial) = 0$. We do not renormalize the mass retained on D_T. The topology loss is the PCE excess

$$\Delta H_T^S = \mathrm{KL}(p_T \| p_T^S) = \frac{1}{2\tau^2} \sum_{u \in D_T} p_T(u) \delta_T^S(u)^2.$$

Here p_T normalizes the teacher persistence lengths, and $g(t) = \frac{t}{1+t}$ weights the contributions used to compute δ_T^S. In ΔH_T^S, the points $u \in D_T$ and their weights $p_T(u)$ are set by the fixed teacher, so the loss measures how well the student explains the teacher's features. The reverse loss ΔH_S^T measures how well the teacher explains the student's features. We add no separate penalty for the unexplained mass.

With $\mathcal{L}_{\mathrm{base}} = \mathcal{L}_{\mathrm{CE}} + 2\mathcal{L}_{\mathrm{KD}}$, we compare three objectives:

$$\begin{aligned} \text{KD:} \quad & \mathcal{L}_{\mathrm{base}}, \\ \text{TopKD:} \quad & \mathcal{L}_{\mathrm{base}} + 5\mathcal{L}_{\mathrm{PI}}, \\ \text{PCE:} \quad & \mathcal{L}_{\mathrm{base}} + 0.1\Delta H_T^S. \end{aligned}$$

Here, $\mathcal{L}_{\mathrm{PI}}$ is the mean squared difference between the 400-dimensional RipsNet outputs. PCE instead computes its loss directly from exact H_0 persistence diagrams, without a pretrained approximation model or a persistence-image grid.

We use CIFAR-100 (Krizhevsky, 2009) with a ResNet56 teacher and a ResNet20 student (He et al., 2016), following TopKD's training protocol (Kim et al., 2024). We report results across five paired seeds. Each method trains for 240 epochs, sharing student initialization and augmented training inputs within each seed. TopKD uses raw features, whereas PCE uses normalized features. We select the coefficient 0.1 using a training-only validation split. For PCE, σ_ℓ and τ are calibrated once per seed using training batches and then fixed. Full settings and an additional comparison with an adaptive-weight variant, $\mathrm{PCE}_{\mathrm{AW}}$, are given in Appendix B.3.

Table 2: CIFAR-100 Top-1 accuracy (%), averaged across five seeds. Best observed test is the highest test accuracy over 240 epochs for each seed; Last-10 mean is the mean over epochs 231–240. Sample standard deviations are given in Table 5.

Method	Best observed test	Last-10 mean
KD	71.18	70.98
TopKD	71.21	70.99
PCE	**71.36**	**71.17**

Adding $0.1\Delta H_T^S$ to KD improves Last-10 accuracy in all five paired seeds, with a mean gain of 0.19 percentage points. PCE also exceeds TopKD by 0.18 points on this measure and performs better on both measures in four of five seeds. These results support the use of PCE as an effective topology loss for neural-network training. In this experiment, it improves knowledge-distillation accuracy without a separately pretrained topology model.

6 Conclusion

We introduced PCE, a new extension of Shannon cross entropy to persistence diagrams that recovers PE when a diagram is compared with itself: $H(p_X, p_X^X) = H(p_X)$. The construction combines a similarity function with persistence weighting to define the probability induced by one diagram on the other, with an additional event for unexplained mass. We developed a theoretical foundation for PCE. We established the main properties of the induced probability and expressed the entropy excess as a KL divergence and a scaled expected squared difference in explanatory power. We proved stability bounds for both, including joint local stability of the entropy excess.

We evaluated PCE empirically through three experiments. The results show that PCE provides finer discriminative power than PE, demonstrates its potential for causal analysis by capturing directional information in dynamical systems, and suggests its potential as a topological loss for knowledge distillation. These results support PCE as both a directional comparison of persistence diagrams and an effective topology loss for learning. Its direction allows the comparison to reflect the different roles of the two diagrams.

REPRODUCIBILITY STATEMENT

Appendix A provides complete proofs of the theoretical results, including all regularity assumptions and constants used in the stability bounds. Appendix B specifies the data-generation procedures, preprocessing steps, parameter values, random seeds, model architectures, optimization settings, and evaluation protocols used in the numerical experiments.

AI USE STATEMENT

OpenAI Codex was used primarily to review the manuscript text and code. It also assisted with routine formatting tasks, including mathematical typesetting and figure layout. The authors reviewed all AI-assisted material and take full responsibility for the final content of the paper.

REFERENCES

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

Marti J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001.

Nieves Atienza, Rocio Gonzalez-Diaz, and Matteo Rucco. Persistent entropy for separating topological features from noise in vietoris–rips complexes. *Journal of Intelligent Information Systems*, 52:637–655, 2019.

Nieves Atienza, Rocío González-Díaz, and Manuel Soriano-Trigueros. On the stability of persistent entropy and new summary functions for topological data analysis. *Pattern Recognition*, 107: 107509, 2020.

Hiroaki Bando, Shizuo Kaji, and Takaharu Yaguchi. Causal inference for empirical dynamical systems based on persistent homology. *JSIAM Letters*, 14:69–72, 2022.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(3):77–102, 2015.

Liangyue Cao, Alistair Mees, and Kevin Judd. Dynamics from multivariate time series. *Physica D: Nonlinear Phenomena*, 121(1–2):75–88, 1998.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

Kit C. Chan, Umar Islambekov, Alexey Luchinsky, and Rebecca Sanders. A computationally efficient framework for vector representation of persistence diagrams. *Journal of Machine Learning Research*, 23(268):1–33, 2022.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. *Journal of Computational Geometry*, 6(2):140–161, Mar 2015.

Harish Chintakunta, Theophilos Gentimis, Rocio Gonzalez-Diaz, Maria-Jose Jimenez, and Hamid Krim. An entropy-based persistence barcode. *Pattern Recognition*, 48(2):391–401, 2015.

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

Thibault de Surrel, Felix Hensel, Mathieu Carrière, Théo Lacombe, Yuichi Ike, Hiroaki Kurihara, Marc Glisse, and Frédéric Chazal. RipsNet: A general architecture for fast and robust estimation of the persistent homology of point clouds. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, volume 196 of *Proceedings of Machine Learning Research*, pp. 96–106. PMLR, 2022.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

Megan Johnson and Jae-Hun Jung. Instability of the betti sequence for persistent homology and a stabilized version of the betti sequence. *Journal of the Korean Society for Industrial and Applied Mathematics*, 25(4):296–311, 2021.

Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological characteristics help in knowledge distillation? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24674–24693. PMLR, 2024.

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

Genki Kusano, Yasuaki Hiraoka, and Kenji Fukumizu. Persistence weighted Gaussian kernel for topological data analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2004–2013. PMLR, 2016.

Christian D. Madsen, Agnese Barbensi, Stephen Y. Zhang, Lucy Ham, Alessia David, Douglas E. V. Pires, and Michael P. H. Stumpf. The topological properties of the protein universe. *Nature Communications*, 16:7503, 2025.

Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The GUDHI library: Simplicial complexes and persistent homology. In *Mathematical Software – ICMS 2014*, volume 8592 of *Lecture Notes in Computer Science*, pp. 167–174. Springer, 2014.

Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.

Jose A. Perea, Elizabeth Munch, and Firas A. Khasawneh. Approximating continuous functions on persistence diagrams using template functions. *Foundations of Computational Mathematics*, 23 (4):1215–1272, 2023.

Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4741–4748, 2015.

Sarah Samorodnitsky, Katie Campbell, Amarise Little, Wodan Ling, Ni Zhao, Yen-Chi Chen, and Michael C. Wu. Detecting clinically relevant topological structures in multiplexed spatial proteomics using TopKAT. *Patterns*, 7(1):101456, 2026.

Suryaka Suresh and Vinayak Abrol. Evaluating generative models via cubical homology based persistent entropy. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 2768–2777. ACM, 2025.

Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young (eds.), *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, pp. 366–381. Springer, Berlin, Heidelberg, 1981.

Sashiel Vagus, Theresa M. Casey, and Uduak Z. George. Persistent entropy links irregularities in daily and weekly rest and activity cycles during gestation week 22 and 32 to maternal and neonate health outcomes: A prospective cohort study. *PLOS One*, 21(3):e0342509, 2026.

A Proofs for Probabilities Induced on the Augmented Event Space

A.1 Regularity assumptions on g

The uniform stability results use a function $g : [0, \infty) \to [0, \infty)$ satisfying

$$g(0) = 0, \qquad \text{and} \qquad 0 \le g(t) \le M_g \quad \text{for every } t \ge 0,$$

for some finite constant M_g. We further assume that g is monotonically increasing and globally Lipschitz: there exists $L_g < \infty$ such that

$$|g(s) - g(t)| \le L_g |s - t| \qquad \text{for every } s, t \ge 0.$$

The condition $g(0) = 0$ makes a point contribute zero when it is matched to the diagonal. Boundedness and global Lipschitz continuity make the constants in the stability estimates independent of the diagrams and their maximum persistence. Together with the constant L_0 from Definition 1, we define

$$L_\phi = L_g + M_g L_0.$$

The function used in the experiments satisfies these assumptions:

$$g(t) = \frac{t}{1+t}.$$

Proposition 3. *Fix $\sigma_b, \sigma_\ell, \tau > 0$, and let $k = \kappa$ be the Gaussian similarity in Section 3.1. Let $g : [0, \infty) \to [0, \infty)$ be an unbounded function with $g(0) = 0$. For this proposition, use the formulas in Definitions 2 and 3 with this g, without the boundedness assumption. Then there exist nonempty finite persistence diagrams D_{X_n}, D_{Y_n} such that*

$$W_1(D_{X_n}, D_{Y_n}) \longrightarrow 0, \qquad \textit{and} \qquad p_{X_n}^{Y_n}(\partial) = 1 - e^{-\frac{1}{2}} > 0 \quad \textit{for every } n.$$

Consequently, neither of the bounds

$$p_X^Y(\partial) \le C W_1(D_X, D_Y)^2, \qquad \textit{and} \qquad d_{\mathrm{TV}}(p_X^Y, p_X^{Y'}) \le C W_1(D_Y, D_{Y'})$$

can hold for all nonempty finite persistence diagrams with a diagram-independent constant $0 \le C < \infty$.

Proof. Since g is unbounded and $g(0) = 0$, choose $t_n > 0$ such that $a_n := g(t_n) > \tau + n$. Set

$$\varepsilon_n = \sigma_b \sqrt{-2 \log \left(1 - \frac{\tau}{a_n}\right)}.$$

Using this separation, define

$$D_{X_n} = \{u_n = (0, t_n)\}, \qquad \text{and} \qquad D_{Y_n} = \{v_n = (\varepsilon_n, t_n)\}.$$

We choose this birth separation so that $a_n(e^{-\frac{\varepsilon_n^2}{2\sigma_b^2}} - 1) = -\tau$. Since $0 < \frac{\tau}{a_n} < 1$ and $\frac{\tau}{a_n} \to 0$, ε_n is well defined and tends to zero. Matching u_n directly to v_n gives

$$W_1(D_{X_n}, D_{Y_n}) \le \|u_n - v_n\| = \varepsilon_n \longrightarrow 0.$$

Since $\ell_{u_n} = \ell_{v_n} = t_n$, the Gaussian similarity reduces to

$$k(u_n, v_n) = e^{-\frac{\varepsilon_n^2}{2\sigma_b^2}} = 1 - \frac{\tau}{a_n}.$$

Using $k(u_n, u_n) = 1$ and $g(t_n) = a_n$, we obtain

$$\alpha_{u_n}(D_{X_n}) = k(u_n, u_n) g(t_n) = a_n,$$

and

$$\alpha_{u_n}(D_{Y_n}) = k(u_n, v_n) g(t_n) = \left(1 - \frac{\tau}{a_n}\right) a_n = a_n - \tau.$$

Thus

$$\delta_{X_n}^{Y_n}(u_n) = (a_n - \tau) - a_n = -\tau, \qquad \text{and} \qquad w_{X_n}^{Y_n}(u_n) = e^{-\frac{\tau^2}{2\tau^2}} = e^{-\frac{1}{2}}.$$

Since D_{X_n} contains one point, $p_{X_n}(u_n) = 1$. Hence

$$p_{X_n}^{Y_n}(u_n) = w_{X_n}^{Y_n}(u_n) p_{X_n}(u_n) = e^{-\frac{1}{2}},$$

and

$$p_{X_n}^{Y_n}(\partial) = \big(1 - w_{X_n}^{Y_n}(u_n)\big) p_{X_n}(u_n) = 1 - e^{-\frac{1}{2}} > 0.$$

If the first bound held with a diagram-independent constant $C \geq 0$, then

$$0 < 1 - e^{-\frac{1}{2}} \leq C W_1(D_{X_n}, D_{Y_n})^2 \leq C\varepsilon_n^2 \longrightarrow 0,$$

which is a contradiction.

For the second bound, set $D_{Y_n'} = D_{X_n}$, so that $p_{X_n}^{X_n} = p_{X_n}$. On the common event space $\{u_n, \partial\}$,

$$\begin{aligned} d_{\mathrm{TV}}(p_{X_n}^{Y_n}, p_{X_n}^{X_n}) &= \frac{1}{2}\left(\left|e^{-\frac{1}{2}} - 1\right| + \left|1 - e^{-\frac{1}{2}} - 0\right|\right) \\ &= 1 - e^{-\frac{1}{2}}. \end{aligned}$$

The second bound would therefore imply

$$0 < 1 - e^{-\frac{1}{2}} \leq C W_1(D_{Y_n}, D_{X_n}) \leq C\varepsilon_n \longrightarrow 0,$$

which is again a contradiction. □

The proof requires neither monotonicity nor Lipschitz continuity of g; it therefore also applies when these conditions are retained. The conclusion concerns constants common to all diagrams and does not rule out stability estimates whose constants depend on D_X.

A.2 Maximum-entropy derivation of the Gaussian similarity function

The Gaussian uniquely maximizes differential entropy among densities with a fixed mean and covariance (Cover & Thomas, 2006). Set $\mathbb{E}_q[U] = u$ and $\Sigma = \mathrm{diag}(\sigma_b^2, \sigma_\ell^2)$. The maximizer is

$$q_u(v) = \frac{1}{2\pi\sigma_b\sigma_\ell} e^{-\frac{1}{2}(v-u)^\top \Sigma^{-1}(v-u)}.$$

For $u = (b_u, \ell_u)$ and $v = (b_v, \ell_v)$, the quadratic form is

$$(v-u)^\top \Sigma^{-1}(v-u) = \left(\frac{b_v - b_u}{\sigma_b}\right)^2 + \left(\frac{\ell_v - \ell_u}{\sigma_\ell}\right)^2.$$

The density is largest at $v = u$, where

$$q_u(u) = \frac{1}{2\pi\sigma_b\sigma_\ell}.$$

Normalizing the density by this maximum removes the dimensional prefactor and gives

$$\begin{aligned} \frac{q_u(v)}{q_u(u)} &= e^{-\frac{1}{2}(v-u)^\top \Sigma^{-1}(v-u)} \\ &= e^{-\frac{1}{2}\left\{\left(\frac{b_v - b_u}{\sigma_b}\right)^2 + \left(\frac{\ell_v - \ell_u}{\sigma_\ell}\right)^2\right\}} \\ &= \kappa(u, v). \end{aligned}$$

Let

$$\rho(u, v)^2 = \left(\frac{b_u - b_v}{\sigma_b}\right)^2 + \left(\frac{\ell_u - \ell_v}{\sigma_\ell}\right)^2.$$

This is the Euclidean metric after coordinate rescaling, and $\kappa(u,v) = h(\rho(u,v))$ for $h(r) = e^{-\frac{r^2}{2}}$, which has all properties required in Definition 1. In particular, $h(0) = 1$, h is strictly decreasing on $(0,\infty)$, and $h(r)$ approaches zero as r approaches infinity.

It remains to verify the Lipschitz condition. Differentiating with respect to the second argument gives

$$\frac{\partial \kappa(u,v)}{\partial b_v} = -\frac{b_v - b_u}{\sigma_b^2}\kappa(u,v), \qquad \text{and} \qquad \frac{\partial \kappa(u,v)}{\partial \ell_v} = -\frac{\ell_v - \ell_u}{\sigma_\ell^2}\kappa(u,v).$$

Let $\sigma_{\min} = \min\{\sigma_b, \sigma_\ell\}$. Then

$$\|\nabla_v \kappa(u,v)\| \le \frac{\rho(u,v)e^{-\frac{\rho(u,v)^2}{2}}}{\sigma_{\min}} \le \frac{1}{\sigma_{\min}},$$

and therefore

$$\begin{aligned}|\kappa(u,v) - \kappa(u,v')| &\le \sup_{t\in[0,1]} \left\|\nabla_v \kappa\big(u,(1-t)v + tv'\big)\right\| \|v - v'\| \\ &\le \frac{1}{\sigma_{\min}}\|v - v'\|.\end{aligned}$$

A.3 Verification of Definition 3

For any finite off-diagonal multiset D, let $\mathrm{supp}(D)$ denote the set of its distinct coordinates. If $u \in \mathrm{supp}(D)$ has multiplicity m_u, we label its occurrences by

$$u_1, \dots, u_{m_u}.$$

We henceforth suppress these labels, with every sum over D understood to count multiplicity. For $u \in D_X$, define

$$p_X(u) = \frac{\ell_u}{\sum_{v\in D_X} \ell_v}, \qquad \text{and} \qquad p_X^Y(u) = w_X^Y(u) p_X(u).$$

The unexplained-event mass is correspondingly

$$p_X^Y(\partial) = \sum_{u\in D_X} \big(1 - w_X^Y(u)\big)\, p_X(u), \qquad \text{and} \qquad p_X(\partial) = 0.$$

Define the finite event space and its sigma-algebra by

$$\Omega_X = D_X \cup \{\partial\}, \qquad \text{and} \qquad \mathcal{F}_X = 2^{\Omega_X}.$$

First, the occurrence-level persistence masses sum to one:

$$\begin{aligned}\sum_{u\in D_X} p_X(u) &= \frac{\sum_{u\in D_X} \ell_u}{\sum_{v\in D_X} \ell_v} \\ &= 1.\end{aligned}$$

Every $p_X(u)$ is positive, and $0 < w_X^Y(u) \le 1$. Therefore all masses assigned by p_X^Y are nonnegative. Their total is

$$\begin{aligned}&\sum_{u\in D_X} p_X^Y(u) + p_X^Y(\partial) \\ &= \sum_{u\in D_X} w_X^Y(u) p_X(u) + \sum_{u\in D_X} \big(1 - w_X^Y(u)\big)\, p_X(u) \\ &= \sum_{u\in D_X} p_X(u) \\ &= 1.\end{aligned}$$

Thus the atom masses extend by finite sums to a probability measure p_X^Y on $(\Omega_X, \mathcal{F}_X)$. Moreover,

$$\sum_{\omega\in\Omega_X} p_X(\omega) = \sum_{u\in D_X} p_X(u) + p_X(\partial) = 1,$$

so p_X is also a probability measure on this space.

A.4 1-WASSERSTEIN DISTANCE

For finite persistence diagrams D and D', we use the diagram 1-Wasserstein distance with the Euclidean ground metric in the birth–persistence coordinates fixed in Section 2 (Mileyko et al., 2011). Following the standard convention, the diagonal Δ is included with infinite multiplicity. Let $\Gamma(D, D')$ be the set of bijections between $D \cup \Delta$ and $D' \cup \Delta$ that have only finitely many nonzero matching costs. Then

$$W_1(D, D') = \inf_{\gamma \in \Gamma(D,D')} \sum_{z \in D \cup \Delta} \|z - \gamma(z)\|.$$

The bijections match occurrences, so repeated points remain distinct. Matching an off-diagonal point to its nearest point in Δ gives its distance to the diagonal and therefore includes unmatched points in the same sum. In birth–persistence coordinates, the diagonal is $\{(t, 0) : t \in \mathbb{R}\}$. The linear change of coordinates $(b, d) \mapsto (b, d - b)$ is invertible, so this distance is equivalent to the usual birth–death version up to fixed multiplicative constants.

A.5 PROOF OF PROPOSITION 1

We prove the three claims in order. As in Appendix A.3, repeated points are treated as distinct atoms of D_X, while occurrence labels are suppressed.

Proof of (i). For every $u \in D_X$,

$$\delta_X^X(u) = \alpha_u(D_X) - \alpha_u(D_X) = 0,$$

and so

$$w_X^X(u) = e^{-\frac{\delta_X^X(u)^2}{2\tau^2}} = 1,$$

and therefore

$$p_X^X(u) = w_X^X(u) p_X(u) = p_X(u),$$

and

$$\begin{aligned} p_X^X(\partial) &= \sum_{v \in D_X} \left(1 - w_X^X(v)\right) p_X(v) \\ &= 0 \\ &= p_X(\partial). \end{aligned}$$

Thus $p_X^X = p_X$ on the entire event space $D_X \cup \{\partial\}$.

Proof of (ii). Suppose first that $\delta_X^Y = 0$ on D_X. Then

$$w_X^Y(u) = e^{-\frac{\delta_X^Y(u)^2}{2\tau^2}} = 1,$$

and so

$$p_X^Y(u) = w_X^Y(u) p_X(u) = p_X(u), \qquad u \in D_X,$$

and

$$p_X^Y(\partial) = \sum_{v \in D_X} \left(1 - w_X^Y(v)\right) p_X(v) = 0 = p_X(\partial).$$

Hence $p_X^Y = p_X$.

Conversely, suppose that $p_X^Y = p_X$. Every off-diagonal persistence point has positive persistence, so $p_X(u) > 0$ for every $u \in D_X$.

$$1 = \frac{p_X^Y(u)}{p_X(u)} = w_X^Y(u) = e^{-\frac{\delta_X^Y(u)^2}{2\tau^2}},$$

and therefore

$$0 = -\log w_X^Y(u) = \frac{\delta_X^Y(u)^2}{2\tau^2}, \qquad u \in D_X.$$

Thus $\delta_X^Y = 0$ on D_X.

Proof of (iii). For probability measures p and q on $\Omega_X = D_X \cup \{\partial\}$, the total variation distance is

$$d_{\mathrm{TV}}(p, q) = \sup_{A \subseteq \Omega_X} |p(A) - q(A)|.$$

For $B \subseteq D_X$, denote the unexplained mass contributed by B by

$$0 \le m(B) := \sum_{u \in B} \big(1 - w_X^Y(u)\big)\, p_X(u) \le m(D_X) = p_X^Y(\partial).$$

Case 1: $\partial \notin A$. Here $A \subseteq D_X$, and

$$\begin{aligned} |p_X^Y(A) - p_X(A)| &= \left| \sum_{u \in A} \big(w_X^Y(u) - 1\big)\, p_X(u) \right| \\ &= |-m(A)| \\ &= m(A) \\ &\le p_X^Y(\partial). \end{aligned}$$

Case 2: $\partial \in A$. Write $A = B \cup \{\partial\}$ with $B \subseteq D_X$. Since $p_X(\partial) = 0$,

$$\begin{aligned} |p_X^Y(A) - p_X(A)| &= \big|p_X^Y(\partial) + p_X^Y(B) - p_X(B)\big| \\ &= \big|p_X^Y(\partial) - m(B)\big| \\ &= p_X^Y(\partial) - m(B) \\ &\le p_X^Y(\partial). \end{aligned}$$

The two cases give

$$\sup_{A \subseteq \Omega_X} |p_X^Y(A) - p_X(A)| \le p_X^Y(\partial),$$

and

$$\begin{aligned} \sup_{A \subseteq \Omega_X} |p_X^Y(A) - p_X(A)| &\ge |p_X^Y(\{\partial\}) - p_X(\{\partial\})| \\ &= |p_X^Y(\partial) - 0| \\ &= p_X^Y(\partial), \end{aligned}$$

and therefore

$$d_{\mathrm{TV}}(p_X^Y, p_X) = p_X^Y(\partial).$$

A.6 Stability lemmas

For the stability proofs, let D be any finite persistence diagram and define

$$\alpha_u(D) = \sum_{v \in D} \phi_u(v).$$

In birth–persistence coordinates, every $v \in \Delta$ has $\ell_v = 0$. For the proofs, the same formula $\delta_X^Y(u) = \alpha_u(D_Y) - \alpha_u(D_X)$ is used for every $u \in \mathbb{R}^2$.

Lemma 1 (Lipschitz continuity of persistence)**.** *Let z and z' be persistence points, possibly on the diagonal. Under the Euclidean norm:*

(i) In birth–persistence coordinates,

$$|\ell_z - \ell_{z'}| \le \|z - z'\|.$$

(ii) In birth–death coordinates,

$$|\ell_z - \ell_{z'}| \le \sqrt{2}\,\|z - z'\|.$$

Proof. *(i)* In birth–persistence coordinates, write $z = (b_z, \ell_z)$ and $z' = (b_{z'}, \ell_{z'})$. Then

$$\begin{aligned} |\ell_z - \ell_{z'}| &\le \sqrt{(b_z - b_{z'})^2 + (\ell_z - \ell_{z'})^2} \\ &= \|z - z'\|. \end{aligned}$$

(ii) In birth–death coordinates, write $z = (b_z, d_z)$ and $z' = (b_{z'}, d_{z'})$. Since $\ell_z = d_z - b_z$ and $\ell_{z'} = d_{z'} - b_{z'}$,

$$\begin{aligned} |\ell_z - \ell_{z'}| &= |(d_z - d_{z'}) - (b_z - b_{z'})| \\ &\le \sqrt{2}\sqrt{(b_z - b_{z'})^2 + (d_z - d_{z'})^2} \\ &= \sqrt{2}\,\|z - z'\|. \end{aligned}$$

□

Lemma 2. *Let D_X, D_Y be finite persistence diagrams. If*

$$L_\phi = L_g + M_g L_0,$$

then, for every $u \in \mathbb{R}^2$,

$$|\alpha_u(D_X) - \alpha_u(D_Y)| \le L_\phi W_1(D_X, D_Y).$$

Proof. Fix $u \in \mathbb{R}^2$ and a matching $\gamma \in \Gamma(D_X, D_Y)$. Because $g(0) = 0$, every diagonal point $v \in \Delta$ satisfies

$$\phi_u(v) = k(u, v)g(\ell_v) = 0.$$

We may therefore add the matched diagonal occurrences to the sums defining $\alpha_u(D_X)$ and $\alpha_u(D_Y)$ without changing either value. Pairs in which both points lie on the diagonal contribute zero and may be omitted.

Consider any remaining matched pair v and $v' = \gamma(v)$. Using $0 < k \le 1$, $g \le M_g$, the Lipschitz assumptions on g and k, Lemma 1, and adding and subtracting $k(u, v)g(\ell_{v'})$, we obtain

$$\begin{aligned} |\phi_u(v) - \phi_u(v')| &= |k(u, v)g(\ell_v) - k(u, v')g(\ell_{v'})| \\ &\le k(u, v)|g(\ell_v) - g(\ell_{v'})| + g(\ell_{v'})|k(u, v) - k(u, v')| \\ &\le L_g|\ell_v - \ell_{v'}| + M_g L_0\|v - v'\| \\ &\le (L_g + M_g L_0)\,\|v - v'\| \\ &= L_\phi\|v - v'\|. \end{aligned}$$

Consequently,

$$\begin{aligned} |\alpha_u(D_X) - \alpha_u(D_Y)| &= \left|\sum_{v \in D_X \cup \Delta} (\phi_u(v) - \phi_u(\gamma(v)))\right| \\ &\le \sum_{v \in D_X \cup \Delta} |\phi_u(v) - \phi_u(\gamma(v))| \\ &\le L_\phi \sum_{v \in D_X \cup \Delta} \|v - \gamma(v)\|, \qquad \forall\, \gamma \in \Gamma(D_X, D_Y), \end{aligned}$$

and therefore

$$\begin{aligned} |\alpha_u(D_X) - \alpha_u(D_Y)| &\le \inf_{\gamma \in \Gamma(D_X, D_Y)} L_\phi \sum_{v \in D_X \cup \Delta} \|v - \gamma(v)\| \\ &= L_\phi W_1(D_X, D_Y). \end{aligned}$$

□

Lemma 3. *We have*

$$\|\delta_X^Y\|_\infty \le L_\phi W_1(D_X, D_Y).$$

Proof. By Definition 2 and Lemma 2,

$$\begin{aligned}\|\delta_X^Y\|_\infty &= \sup_{u\in D_X} |\delta_X^Y(u)| \\ &= \sup_{u\in D_X} |\alpha_u(D_Y) - \alpha_u(D_X)| \\ &\le \sup_{u\in D_X} L_\phi W_1(D_X, D_Y) \\ &= L_\phi W_1(D_X, D_Y).\end{aligned}$$

□

A.7 Proof of Theorem 1

For the first claim, set $R := L_\phi W_1(D_X, D_Y)$. By Lemma 3, $|\delta_X^Y(u)| \le R$ for every $u \in D_X$. Using $1 - e^{-t} \le t$ for $t \ge 0$,

$$\begin{aligned}d_{\mathrm{TV}}(p_X^Y, p_X) &= p_X^Y(\partial) \\ &= \sum_{u\in D_X} p_X(u)\left(1 - e^{-\frac{\delta_X^Y(u)^2}{2\tau^2}}\right) \\ &\le \sum_{u\in D_X} p_X(u)\frac{\delta_X^Y(u)^2}{2\tau^2} \\ &\le \sum_{u\in D_X} p_X(u)\frac{R^2}{2\tau^2} \\ &= \frac{R^2}{2\tau^2}\sum_{u\in D_X} p_X(u) \\ &= \frac{R^2}{2\tau^2} \\ &= \frac{L_\phi^2}{2\tau^2} W_1(D_X, D_Y)^2.\end{aligned}$$

For the second claim, fix D_X and define

$$f(a) := e^{-\frac{a^2}{2\tau^2}},$$

and differentiation gives

$$f'(a) = -\frac{a}{\tau^2} e^{-\frac{a^2}{2\tau^2}}.$$

For $q(x) := xe^{-\frac{x^2}{2}}$ on $[0, \infty)$,

$$q'(x) = (1 - x^2)e^{-\frac{x^2}{2}}.$$

Thus q attains its maximum at $x = 1$, and therefore

$$\sup_{x\ge 0} q(x) = q(1) = e^{-\frac{1}{2}}.$$

Hence

$$\sup_{a\in\mathbb{R}} |f'(a)| = \frac{1}{\tau}\sup_{x\ge 0} q(x) = \frac{1}{\tau\sqrt{e}}.$$

Moreover, Lemma 2 gives, for every $u \in \mathbb{R}^2$,

$$\begin{aligned}|\delta_X^Y(u) - \delta_X^{Y'}(u)| &= |\alpha_u(D_Y) - \alpha_u(D_X) - \alpha_u(D_{Y'}) + \alpha_u(D_X)| \\ &= |\alpha_u(D_Y) - \alpha_u(D_{Y'})| \\ &\le L_\phi W_1(D_Y, D_{Y'}).\end{aligned}$$

The mean value theorem applied to f gives

$$\begin{aligned}|w_X^Y(u) - w_X^{Y'}(u)| &= |f(\delta_X^Y(u)) - f(\delta_X^{Y'}(u))| \\ &\le \frac{1}{\tau\sqrt{e}} |\delta_X^Y(u) - \delta_X^{Y'}(u)| \\ &\le \frac{L_\phi}{\tau\sqrt{e}} W_1(D_Y, D_{Y'}).\end{aligned}$$

Therefore, for every $u \in D_X$,

$$\begin{aligned}|p_X^Y(u) - p_X^{Y'}(u)| &= p_X(u)|w_X^Y(u) - w_X^{Y'}(u)| \\ &\le p_X(u) \frac{L_\phi}{\tau\sqrt{e}} W_1(D_Y, D_{Y'}).\end{aligned}$$

Hence

$$\begin{aligned}S &:= \sum_{u \in D_X} |p_X^Y(u) - p_X^{Y'}(u)| \\ &\le \frac{L_\phi}{\tau\sqrt{e}} W_1(D_Y, D_{Y'}) \sum_{u \in D_X} p_X(u) \\ &= \frac{L_\phi}{\tau\sqrt{e}} W_1(D_Y, D_{Y'}).\end{aligned}$$

The difference between the two boundary masses satisfies

$$\begin{aligned}|p_X^Y(\partial) - p_X^{Y'}(\partial)| &= \left| \sum_{u \in D_X} p_X(u) \left(w_X^{Y'}(u) - w_X^Y(u) \right) \right| \\ &\le \sum_{u \in D_X} p_X(u) |w_X^{Y'}(u) - w_X^Y(u)| \\ &= S.\end{aligned}$$

For probability measures on a finite event space, total variation is one half of their ℓ^1 difference. Since both measures are defined on Ω_X, we obtain

$$\begin{aligned}d_{\mathrm{TV}}(p_X^Y, p_X^{Y'}) &= \frac{1}{2} \left(S + |p_X^Y(\partial) - p_X^{Y'}(\partial)| \right) \\ &\le S \\ &\le \frac{L_\phi}{\tau\sqrt{e}} W_1(D_Y, D_{Y'}).\end{aligned}$$

Thus both claims hold with the diagram-independent constant

$$C = \max \left\{ \frac{L_\phi^2}{2\tau^2}, \frac{L_\phi}{\tau\sqrt{e}} \right\}.$$

A.8 Proof of Proposition 2

For every $u \in D_X$, Definition 3 and the positivity of $p_X(u)$ give

$$\begin{aligned}\log \frac{p_X(u)}{p_X^Y(u)} &= \log \frac{1}{w_X^Y(u)} \\ &= -\log w_X^Y(u) \\ &= \frac{\delta_X^Y(u)^2}{2\tau^2}.\end{aligned}$$

The KL divergence on Ω_X is

$$\mathrm{KL}(p_X \| p_X^Y) = \sum_{\omega \in \Omega_X} p_X(\omega) \log \frac{p_X(\omega)}{p_X^Y(\omega)}.$$

The boundary atom contributes zero because $p_X(\partial) = 0$, using the standard convention that $0 \log \left(\frac{0}{q}\right) = 0$. Thus

$$\begin{aligned}
\mathrm{KL}(p_X \| p_X^Y) &= \sum_{u \in D_X} p_X(u) \log \frac{p_X(u)}{p_X^Y(u)} \\
&= \sum_{u \in D_X} p_X(u) \frac{\delta_X^Y(u)^2}{2\tau^2} \\
&= \frac{1}{2\tau^2} \mathbb{E}_{u \sim p_X} \left[\delta_X^Y(u)^2\right].
\end{aligned}$$

On the other hand, expanding the logarithm of the ratio gives

$$\begin{aligned}
\mathrm{KL}(p_X \| p_X^Y) &= \sum_{u \in D_X} p_X(u) \log p_X(u) - \sum_{u \in D_X} p_X(u) \log p_X^Y(u) \\
&= -H(p_X) + H(p_X, p_X^Y) \\
&= H(p_X, p_X^Y) - H(p_X).
\end{aligned}$$

Therefore,

$$H(p_X, p_X^Y) - H(p_X) = \mathrm{KL}(p_X \| p_X^Y) = \frac{1}{2\tau^2} \mathbb{E}_{u \sim p_X} \left[\delta_X^Y(u)^2\right] \geq 0.$$

A.9 Proof of Theorem 2

Proposition 2 and Lemma 3 give

$$\begin{aligned}
0 \leq \Delta H_X^Y &= \frac{1}{2\tau^2} \sum_{u \in D_X} p_X(u) \delta_X^Y(u)^2 \\
&\leq \frac{1}{2\tau^2} \sum_{u \in D_X} p_X(u) \|\delta_X^Y\|_\infty^2 \\
&= \frac{1}{2\tau^2} \|\delta_X^Y\|_\infty^2 \sum_{u \in D_X} p_X(u) \\
&= \frac{1}{2\tau^2} \|\delta_X^Y\|_\infty^2 \\
&\leq \frac{1}{2\tau^2} \left(L_\phi W_1(D_X, D_Y)\right)^2 \\
&= \frac{L_\phi^2}{2\tau^2} W_1(D_X, D_Y)^2.
\end{aligned}$$

Since

$$\frac{L_\phi^2}{2\tau^2} \leq C = \max\left\{\frac{L_\phi^2}{2\tau^2}, \frac{L_\phi}{\tau\sqrt{e}}\right\},$$

the same diagram-independent constant C as in Theorem 1 gives the claimed bound.

A.10 Proof of Theorem 3

For a finite persistence diagram D, define

$$T(D) = \sum_{z \in D} \ell_z, \qquad \text{and} \qquad G(D) = \sum_{z \in D} g(\ell_z).$$

Set

$$T_* := \min\{T(D_X), T(D_{X'})\},$$
$$B_* := \max\left\{\|\delta_X^Y\|_\infty, \|\delta_{X'}^{Y'}\|_\infty\right\},$$
$$G_* := \max\left\{G(D_X) + G(D_Y), G(D_{X'}) + G(D_{Y'})\right\}.$$

These quantities are finite, and $T_* > 0$ because the diagrams are nonempty and contain only off-diagonal points.

We first control the change in δ when its evaluation point also moves. By Definition 1,

$$k(u, v) = h(\rho(u, v)) = h(\rho(v, u)) = k(v, u),$$

and therefore

$$|k(u, v) - k(u', v)| = |k(v, u) - k(v, u')| \le L_0\|u - u'\|.$$

For any $z \in D$,

$$\begin{aligned}|\phi_u(z) - \phi_{u'}(z)| &= g(\ell_z)|k(u, z) - k(u', z)| \\ &\le L_0 g(\ell_z)\|u - u'\|,\end{aligned}$$

and summing over the diagram gives

$$\begin{aligned}|\alpha_u(D) - \alpha_{u'}(D)| &= \left|\sum_{z\in D}(\phi_u(z) - \phi_{u'}(z))\right| \\ &\le \sum_{z\in D}|\phi_u(z) - \phi_{u'}(z)| \\ &\le L_0 G(D)\|u - u'\|.\end{aligned}$$

We use this pointwise bound together with Lemma 2:

$$\begin{aligned}|\alpha_u(D_Y) - \alpha_{u'}(D_{Y'})| &\le |\alpha_u(D_Y) - \alpha_u(D_{Y'})| + |\alpha_u(D_{Y'}) - \alpha_{u'}(D_{Y'})| \\ &\le L_\phi W_1(D_Y, D_{Y'}) + L_0 G(D_{Y'})\|u - u'\|,\end{aligned}$$

and similarly

$$|\alpha_u(D_X) - \alpha_{u'}(D_{X'})| \le L_\phi W_1(D_X, D_{X'}) + L_0 G(D_{X'})\|u - u'\|.$$

Since $\delta_X^Y(u) = \alpha_u(D_Y) - \alpha_u(D_X)$, the triangle inequality now yields

$$\begin{aligned}|\delta_X^Y(u) - \delta_{X'}^{Y'}(u')| &\le L_\phi\left(W_1(D_X, D_{X'}) + W_1(D_Y, D_{Y'})\right) \\ &\quad + L_0\left(G(D_{X'}) + G(D_{Y'})\right)\|u - u'\| \\ &\le L_\phi\left(W_1(D_X, D_{X'}) + W_1(D_Y, D_{Y'})\right) + L_0 G_*\|u - u'\|.\end{aligned}$$

Take optimal matchings between D_X and $D_{X'}$, and between D_Y and $D_{Y'}$, allowing points to be matched to the diagonal. Let $\mathcal{M}_X$ be the finite set of active pairs (u, u') in the first matching, and assign persistence zero to a diagonal point. For brevity, $\sum_{\mathcal{M}_X}$ denotes a sum over all $(u, u') \in \mathcal{M}_X$. Let

$$R_X = W_1(D_X, D_{X'}), \qquad \text{and} \qquad R_Y = W_1(D_Y, D_{Y'}),$$

and, for $(u, u') \in \mathcal{M}_X$, set

$$a(u) = \frac{\ell_u}{T(D_X)}, \qquad \text{and} \qquad a'(u') = \frac{\ell_{u'}}{T(D_{X'})}.$$

By Lemma 1, including for pairs with one point on the diagonal,

$$|\ell_u - \ell_{u'}| \le \|u - u'\|, \qquad (u, u') \in \mathcal{M}_X,$$

and summing over the matching gives

$$\sum_{\mathcal{M}_X}|\ell_u - \ell_{u'}| \le \sum_{\mathcal{M}_X}\|u - u'\| = R_X,$$

and therefore

$$\begin{aligned}|T(D_X) - T(D_{X'})| &= \left|\sum_{\mathcal{M}_X} (\ell_u - \ell_{u'})\right| \\ &\le \sum_{\mathcal{M}_X} |\ell_u - \ell_{u'}| \\ &\le R_X.\end{aligned}$$

We next bound the change in the normalized persistence weights. For each pair,

$$\begin{aligned}|a(u) - a'(u')| &= \left|\frac{\ell_u}{T(D_X)} - \frac{\ell_{u'}}{T(D_{X'})}\right| \\ &\le \frac{|\ell_u - \ell_{u'}|}{T(D_X)} + \ell_{u'}\left|\frac{1}{T(D_X)} - \frac{1}{T(D_{X'})}\right|,\end{aligned}$$

and summing over all matched pairs gives

$$\begin{aligned}\sum_{(u,u')\in\mathcal{M}_X} |a(u) - a'(u')| &\le \frac{1}{T(D_X)} \sum_{(u,u')\in\mathcal{M}_X} |\ell_u - \ell_{u'}| + \left|\frac{1}{T(D_X)} - \frac{1}{T(D_{X'})}\right| \sum_{(u,u')\in\mathcal{M}_X} \ell_{u'} \\ &= \frac{1}{T(D_X)} \sum_{(u,u')\in\mathcal{M}_X} |\ell_u - \ell_{u'}| + \frac{|T(D_X) - T(D_{X'})|}{T(D_X)T(D_{X'})} T(D_{X'}) \\ &= \frac{1}{T(D_X)} \sum_{(u,u')\in\mathcal{M}_X} |\ell_u - \ell_{u'}| + \frac{|T(D_X) - T(D_{X'})|}{T(D_X)} \\ &\le \frac{2R_X}{T(D_X)} \\ &\le \frac{2R_X}{T_*}.\end{aligned}$$

In particular,

$$\sum_{(u,u')\in\mathcal{M}_X} a(u) = \sum_{(u,u')\in\mathcal{M}_X} a'(u') = 1.$$

Set

$$r(u) := \begin{cases}\delta_X^Y(u)^2, & u \notin \Delta, \\ 0, & u \in \Delta,\end{cases} \qquad r'(u') := \begin{cases}\delta_{X'}^{Y'}(u')^2, & u' \notin \Delta, \\ 0, & u' \in \Delta.\end{cases}$$

The diagonal values do not change either expectation because the corresponding normalized persistence weights are zero. Moreover,

$$0 \le r(u) \le B_*^2, \qquad \text{and} \qquad 0 \le r'(u') \le B_*^2.$$

Consider a pair (u, u') for which both points are off the diagonal. Factoring the difference of squares gives

$$\begin{aligned}|r(u) - r'(u')| &= \left|\delta_X^Y(u)^2 - \delta_{X'}^{Y'}(u')^2\right| \\ &= \left|\delta_X^Y(u) - \delta_{X'}^{Y'}(u')\right| \left|\delta_X^Y(u) + \delta_{X'}^{Y'}(u')\right| \\ &\le 2B_* \left|\delta_X^Y(u) - \delta_{X'}^{Y'}(u')\right| \\ &\le 2B_* L_\phi (R_X + R_Y) + 2B_* L_0 G_* \|u - u'\|.\end{aligned}$$

For a fixed pair, set $c = c(u, u') := \min\{a(u), a'(u')\}$. Then

$$\begin{aligned}|a(u)r(u) - a'(u')r'(u')| &= |(a(u) - c)\, r(u) - (a'(u') - c)\, r'(u') + c\,(r(u) - r'(u'))| \\ &\le B_*^2\,(|a(u) - c| + |a'(u') - c|) + c|r(u) - r'(u')| \\ &= B_*^2 |a(u) - a'(u')| + c|r(u) - r'(u')|.\end{aligned}$$

For a pair containing a diagonal point, the corresponding normalized weight and hence c are zero. Thus, for every $(u, u') \in \mathcal{M}_X$,

$$c(u,u')|r(u) - r'(u')| \le c(u,u')\left(2B_*L_\phi(R_X + R_Y) + 2B_*L_0G_*\|u - u'\|\right).$$

Moreover, the definition of c gives

$$0 \le \sum_{\mathcal{M}_X} c(u,u') \le \sum_{\mathcal{M}_X} a(u) = 1,$$

and

$$0 \le \sum_{\mathcal{M}_X} c(u,u')\|u - u'\| \le \sum_{\mathcal{M}_X} \|u - u'\| = R_X.$$

Using these estimates, denote the difference between the two weighted sums by

$$\mathcal{E}_X := \left| \sum_{\mathcal{M}_X} a(u)r(u) - \sum_{\mathcal{M}_X} a'(u')r'(u') \right|,$$

and the preceding estimates yield

$$\begin{aligned}
\mathcal{E}_X &\le B_*^2 \sum_{\mathcal{M}_X} |a(u) - a'(u')| + \sum_{\mathcal{M}_X} c(u,u')|r(u) - r'(u')| \\
&\le B_*^2 \frac{2R_X}{T_*} + 2B_*L_\phi(R_X + R_Y) \sum_{\mathcal{M}_X} c(u,u') \\
&\quad + 2B_*L_0G_* \sum_{\mathcal{M}_X} c(u,u')\|u - u'\| \\
&\le B_*^2 \frac{2R_X}{T_*} + 2B_*L_\phi(R_X + R_Y) + 2B_*L_0G_*R_X \\
&= \left(\frac{2B_*^2}{T_*} + 2B_*(L_\phi + L_0G_*) \right) R_X + 2B_*L_\phi R_Y.
\end{aligned}$$

The matching includes each off-diagonal occurrence exactly once. Therefore,

$$\sum_{\mathcal{M}_X} a(u)r(u) = \mathbb{E}_{u \sim p_X}\left[\delta_X^Y(u)^2\right],$$

and

$$\sum_{\mathcal{M}_X} a'(u')r'(u') = \mathbb{E}_{u' \sim p_{X'}}\left[\delta_{X'}^{Y'}(u')^2\right].$$

Proposition 2 consequently gives

$$\begin{aligned}
|\Delta H_X^Y - \Delta H_{X'}^{Y'}| &= \frac{\mathcal{E}_X}{2\tau^2} \\
&\le C_X R_X + C_Y R_Y,
\end{aligned}$$

where

$$C_X = \frac{1}{\tau^2}\left(\frac{B_*^2}{T_*} + B_*(L_\phi + L_0G_*) \right), \qquad \text{and} \qquad C_Y = \frac{B_*L_\phi}{\tau^2}.$$

To make the constant uniform on a neighborhood of (D_X, D_Y), first note that $g(0) = 0$ allows us to extend $g(\ell)$ by zero to diagonal points. For any optimal matching $\gamma \in \Gamma(D, D')$, the Lipschitz condition on g and Lemma 1 give

$$\begin{aligned}
|G(D) - G(D')| &= \left| \sum_{z \in D \cup \Delta} \left(g(\ell_z) - g(\ell_{\gamma(z)}) \right) \right| \\
&\le \sum_{z \in D \cup \Delta} |g(\ell_z) - g(\ell_{\gamma(z)})| \\
&\le L_g \sum_{z \in D \cup \Delta} |\ell_z - \ell_{\gamma(z)}| \\
&\le L_g \sum_{z \in D \cup \Delta} \|z - \gamma(z)\| \\
&= L_g W_1(D, D').
\end{aligned}$$

Set $r_{X,Y} := \frac{T(D_X)}{2}$ and suppose that $R_X + R_Y \leq r_{X,Y}$. Then

$$R_X \leq R_X + R_Y \leq \frac{T(D_X)}{2},$$

and so

$$\begin{aligned} T(D_{X'}) &\geq T(D_X) - R_X \\ &\geq \frac{T(D_X)}{2}, \end{aligned}$$

and therefore

$$T_* = \min\{T(D_X), T(D_{X'})\} \geq \frac{T(D_X)}{2}.$$

Set

$$\overline{G}_{X,Y} := G(D_X) + G(D_Y) + L_g r_{X,Y}.$$

The Lipschitz estimate for G gives

$$\begin{aligned} G(D_{X'}) + G(D_{Y'}) &\leq G(D_X) + L_g R_X + G(D_Y) + L_g R_Y \\ &= G(D_X) + G(D_Y) + L_g(R_X + R_Y) \\ &\leq \overline{G}_{X,Y}, \end{aligned}$$

and therefore

$$\begin{aligned} G_* &= \max\left\{G(D_X) + G(D_Y), G(D_{X'}) + G(D_{Y'})\right\} \\ &\leq \overline{G}_{X,Y}. \end{aligned}$$

Finally, $0 < k(u,z) \leq 1$ implies

$$\begin{aligned} 0 \leq \alpha_u(D) &= \sum_{z \in D} k(u,z) g(\ell_z) \\ &\leq \sum_{z \in D} g(\ell_z) \\ &= G(D). \end{aligned}$$

Therefore, the triangle inequality in the definition of δ gives

$$\|\delta_X^Y\|_\infty \leq G(D_X) + G(D_Y),$$

and

$$\|\delta_{X'}^{Y'}\|_\infty \leq G(D_{X'}) + G(D_{Y'}).$$

By the definitions of B_* and G_*,

$$B_* \leq G_* \leq \overline{G}_{X,Y}.$$

Substituting these estimates and the lower bound for T_* into C_X and C_Y gives

$$\begin{aligned} C_X &\leq \frac{1}{\tau^2}\left(\frac{\overline{G}_{X,Y}^2}{\frac{T(D_X)}{2}} + \overline{G}_{X,Y}\left(L_\phi + L_0 \overline{G}_{X,Y}\right)\right) \\ &= \frac{1}{\tau^2}\left(\frac{2\overline{G}_{X,Y}^2}{T(D_X)} + \overline{G}_{X,Y}\left(L_\phi + L_0 \overline{G}_{X,Y}\right)\right) \\ &=: C_{X,Y}, \end{aligned}$$

and

$$\begin{aligned} C_Y &\leq \frac{\overline{G}_{X,Y} L_\phi}{\tau^2} \\ &\leq C_{X,Y}. \end{aligned}$$

The constant $C_{X,Y}$ depends only on the fixed center diagrams D_X, D_Y and on the fixed functions and scales. It does not depend on $D_{X'}$ or $D_{Y'}$ as long as $R_X + R_Y \le r_{X,Y}$. Hence

$$\begin{aligned}|\Delta H_X^Y - \Delta H_{X'}^{Y'}| &\le C_X R_X + C_Y R_Y \\ &\le C_{X,Y}(R_X + R_Y) \\ &= C_{X,Y}\left(W_1(D_X, D_{X'}) + W_1(D_Y, D_{Y'})\right).\end{aligned}$$

For $D_{X'} = D_X$, $R_X = 0$, and hence

$$\begin{aligned}|\Delta H_X^Y - \Delta H_X^{Y'}| &\le C_{X,Y}(0 + R_Y) \\ &= C_{X,Y} W_1(D_Y, D_{Y'}).\end{aligned}$$

B Additional Experimental Details

For every $u \in D_X$, the numerical implementation evaluates

$$\alpha_u(D_Y) = \sum_{v \in D_Y} k(u,v) g(\ell_v), \qquad \delta_X^Y(u) = \alpha_u(D_Y) - \alpha_u(D_X),$$

then sets $p_X^Y(u) = e^{-\frac{\delta_X^Y(u)^2}{2\tau^2}} p_X(u)$ and assigns the remaining mass to ∂. No further normalization is required. Every experiment below uses

$$g(t) = \frac{t}{1+t},$$

which satisfies the regularity assumptions in Appendix A.1. The reference probability itself remains the standard persistence probability $p_X(u) = \frac{\ell_u}{\sum_{v \in D_X} \ell_v}$.

B.1 Two-loop configuration

The point clouds X, Y_1, Y_2, Y_3, Y_4 in Figure 1 are generated with random seeds $7, 107, 207, 307, 407$, respectively.

The large and small loops are centered at $(-1.25, 0)$ and $(1.20, 0)$, with radii 0.80 and 0.42, respectively. We sample approximately equally spaced angles on each circle and add Gaussian perturbations to both the angles and the planar coordinates. For Y_1, Y_2, Y_3, we also sample nearby concentric circles to give the loops finite thickness. Thus Y_1 contains both thickened loops, Y_2 only the large loop, and Y_3 only the small loop. The reference cloud X contains both base loops together with background noise. This background consists of eight small, perturbed circular components and points sampled uniformly from $[-2.2, 2.2] \times [-1.6, 1.6]$. Y_4 contains only these types of background components. We use the generated coordinates without centering or rescaling.

We compute finite H_1 diagrams using Ripser with Euclidean distances and Vietoris–Rips filtrations. Birth–death pairs are converted to birth–persistence coordinates, and features with persistence at most 10^{-3} are discarded. The Gaussian similarity is evaluated in birth–persistence coordinates with $\sigma_b = \sigma_\ell = 0.50$. A single response scale is used for all four comparisons:

$$\tau = 0.298499.$$

It is the median of the positive values $|\delta_X^{Y_i}(u)|$, pooled over every $u \in D_X$ and $i \in \{1,2,3,4\}$. It is calibrated once and is not adjusted separately for individual pairs.

Table 3 lists the sizes of the point clouds and diagrams together with the PE values before rounding.

For the large- and small-loop points $u_L, u_S \in D_X$, the reference probabilities are $p_X(u_L) \simeq 0.469$ and $p_X(u_S) \simeq 0.266$. The retained masses at u_S are $p_X^{Y_2}(u_S) \simeq 0.082$ and $p_X^{Y_3}(u_S) \simeq 0.083$. From Y_2 to Y_3, the u_L terms increase by 0.257 in unexplained mass and 0.733 in PCE; the changes at all other points sum to -0.004 and -0.017, respectively.

The augmented probabilities satisfy the numerical identities expected from the construction: the mass on D_X plus the mass at ∂ equals one, total variation equals the boundary mass, and entropy excess equals $\frac{1}{2\tau^2}\mathbb{E}_{u\sim p_X}[\delta_X^Y(u)^2]$, up to floating-point precision.

Table 3: Two-loop data and persistence-diagram summary.

Data	Point-cloud size	Number of H_1 points	PE
X	239	12	1.500261
Y_1	758	64	1.500305
Y_2	695	70	1.500447
Y_3	605	59	1.500307
Y_4	84	13	1.500474

B.2 Spring–mass configuration

We integrate the spring–mass equations in Section 5.2 using fourth-order Runge–Kutta, with time step 0.01 and initial state $(x, \dot{x}, y, \dot{y}) = (1, 0, 0, 1)$. From 10,000 updates, we retain 1,000 observations at intervals of 0.1, starting with the first post-update state. We discard no transient and add no observation noise. The grid $\alpha, \beta \in \{0, 0.1, \ldots, 0.8\}$ gives 81 settings: one independent case, eight cases in each one-way direction, and 64 bidirectional cases. Here α controls $B \to A$ and β controls $A \to B$.

Using the raw observations without centering or rescaling, we form four-dimensional delay vectors with lag $q = 1$:

$$X_A = \{(x_t, x_{t+1}, x_{t+2}, x_{t+3})\}_t,$$
$$X_B = \{(y_t, y_{t+1}, y_{t+2}, y_{t+3})\}_t,$$
$$X_{AB} = \{(x_t, x_{t+1}, y_t, y_{t+1})\}_t.$$

The joint reconstruction follows Cao et al. (1998). One lag equals 0.1 in physical time, and all three clouds use the first 997 common time indices.

We compute finite H_1 diagrams from alpha complexes in GUDHI 3.11.0 (Maria et al., 2014), taking square roots of its squared filtration radii. In each diagram, we divide both birth and persistence coordinates by that diagram's maximum persistence, so $\max_{u \in D} \ell_u = 1$ separately for D_A, D_B, D_{AB}. No minimum-persistence threshold is applied. We follow the published setup of Bando et al. (2022); their unavailable implementation and unspecified numerical details prevent an exact reproduction claim.

Table 4 summarizes all nine quantities in Figure 2. For landscapes and silhouettes, each birth–death interval (b, d) gives the triangular function $T_{b,d}(t) = \max\{0, \min(t - b, d - t)\}$. Betti curves, landscapes, silhouettes, and images are sampled on equally spaced grids and compared by

$$d(D_1, D_2) = \frac{\|v(D_1) - v(D_2)\|_2}{\sqrt{m}},$$

where $v(D) \in \mathbb{R}^m$ is the sampled vector, including all layers or grid points. Bottleneck and 2-Wasserstein distances use the ℓ_∞ ground metric in birth–death coordinates and allow matching to the diagonal.

The two PCE quantities use the Gaussian similarity in Section 3.1 and $g(t) = \frac{t}{1+t}$. We set τ to the median of $|\delta_A^B|$ and $|\delta_B^A|$, pooled over all points and all 81 settings.

For the persistence scale-space kernel, we use

$$k_\sigma(D_1, D_2) = \frac{1}{8\pi\sigma} \sum_{u \in D_1} \sum_{v \in D_2} \left(e^{-\frac{\|u - v\|_2^2}{8\sigma}} - e^{-\frac{\|u - \bar{v}\|_2^2}{8\sigma}} \right),$$

where u, v are in birth–death coordinates and $\bar{v}$ is the reflection of v across the diagonal. Its distance is

$$d_\sigma(D_1, D_2) = \left[k_\sigma(D_1, D_1) + k_\sigma(D_2, D_2) - 2k_\sigma(D_1, D_2) \right]^{\frac{1}{2}}. \tag{1}$$

The scale σ is one eighth of the median squared distance between points pooled from D_A, D_B, D_{AB} over all 81 settings, estimated using at most 250,000 pairs with seed 7. All grids and scales are fixed across the full experiment.

Table 4: The nine quantities used in the spring–mass comparison. The first two compare D_A and D_B directly; the other seven compare D_{AB} with each individual diagram.

Method	Quantity or representation	Settings
PCE entropy excess	PCE above PE: $\Delta H_X^Y = \frac{1}{2\tau^2}\sum_{u\in D_X} p_X(u)\delta_X^Y(u)^2$.	$\sigma_b = 0.40$, $\sigma_\ell = 0.15$; $\tau = 2.001779$.
Unexplained mass	Mass assigned to ∂: $p_X^Y(\partial) = 1 - \sum_{u\in D_X} p_X^Y(u)$.	Same settings as the entropy excess.
Bottleneck distance	Smallest maximum distance over diagram matchings.	Exact matching; ℓ_∞ ground metric.
2-Wasserstein distance	Smallest square root of the sum of squared distances over diagram matchings.	Order 2; ℓ_∞ ground metric.
Betti curve	Number of intervals containing each grid point t.	256 points on $[0, 1.8]$.
Persistence landscape	The kth layer is the kth largest value of $T_{b,d}(t)$ at each t.	First 5 layers; 256 points on $[0, 1.8]$.
Persistence silhouette	Weighted average of $T_{b,d}(t)$ over the diagram's intervals.	Uniform weights; 100 points on $[0, 1.8]$.
Persistence scale-space kernel	Gaussian kernel with diagonal reflection, converted to a distance by Eq. (1).	$\sigma = 0.016325$; birth–death coordinates.
Persistence image	Persistence-weighted Gaussian surface evaluated at grid points in birth–persistence coordinates.	32×32 grid on $[0, 1.4]^2$; Gaussian width 0.08.

The horizontal and vertical coordinates in Figure 2 are $(\Delta H_B^A, \Delta H_A^B)$ for entropy excess, $(p_B^A(\partial), p_A^B(\partial))$ for unexplained mass, and $(d(D_{AB}, D_A), d(D_{AB}, D_B))$ for every symmetric baseline.

Regime-separation score. For each method, let $z_i \in \mathbb{R}^2$ be the plotted coordinate pair for setting i, and let $G_1, \ldots, G_K$ partition the N settings by coupling regime, with $n_k = |G_k|$. We use Euclidean distances between these pairs without further coordinate rescaling. The total and within-group sums of squares are

$$\mathrm{SS}_{\mathrm{total}} = \frac{1}{N}\sum_{i<j}\|z_i - z_j\|_2^2, \qquad \text{and} \qquad \mathrm{SS}_{\mathrm{within}} = \sum_{k=1}^{K}\frac{1}{n_k}\sum_{\substack{i<j\\ i,j\in G_k}}\|z_i - z_j\|_2^2.$$

The PERMANOVA (Anderson, 2001) separation score reported in Table 1 is then

$$R^2 = \frac{\mathrm{SS}_{\mathrm{total}} - \mathrm{SS}_{\mathrm{within}}}{\mathrm{SS}_{\mathrm{total}}}.$$

Thus R^2 measures the fraction of total variation attributable to between-group separation. The four-regime comparison uses $N = 81$ and $K = 4$, with group sizes $1, 8, 8, 64$; the singleton group contributes zero to the within-group sum. The two-direction comparison uses only the $N = 16$ one-way settings, with two groups of eight. These are descriptive separation scores, not classification accuracies or permutation-test p-values.

B.3 Knowledge-distillation configuration

The knowledge-distillation experiment uses CIFAR-100 with 50,000 training images and 10,000 test images. We use the pretrained ResNet56 teacher model and the ResNet20 student architecture from the TopKD implementation (Kim et al., 2024). We report results across five paired seeds. All four methods use batch size 64 and 240 epochs of SGD with initial learning rate 0.05, momentum 0.9, and weight decay 5×10^{-4}. The learning rate is multiplied by 0.1 after epochs 150, 180, and 210. Standard random cropping with four-pixel padding and random horizontal flipping are used during training. All methods share the same student initialization and augmented training inputs

within each seed. The KD term is the teacher-to-student logit KL with temperature $T = 4$ and the usual T^2 factor. Its coefficient is 2, so $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{CE}} + 2\mathcal{L}_{\text{KD}}$.

For a mini-batch of 64 images, let F_T and F_S denote the teacher and student feature clouds at the input to the final fully connected layer. TopKD feeds these raw features into its frozen RipsNet, whereas both PCE methods ℓ_2-normalize each feature vector. We compute exact finite H_0 Vietoris–Rips persistence by a minimum spanning tree. Its 63 edge lengths are the finite persistence lifetimes; all births are zero. The teacher features and D_T are detached, while gradients pass through the selected student MST edge lengths. For teacher lifetimes t_i and student lifetimes s_j, we use

$$p_T(t_i) = \frac{t_i}{\sum_j t_j}, \qquad \kappa(t,s) = e^{-\frac{(t-s)^2}{2\sigma_\ell^2}}, \qquad g(t) = \frac{t}{1+t}.$$

The birth bandwidth has no effect because all births are zero. The function g weights the contributions in $\alpha_u(D)$; it does not replace the persistence lengths in p_T. We use the probability from Section 3, including ∂, without renormalizing the retained mass on D_T. The loss is exactly ΔH_T^S from Proposition 2, with no additional boundary penalty. We record $p_T^S(\partial)$ as a training diagnostic, separately from test accuracy.

Let R denote the frozen pretrained RipsNet. Its persistence-image predictions for the teacher and student feature clouds are

$$\widehat{\text{PI}}_T = R(F_T) \in \mathbb{R}^{400}, \qquad \text{and} \qquad \widehat{\text{PI}}_S = R(F_S) \in \mathbb{R}^{400},$$

respectively. The hat indicates a RipsNet prediction of a persistence image. The terms $\widehat{\text{PI}}_{T,j}$ and $\widehat{\text{PI}}_{S,j}$ denote the j-th components of these vectors, so $j = 1, \ldots, 400$ indexes output coordinates, not the images in the mini-batch. We define their mean squared difference by

$$\mathcal{L}_{\text{PI}} = \frac{1}{400} \sum_{j=1}^{400} \left(\widehat{\text{PI}}_{T,j} - \widehat{\text{PI}}_{S,j} \right)^2.$$

The factor $\frac{1}{400}$ averages the squared differences over all output coordinates. TopKD minimizes $\mathcal{L}_{\text{base}} + 5\mathcal{L}_{\text{PI}}$.

PCE, our main method, uses $\mathcal{L}_{\text{base}} + 0.1\Delta H_T^S$ throughout training and across all seeds. It does not load RipsNet or compute a PI loss. As an auxiliary comparison, we evaluate PCE_{AW} (adaptive weight), which uses $\mathcal{L}_{\text{base}} + c_{\text{AW}}\Delta H_T^S$, where

$$c_{\text{AW}} = 0.5 \, \frac{\max\{\|\nabla_{F_S}(5\mathcal{L}_{\text{PI}})\|_F, 10^{-12}\}}{\max\{\|\nabla_{F_S}\Delta H_T^S\|_F, 10^{-12}\}}.$$

Here $\|\cdot\|_F$ is the Euclidean norm over all feature entries. We treat c_{AW} as constant during backpropagation. When both norms exceed the numerical floor, the PCE topology gradient has half the norm of $\nabla_{F_S}(5\mathcal{L}_{\text{PI}})$. Its direction is determined by PCE. This comparison still evaluates the frozen RipsNet to compute the coefficient.

For each PCE method and seed, calibration uses eight training mini-batches before optimization. In each batch, we take the median of the pairwise teacher lifetime gaps $|t_i - t_j|$, $i < j$, greater than 10^{-12}, using a floor of 10^{-3}; if no such gap exists, we use the floor. The median of these eight values gives σ_0. We set $\sigma_\ell = \sigma_0$ for PCE and $\sigma_\ell = 0.5\sigma_0$ for PCE_{AW}. After fixing σ_ℓ, we pool the values $|\delta_T^S| > 10^{-10}$ from the same eight diagram pairs. Their median, bounded below by 10^{-3}, gives τ; if there are no such values, we set $\tau = \sigma_\ell$. These parameters remain fixed during each run but can differ across seeds and between the two PCE methods. For example, PCE at seed 0 uses $\sigma_\ell = 0.0785614$ and $\tau = 15.7586851$. The two methods differ in bandwidth as well as loss weighting, so their comparison does not isolate the effect of the loss coefficient.

Parameter selection uses a deterministic $45{,}000/5{,}000$ split of the CIFAR-100 training set, formed by a random permutation with seed 20260919. The 5,000 validation images are held out from student training; the released teacher was pretrained on all 50,000 training images. For PCE, we first screen loss coefficients 0.1, 1, and 10 for 30 epochs on 20,000 training images with seed 2026. The best value is 0.1, so the search adds 0.01 as the one allowed extension. The best two candidates, 0.01 and 0.1, are then compared for 60 epochs on all 45,000 training images with seeds 2026 and

2027. Selection uses the mean last-10-epoch validation Top-1 accuracy across these two seeds, with ties resolved in favor of the smaller coefficient. The coefficients 0.01 and 0.1 give 68.622% and 68.930%, respectively, so we choose 0.1. TopKD gives 70.005% under the same confirmation budget; the selected PCE does not outperform TopKD on this validation comparison.

$\mathrm{PCE}_{\mathrm{AW}}$ uses a separate, earlier parameter selection. With gradient ratio 1, it screens bandwidth multipliers 0.5, 1, and 2. At the selected multiplier 0.5, it also screens ratios 0.5 and 2. The best two settings, $(0.5, 0.5)$ and $(0.5, 1)$, are checked under the same $45{,}000$-image, 60-epoch, two-seed confirmation protocol. The selected bandwidth multiplier and gradient ratio are both 0.5. All parameter choices are fixed before the corresponding final runs, and official test scores are not used to choose them. Some earlier seed-0 test results existed before the four-method study was set up; the parameter-selection runs themselves use training data only.

For each method and seed, Best observed test is the highest official-test Top-1 accuracy over 240 epochs, and Last-10 mean is the mean over epochs 231–240. Table 2 reports the means for the three main methods across five seeds. Table 5 includes all four methods, giving the per-seed values and their means and sample standard deviations; the sample variance uses divisor $5 - 1$. The original study record specifies Last-10 mean as the primary measure. We also report Best observed test for comparison with the TopKD reporting convention; it is a descriptive additional measure, not a validation-selected checkpoint result.

Table 5: Per-seed CIFAR-100 test Top-1 accuracy (%) for five paired seeds. Best observed test is the maximum over 240 epochs; Last-10 mean averages epochs 231–240. Summary rows give the mean $\pm$ sample standard deviation across seeds.

Seed	KD	TopKD	$\mathrm{PCE}_{\mathrm{AW}}$	PCE
Best observed test				
0	71.17	71.29	71.44	71.32
1	71.29	70.96	71.41	71.19
2	71.30	71.16	71.41	71.48
3	71.17	71.26	71.20	71.76
4	70.99	71.37	71.24	71.03
Mean $\pm$ SD	71.18 ± 0.13	71.21 ± 0.16	71.34 ± 0.11	71.36 ± 0.28
Last-10 mean				
0	71.001	71.039	71.251	71.115
1	71.006	70.755	71.197	71.032
2	71.138	71.001	71.060	71.272
3	71.007	70.974	71.058	71.670
4	70.743	71.203	71.040	70.771
Mean $\pm$ SD	70.98 ± 0.14	70.99 ± 0.16	71.12 ± 0.10	71.17 ± 0.33

The mean gains of PCE over $\mathrm{PCE}_{\mathrm{AW}}$ are only 0.02 percentage points for Best observed test and 0.05 for Last-10 mean. The adaptive-weight variant has smaller sample standard deviations on both measures.

The Top-1 gains do not extend to every measure. At the best-Top-1 checkpoints, mean Top-5 accuracy is 92.626% for KD, 92.508% for TopKD, 92.660% for $\mathrm{PCE}_{\mathrm{AW}}$, and 92.546% for PCE.

Runs execute sequentially on a MacBook Pro with an Apple M1 Pro chip and 16 GB of unified memory, using PyTorch $2.8.0$ with the MPS backend. This is a local implementation comparison, not a reproduction of the original authors' full pretraining process and computing environment. Before including a completed paired seed, we check the initial weights and all augmented training inputs, the loss decomposition, PCE probability identities, fixed coefficients, complete training histories, and saved best-checkpoint accuracy. All five reported paired seeds passed these checks. Resume checks preserve the optimizer and random-number states.