Concept Flow: A Differential-Form Theory of Attribution on Hyperbolic Representations

Jose Marie Antonio Minoza JMINOZA@UPD.EDU.PH
Center for AI Research PH

Editors: Preprint

Abstract

Concept-based interpretability explains predictions through human concepts. It most often uses TCAV, which scores a class by the sensitivity of its logit to a concept direction in Euclidean space. Representations for hierarchical data, however, are increasingly hyperbolic, where that flat sensitivity has no direct analogue. Concept Flow brings attribution to hyperbolic representation learning as a score in differential form on the curved (Riemannian) geometry, $S[W] = \mathrm{d}h(W)$, the logit's differential $\mathrm{d}h$ paired with a concept vector field W that flows toward its centroid. The integral of $\mathrm{d}h$ along a path depends only on its endpoints, so aggregating single-concept sensitivity along any path or loop is fixed by the endpoint logits (Integrated-Gradients completeness) and carries no curvature. A curvature-dependent signal must therefore come from the connection. On a hyperbolic (Lorentz-model) representation that signal is *holonomy*. A concept direction carried around a loop comes back rotated by $c \cdot$ Area, the curvature magnitude times the area the loop encloses. The rotation is $O(c\,d^2)$, the leading-order (in the curvature c) contraction of the Riemann tensor, and vanishes in the flat limit. A loop-free multi-point comparator derived here reads the same contraction. The score is $O^+(n, 1)$-invariant and recovers Euclidean TCAV as $c \to 0$. A control lifts L_2-normalized Euclidean features to the hyperboloid and pins them to a constant-radius sphere. There holonomy collapses to a fixed function of angle, so a hierarchy signal appears only when the geometry is natively hyperbolic. On a tree hierarchy embedded with controlled curvature, holonomy tracks concept depth and grows with curvature (Spearman up to 0.87), the constant-radius control is uninformative, and single-concept sensitivity gives no comparable gain, as the theory predicts.

Keywords: concept attribution, TCAV, hyperbolic representations, holonomy, curvature, integrated gradients, interpretability

1. Introduction

Concept-based interpretability asks how a model's output responds to a human-interpretable concept. TCAV (Kim et al., 2018) represents a concept by a *concept activation vector* (CAV), the normal to a linear probe that separates activations containing the concept from those that do not, so it points in the activation-space direction of "more of concept C". It scores a class by how often nudging an input along that direction raises the class logit, a sign statistic judged significant against random concepts. The construction is Euclidean: a single global direction and a flat derivative. Representations that encode hierarchy, however, are increasingly hyperbolic (Nickel and Kiela, 2017, 2018; Ganea et al., 2018; Desai et al., 2023; Pal et al., 2025), and there the tools stop at concept *presence*, via cone-containment margins (Uyterlinde et al., 2026) or transport-based steering (Briglia et al., 2026), with no sensitivity score and no significance test.

Concept Flow's score is the pairing of the logit's differential $\mathrm{d}h$ with a concept vector field, which makes the role of geometry precise. Since $\mathrm{d}h$ is exact, aggregating single-concept sensitivity along any path telescopes to an endpoint logit difference, the manifold form of Integrated-Gradients completeness (Sundararajan et al., 2017), so it carries no curvature information. The curvature-driven, "emergent" part of concept structure arises one order up, in the connection, as the holonomy of transporting a concept direction around a loop of concepts (Figure 1).

Representation holonomy has recently been shown to be a meaningful, gauge-invariant statistic of learned geometry (Sevetlidis and Pavlidis, 2026). That work measures holonomy as a diagnostic of an unknown, emergent curvature, with no concepts, attribution, or hyperbolic structure. This work gives holonomy semantic content as a concept-attribution signal on a *known* hyperbolic geometry, and establishes what it measures, why a holonomy-type signal is required, and when it carries hierarchical information.

Contributions. (i) Concept attribution is formulated as the differential form $S[W] = \mathrm{d}h(W)$, and a factorization no-go shows that path and loop aggregation of single-concept sensitivity is fixed by the pointwise score field and is curvature-blind (Theorems 2 and 3). (ii) On $\mathbb{L}_c^n$ the loop holonomy $c \cdot \text{Area} = O(c\,d^2)$ is the canonical curvature-driven signal. It has an exact zero flat baseline and is h-independent, so it escapes the impossibility class of Bilodeau et al. (2024), and a Jacobi-field expansion identifies it with a loop-free comparator as the same first-order Riemann contraction (Theorems 4 and 5). (iii) The score is $O^+(n,1)$-invariant, reduces to Euclidean TCAV as $c \to 0$, and admits a unique Fréchet mean. (iv) A control theorem shows the signal collapses under a constant-radius (Euclidean-lift) representation (Theorem 8), and a controlled tree-embedding study confirms the holonomy-versus-depth prediction, and its predicted growth with curvature, against that control.

2. Background

Embeddings are points of the Lorentz model $\mathbb{L}_c^n = \{x \in \mathbb{R}^{n+1} : \langle x, x\rangle_{\mathbb{L}} = -1/c,\ x_0 > 0\}$, the upper sheet of a hyperboloid in dimension n and curvature $-c$, with $c > 0$ the curvature magnitude and $\langle x, y\rangle_{\mathbb{L}} = -x_0 y_0 + \sum_{j\geq 1} x_j y_j$ the Minkowski inner product (x_0 the time coordinate). Each input is embedded at a point $z \in \mathbb{L}_c^n$, and $o = (1/\sqrt{c}, 0, \ldots, 0)$, the apex of the sheet, serves as the origin. The tangent space $T_z\mathbb{L}_c^n = \{u : \langle z, u\rangle_{\mathbb{L}} = 0\}$, the directions in which a point can move off from z, carries the metric $g_z = \langle \cdot, \cdot \rangle_{\mathbb{L}}$, and $\Pi_z(v) = v + c\langle z, v\rangle_{\mathbb{L}} z$ projects an ambient vector onto it. The exponential map $\exp_x$ follows the geodesic leaving x in a given tangent direction, and its inverse $\log_x(y) \in T_x\mathbb{L}_c^n$ is the initial velocity of the unit-time geodesic from x to y, so it points along that geodesic with length the distance from x to y. Parallel transport $P_{x\to y}$ carries a tangent vector along the geodesic from x to y, and $d(x, y)$ is geodesic distance, written d when a single representative magnitude is meant. Their closed forms are collected in Appendix A. A concept C has a centroid $\mu_C \in \mathbb{L}_c^n$, its Fréchet mean (the point minimizing the summed squared geodesic distance to its examples), which corresponds to the Euclidean mean in flat space. The Riemannian gradient of a logit $h : \mathbb{L}_c^n \to \mathbb{R}$ (the model's scalar score for a class) is $\operatorname{grad} h(z) = \Pi_z(J\nabla h(z))$. Here ∇h is the ordinary gradient in the surrounding $\mathbb{R}^{n+1}$, $J = \operatorname{diag}(-1, 1, \ldots, 1)$ turns it into a tangent vector under the Minkowski inner product, and Π_z projects the result onto $T_z\mathbb{L}_c^n$.

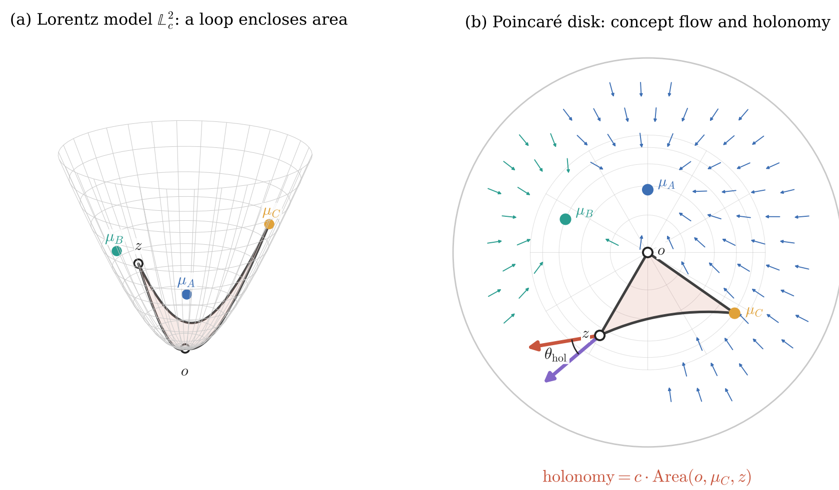

Figure 1: **Concept flow on a hyperbolic representation.** *(a)* The Lorentz model $\mathbb{L}_c^n$ with the geodesic loop $o \to \mu_C \to z$. *(b)* The same points in the Poincaré disk: the concept field $W(z) = \log_z \mu$ (arrows) flowing toward two example concept centroids μ_A, μ_B, and at z the holonomy angle θ_{hol} for the loop of *(a)*, between transport around it and along the direct path $o \to z$.

Euclidean TCAV (Kim et al., 2018) scores class k by the fraction of inputs whose logit rises along the CAV v_C, that is by sign statistics of $\nabla h_k \cdot v_C$. The concept counts only when this score exceeds those of random concepts. Integrated Gradients (Sundararajan et al., 2017) instead integrates the gradient along a path from a baseline and satisfies *completeness*, so the attributions sum to the difference of logits at the two endpoints. Both facts reappear below as the flat, zeroth-order face of the theory.

3. Concept Flow

Definition 1 (Concept-flow score) *For $h \in C^\infty(\mathbb{L}_c^n)$ and a tangent vector field W, the concept-flow score is $S[W](z) = g_z(\operatorname{grad} h(z), W(z)) = \mathrm{d}h_z(W(z))$.*

The concept direction $w_C(z) := \log_z \mu_C / \|\log_z \mu_C\|_z$ is the unit geodesic direction from z toward the centroid. TCAV fits a linear probe on concept and random examples for its direction. The hyperbolic direction is this geodesic instead, taken from the concept centroid with no probe to fit. With $W = w_C$ the score becomes the hyperbolic TCAV score $S^{\mathbb{H}}_{C,k}(z) = S[w_C](z) = g_z(\operatorname{grad} h_k(z), w_C(z))$ for the class-k logit h_k. A relation field $\log_z \mu_B - \log_z \mu_A$ and a trajectory velocity $\dot{z}(t)$ are further instances of W. Everything below rests on $\mathrm{d}h$ being *exact*: it is the differential of the scalar h, so its integral along a path depends only on the endpoints, and its integral around any closed loop is zero.

The pointwise sensitivity, integrated along the whole geodesic from an input toward its centroid, gains nothing, because the integral telescopes to its endpoints.

Theorem 2 (Completeness / telescoping) *For any piecewise-C^1 curve γ, $\int_\gamma \mathrm{d}h(\dot\gamma)\,\mathrm{d}t = h(\gamma_1) - h(\gamma_0)$ and $\oint \mathrm{d}h(\dot\gamma)\,\mathrm{d}t = 0$, identically in the metric. In particular the "integrated" hyperbolic sensitivity equals $d(z,\mu_C)\, S[\,\widehat{\log_z \mu_C}\,](z)$ exactly, with $\widehat{u} := u/\|u\|_z$ the unit-normalized vector, and carries the same sign statistic as the local score.*

This is the manifold form of IG completeness. The integrand $(h \circ \gamma)'$ involves neither the metric nor the curvature. Aggregation therefore adds nothing geometric, and any curvature-dependent signal must arise elsewhere.

Theorem 3 (Factorization no-go) *Any functional formed by aggregating $S[W]$ along paths or loops depends only on the zeroth-order score field along the curve, carrying no curvature beyond it. A curvature-isolating attribution signal therefore requires the connection acting on W (holonomy or covariant derivative). The scope is sharp. Call a functional length-homogeneous of degree w if it scales by λ^w when all lengths scale by λ. Its flat-limit defect is then $O(c\,d^{2+w})$ unconditionally, and $\Theta(c\,d^{2+w})$ exactly when its leading curvature term $F^{(1)}$ (Theorem 5) is nonzero. Holonomy is dimensionless ($w = 0$) with defect $\Theta(c\,d^2)$, and the raw comparator has $w = 1$. This rate is not universal. It needs $F^{(1)} \neq 0$, and fails for θ^2_{hol}, which has $F^{(1)} = 0$ and defect $\Theta(c^2 d^4)$.*

A single concept, probed pointwise or integrated toward its centroid, carries no more curvature information than the flat Euclidean score. Hyperbolic geometry begins to help only when the probe encircles more than one concept, so that the loop encloses area.

4. Curvature via Holonomy

The pointwise score is zeroth order, a value at each point. The connection, which transports a direction along the manifold, is first order. The curvature, the rotation left after transport around a loop, is second order.

Parallel transport slides a tangent vector along a curve, keeping it as constant as the curved geometry permits. On a flat space, transporting a vector around any closed loop returns it unchanged. On a curved space it comes back rotated, and that rotation angle, the *holonomy* of the loop, measures the curvature enclosed by it. On a globe, for example, an arrow carried around a triangle, always held parallel to itself, comes back pointing in a new direction, rotated by the enclosed area times the curvature though no one ever turns it. For constant curvature $-c$, Gauss–Bonnet makes this exact, and the holonomy around a geodesic triangle equals c times its area. Curvature is a local quantity, defined at each point, and holonomy is its global counterpart, the curvature integrated over the region a loop encloses (Ambrose–Singer). A concept direction $\log_z \mu_C$ is itself a tangent vector, so transporting it around a loop of concepts turns this classical fact into an attribution signal.

Theorem 4 (Holonomy equals curvature) *The holonomy of transporting a concept direction around a simple geodesic loop of nonzero signed area in $\mathbb{L}^n_c$ has magnitude $|\theta_{\mathrm{hol}}| = c \cdot \mathrm{Area} = O(c\,d^2)$ and vanishes as $c \to 0$. The two-path realization (transport along $o \to z$ versus $o \to \mu_C \to z$) defines the holonomy gap $\theta_{\mathrm{hol}}(z, \mu_C)$. The raw angle between the local and transported directions does* not *isolate curvature, because it retains a nonzero flat baseline that only the closed loop cancels.*

Theorem 5 (Curvature-contraction invariant) *Expand a functional's flat-limit defect in the curvature as* $F_c = \sum_{k\geq 0} c^k F^{(k)}$*, with* F_0 *its value at* $c = 0$*. For a curvature-isolating functional built from the geometry primitives* $F_0 \equiv 0$*, so the leading defect is* $c\,F^{(1)}$*, with* $F^{(1)}$ *a linear contraction of the Riemann tensor, from the expansion* $\log_z m = \frac{a+b}{2} - \frac{c}{12}|v|^2 a_\perp + O(c^2 d^5)$ *(*m *the midpoint of* μ_A, μ_B*, with* $a = \log_z \mu_A$*,* $b = \log_z \mu_B$*,* $v = b - a$*, and* $a_\perp$ *the part of* a *orthogonal to* v*). Holonomy (*$F^{(1)} = \pm\frac{1}{2}|a \wedge b|$*, the area spanned by* a *and* b*,* h*-independent) and the loop-free comparator* $W_{\text{comp}} = \log_z \mu_A + \log_z \mu_B - 2\log_z m = +\frac{c}{6}|v|^2 a_\perp$ *(*$F^{(1)} = -\frac{1}{6}\langle \operatorname{grad} h, R(a,v)v\rangle$*) are the same first-order Riemann contraction, holonomy read against the area bivector* $a \wedge b$ *and the comparator against* $\operatorname{grad} h$*.*

The comparator is a discrete second derivative of three concept points. In flat space the log map is affine, so the midpoint m of μ_A and μ_B satisfies $\log_z m = \frac{1}{2}(\log_z \mu_A + \log_z \mu_B)$ and W_{comp} vanishes. Curvature breaks this midpoint identity, and the defect is exactly the Jacobi-field term $\frac{c}{6}|v|^2 a_\perp$, so three centroids and their log maps detect curvature with no loop and no gradient. This is the pointwise, loop-free counterpart of the holonomy (Theorem 4).

Bilodeau et al. (2024) show that any attribution that is complete and linear in the model can be driven to fail simple recovery tasks. Because holonomy is h-independent, it falls outside that class. The comparator is cheaper but linear in $\mathrm{d}h$, so it does not escape that class. A first-order, trajectory-based sibling of holonomy, benchmarked in Table 3, is the concept angular velocity $\omega_C = \sqrt{c}\coth(\sqrt{c}\,d)\sin\theta$, with $d = d(z, \mu_C)$ and θ the angle at z between the trajectory velocity and the concept direction. It has flat limit $\sin\theta/d$ and loop aggregate $\oint \omega_C\,\mathrm{d}s = 2\pi\,\mathrm{wind}(\mu_C) + c \cdot \mathrm{Area}$, where $\mathrm{wind}(\mu_C)$ counts how many times the loop winds around μ_C.

Computing the holonomy gap. The gap follows directly from the closed forms of Appendix A, with no gradient of h. A concept C fixes a centroid μ_C, the Fréchet mean of its examples and the point minimizing the summed squared geodesic distance to them, which sets a reference direction $w_0 = \log_o \mu_C / \|\log_o \mu_C\|_o$ at the origin o, the unit initial velocity of the geodesic from o toward μ_C. Parallel transport carries w_0 to the input z two ways, directly along $o \to z$ as $w_{\text{direct}} = P_{o\to z} w_0$ and around the concept along $o \to \mu_C \to z$ as $w_{\text{loop}} = P_{\mu_C \to z} P_{o \to \mu_C} w_0$, and the gap is the angle between the two arrivals, $\theta_{\text{hol}}(z, \mu_C) = \arccos g_z(w_{\text{direct}}, w_{\text{loop}})$. Because transport is a g-isometry, both arrivals stay unit vectors, and the gap is a direct arc cosine of their Minkowski inner product. The same test carries over. With random-concept centroids the construction gives a baseline of gaps, and a concept counts only when its gap exceeds that baseline. Each pair costs a few exponential and log maps, evaluated with the numerically stable forms of Appendix A (the clamped $\alpha \geq 1$ and the fractional $P_{x\to y}$) to avoid the near-coincidence instability of a log-based transport.

5. Invariance and the Flat Limit

The score depends only on the geometry, independent of the coordinates chosen to describe it, and it recovers Euclidean TCAV in the flat limit.

Theorem 6 (Isometry invariance) *For $Q \in O^{+}(n,1)$, with $z \mapsto Qz$, $\mu_C \mapsto Q\mu_C$, $h \mapsto h \circ Q^{-1}$, the score is unchanged: $S^{\mathbb{H}}_{C,k}(Qz) = S^{\mathbb{H}}_{C,k}(z)$. It is independent of the choice of origin and Lorentz frame.*

The score is a g-inner product of two geometric objects, the gradient and the concept direction, and an isometry preserves that inner product, so the value cannot depend on the frame.

Proposition 7 (Flat limit) *In spatial coordinates, with h a fixed function of those coordinates, as $c \to 0$ the score converges to the Euclidean TCAV directional derivative at rate $O(c)$. The Fréchet mean exists and is unique, and the score equals the geodesic directional derivative $D_{w_C} h$.*

As the curvature flattens, the metric and the log map become Euclidean, so the score keeps no memory of the geometry and reduces to the ordinary directional derivative. The manifold also holds many concepts at once. The number of centroids that can be kept at least δ apart within a geodesic ball of radius R grows as $\log N = (n-1)\sqrt{c}\,R + \Theta(1)$, exponentially in R against only polynomially in Euclidean space, so the space has far more room for distinct concepts. Across M concepts and K classes the many tests are controlled by the Benjamini–Hochberg procedure, which caps the false discovery rate, with a random-concept null drawn for each test as in TCAV.

6. The Constant-Radius Control

A control that removes native curvature isolates its role.

Theorem 8 (Constant-radius collapse) *The exponential lift of L_2-normalized Euclidean features ($\|f\| = \rho$) by $\exp_o$ places every embedding on the geodesic sphere $S_\rho(o)$, isometric to the round sphere $S^{n-1}(\sinh(\sqrt{c}\rho)/\sqrt{c})$. The radial (hierarchy) coordinate has zero variance. The score reduces to a positively-rescaled angular derivative, so $\mathrm{TCAV}^{\mathbb{H}}$, the TCAV sign statistic formed from $S^{\mathbb{H}}$, equals Euclidean TCAV and is independent of c, and the holonomy signal collapses to the deterministic curve $\pi - \Delta - 2\beta(\Delta)$, for an explicit function β of the angular separation Δ given in the proof, with zero conditional variance given Δ. Hierarchy information requires natively-trained hyperbolic curvature.*

Any gain that survives the constant-radius lift is therefore attributable to native curvature, the rigorous form of the near-Euclidean collapse reported for publicly released hyperbolic models, whose learned curvature is often close to zero (Kim et al., 2026a). A fixed-length normalization pins every embedding to a sphere of one radius, so the depth coordinate a hierarchy needs is discarded before the score is computed, leaving only an angle that a Euclidean method already captures.

7. Experiments

A balanced b-ary tree ($b = 3$, depth 4, 121 nodes) is embedded in $\mathbb{L}^n_c$ by placing each node at a radius proportional to its depth and spreading siblings in angle, so a parent always lies closer to the apex o than its children and the radial coordinate is exact hierarchy depth.

Table 1: Spearman correlation between the holonomy score and hierarchy depth (non-thin triangles, mean over five seeds, per-seed standard deviation ≤ 0.06). In the native embedding the correlation rises with the curvature magnitude c at $n = 8$. At $n = 2$ the rise is among non-thin triangles, whose excluded near-collinear fraction itself grows with c. The constant-radius control stays uninformative at every c.

dim	arm	$c = 0.25$	$c = 0.5$	$c = 1$	$c = 2$
$n = 2$	native	$+0.60$	$+0.62$	$+0.68$	$+0.79$
$n = 2$	control	$+0.00$	$+0.00$	-0.01	-0.02
$n = 8$	native	$+0.75$	$+0.79$	$+0.84$	$+0.87$
$n = 8$	control	-0.16	-0.15	-0.11	$+0.02$

The curvature c is chosen, not learned, so the ground truth is known and the holonomy prediction can be checked against true depth. The native arm keeps this depth-encoding radius. The constant-radius control (the Euclidean lift of Theorem 8) moves every node to one radius and keeps only the angles, which removes native curvature. For each concept–node pair the holonomy gap θ_{hol} is computed by transporting a concept direction around the loop $o \to \mu_C \to z$ against the direct path $o \to z$ and recording the rotation, and it is correlated (Spearman) with hierarchy depth across curvatures $c \in \{0.25, 0.5, 1, 2\}$, dimensions $n \in \{2, 8\}$, and five seeds.

The holonomy signal tracks depth (Table 1). In the native $n = 8$ embedding the Spearman correlation rises with curvature from 0.75 to 0.87 as the larger intrinsic size $\sqrt{c}\,r$ orders the enclosed area more sharply by depth, up to the saturation of Theorem 4. At $n = 2$ the correlation among non-thin triangles rises from 0.60 to 0.79, but the near-collinear fraction the thin threshold removes grows with c from a sixth to over a half, so on all pairs the $n = 2$ number declines and the rise belongs to the surviving subset. The constant-radius control is uninformative at every curvature ($|\text{Spearman}| \leq 0.16$, and within 0.02 of zero for $n = 2$), with per-seed standard deviation at most 0.06. Conditioning on angular separation, holonomy's variance is 0.84 in the native arm and 0.03 in the control, so freezing the radius turns holonomy into an almost deterministic function of angle (Figure 2). The $n = 8$ embedding avoids thin (near-collinear) triangles because distinct concepts occupy near-orthogonal axes, which is why it outscores $n = 2$.

On this tree the single-concept scores track depth far more weakly than the second-order probes at every curvature, and the loop-free comparator matches the holonomy to first order in the curvature (Theorem 5), so the measured gain is second-order (Table 3). The gap between the native and control arms is the empirical signature of Theorem 8. With the radius frozen, holonomy reduces to a deterministic function of angular separation and no longer tracks depth.

The same prediction holds on a real, irregular hierarchy. Four is-a subtrees of WordNet (Miller, 1995), rooted at carnivore, mammal, primate, and aquatic mammal (60 to 400 synsets with uneven branching and depth), are embedded with radius equal to WordNet depth (Figure 3a), and holonomy tracks that depth with Spearman near 0.9 at moderate

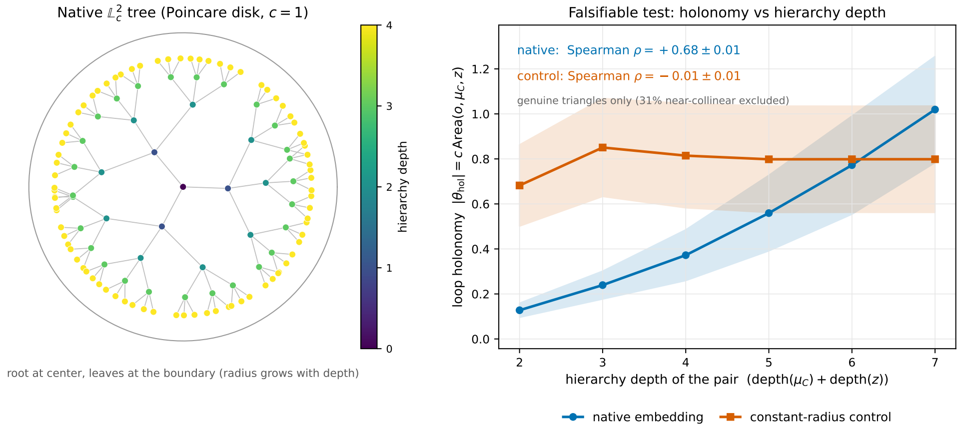

Figure 2: Holonomy tracks hierarchy depth on a controlled tree embedding and collapses under the constant-radius control. Left: the $\mathbb{H}^2$ ($n = 2$) tree colored by depth. Right: holonomy versus depth for the plotted $n = 2$ embedding at $c = 1$ (native Spearman 0.68 on non-thin triangles, control ≈ 0). The $n = 8$ embedding, which avoids thin triangles, reaches 0.84.

curvature (Figure 3b). The correlation dips at $c = 2$, the larger trees first, the area saturation of Theorem 4 once $\sqrt{c}\,d$ leaves the small-triangle regime. Radius is set to WordNet depth rather than trained, so the test isolates whether the holonomy–depth law survives real irregular branching, which it does.

8. Related Work

Holonomy in interpretability. Sevetlidis and Pavlidis (2026) establish representation holonomy as a gauge-invariant (frame-independent) diagnostic of learned geometry. Javidnia (2026) treat holonomy as an obstruction to global interpretability, and Richards (2026) apply the instrument to sparse features. None attaches holonomy to concepts, attribution, or hyperbolic geometry. This work supplies all three. The connection is known and hyperbolic, and its holonomy can be thought of as a concept-attribution signal.

Hyperbolic concepts. Uyterlinde et al. (2026) give concept presence via entailment cones and Briglia et al. (2026) steer concepts by parallel transport. Neither computes a sensitivity score or a significance test.

Riemannian attribution. Manifold and geodesic Integrated Gradients (Zaher et al., 2024; Salek and Enguehard, 2025; Martino and Tschiatschek, 2026; Kim et al., 2026b) integrate the gradient along geodesics of a chosen metric, so geometry enters through the integration path. By Theorem 3 that path aggregation carries no curvature information. These methods change the route of integration, whereas the holonomy changes the measured object, from a gradient pairing to a transport defect of the connection.

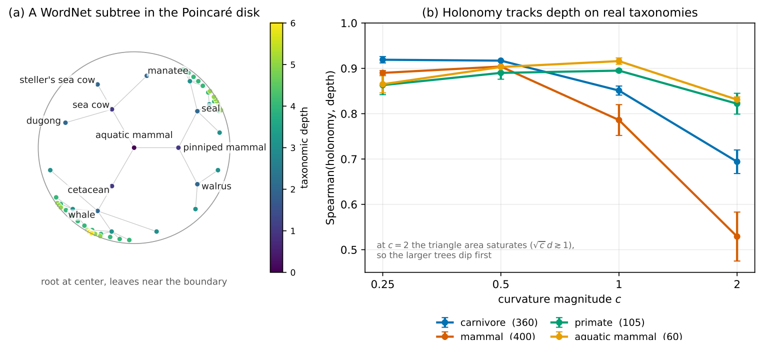

Figure 3: Concept Flow on real WordNet taxonomies. (a) An is-a subtree (aquatic mammal, 60 synsets) embedded in $\mathbb{H}^2$ with radius equal to taxonomic depth, higher ranks near the centre and species near the boundary. (b) Spearman correlation between holonomy and taxonomic depth for four subtrees across curvature, strong at moderate c and dipping at $c = 2$ where the triangle area saturates, the larger trees first.

Foundations. This work generalizes TCAV (Kim et al., 2018) and its directional derivative, uses the completeness axiom of Sundararajan et al. (2017), and positions the holonomy probe against the impossibility theorems of Bilodeau et al. (2024).

9. Discussion

The holonomy signal is measurable when the geodesic triangle is non-thin and its size satisfies $\sqrt{c}\,d \lesssim 1$, beyond which the enclosed area saturates. Publicly released hyperbolic image–text models such as MERU (Desai et al., 2023) and HyCoCLIP (Pal et al., 2025) often train to such low effective curvature that their embeddings are nearly flat (Kim et al., 2026a), and there c·Area is too small to detect. Theorem 8 also shows how to avoid it. Fixing the feature norm discards the radial axis, so hierarchy depth must sit on the geodesic radius for the signal to survive. Validation here therefore uses a tree embedded with controlled curvature, and the same diagnostic transfers to a high-curvature embedding trained from scratch. A relation field $\log_z \mu_B - \log_z \mu_A$ is the tangent vector at z from one concept toward another, the hyperbolic counterpart of the word-vector difference a Euclidean analogy adds and subtracts. An analogy transports that relation to a third concept, and because transport on a curved space is path-dependent, the analogy fails to close by the holonomy $c \cdot \text{Area}$, with the flat subtraction recovered only as $c \to 0$. Entailment cones, which encode an is-a hierarchy as an asymmetric containment order, suggest a directed score with no Euclidean analogue, one that tests whether one concept entails another. Because holonomy adds over adjacent loops, a composition of several concepts follows from a single loop that visits them, the same path-ordered quantity that measures gauge flux around a Wilson loop or the winding of a knot. The probe reaches any hierarchy carried in hyperbolic space, a protein or molecular taxonomy or a chain of analogical reasoning, where a nonzero loop holonomy marks relations that fail to commute.

References

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences (PNAS)*, 121(2), 2024. doi: 10.1073/pnas.2304406120.

Maria Rosaria Briglia, Simone Facchiano, Paolo Cursi, Alessio Sampieri, Emanuele Rodolà, Guido Maria D'Amely di Melendugno, Luca Franco, Fabio Galasso, and Iacopo Masi. Not all latent spaces are flat: Hyperbolic concept control. *arXiv preprint arXiv:2603.14093*, 2026.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Hossein Javidnia. A gauge theory of superposition: Toward a sheaf-theoretic atlas of neural representations. *arXiv preprint arXiv:2603.00824*, 2026.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Jaeyoung Kim, Eunseok Kim, and Dongsuk Jang. Is the geometry doing the work? an operating-point audit of hierarchy in hyperbolic vision-language models. *arXiv preprint arXiv:2607.05268*, 2026a.

Soyeon Kim, Seongwoo Lim, Kyowoon Lee, and Jaesik Choi. Spectral integrated gradients for coarse-to-fine feature attribution. *arXiv preprint arXiv:2605.19607*, 2026b.

Gabriele Martino and Sebastian Tschiatschek. Fringe: Distribution-space integrated gradients with fisher-rao geometry. *arXiv preprint arXiv:2605.06404*, 2026.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Avik Pal, Max van Spengler, Guido Maria d'Incau, Raphael Memmesheimer, Deepak Bhattacharjee, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

L. Richards. Do active sae feature planes carry more holonomy? a preregistered study on gemma. *arXiv preprint arXiv:2607.20522*, 2026.

Sina Salek and Joseph Enguehard. Using the path of least resistance to explain deep networks. *arXiv preprint arXiv:2502.12108*, 2025.

Vasileios Sevetlidis and George Pavlidis. Gauge-invariant representation holonomy. In *International Conference on Learning Representations (ICLR)*, 2026.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 3319–3328, 2017.

Zhiyuan Tao, Srikumar Sastry, Matthew J. Thompson, Yu Su, Elizabeth G. Campolongo, Net Zhang, Ziheng Zhang, Hilmar Lapp, Tanya Berger-Wolf, Nathan Jacobs, Wei-Lun Chao, and Jianyang Gu. Beyond flat labels: Level-restricted contrastive learning for hierarchical fine-grained vision classification. *arXiv preprint arXiv:2606.21838*, 2026.

Daniel Uyterlinde, Swasti Shreya Mishra, and Pascal Mettes. Hyperbolic concept bottleneck models. *arXiv preprint arXiv:2605.06440*, 2026.

Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Manifold integrated gradients: Riemannian geometry for feature attribution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Appendix A. Lorentz-model closed forms and stability

Write $\alpha := -c\langle x, y\rangle_{\mathbb{L}} = \cosh(\sqrt{c}\, d(x, y)) \geq 1$ and $\|u\|_x = \sqrt{\langle u, u\rangle_{\mathbb{L}}}$ for $u \in T_x\mathbb{L}^n_c$. The verified closed forms are

$$\begin{aligned} d(x, y) &= \tfrac{1}{\sqrt{c}} \operatorname{arccosh}(\alpha), & \exp_x(u) &= \cosh(a)\, x + \operatorname{sinhc}(a)\, u, \quad a = \sqrt{c}\, \|u\|_x, \\ \log_x(y) &= \tfrac{\operatorname{arccosh}(\alpha)}{\sqrt{\alpha^2-1}}\, (y - \alpha x), & P_{x\to y}(v) &= v + \tfrac{c\,\langle y,v\rangle_{\mathbb{L}}}{1+\alpha}\, (x + y), \end{aligned}$$

with $\Pi_x(v) = v + c\langle x, v\rangle_{\mathbb{L}} x$ and $\operatorname{grad} h(x) = \Pi_x(J\nabla h(x))$, $J = \operatorname{diag}(-1, 1, \ldots, 1)$. Here $\operatorname{sinhc}(a) = \sinh(a)/a$ and the coefficient $\operatorname{arccosh}(\alpha)/\sqrt{\alpha^2 - 1}$ both tend to 1 at the removable singularities ($a = 0$, $\alpha = 1$). Evaluate them by their series there, clamp $\alpha \geq 1$ against rounding, and use the form above for $P_{x\to y}$ (denominator $1 + \alpha \geq 2$), which avoids the near-coincidence instability of a log-based form. For very large radius replace the arccosh distance by the chordal form

$$d = \tfrac{2}{\sqrt{c}} \operatorname{arcsinh}\Big(\tfrac{\sqrt{c}}{2} \sqrt{-\langle x - y, x - y\rangle_{\mathbb{L}}}\Big),$$

which avoids the catastrophic cancellation of $\operatorname{arccosh}(-c\langle x, y\rangle_{\mathbb{L}})$ beyond $\sqrt{c}\, R \approx 16$.

Appendix B. Proofs

Proof [Theorem 2] Let γ be the geodesic from z to μ_C, so $\gamma'(0) = \log_z \mu_C$ and $\|\gamma'\| \equiv d$. Fix any $G \in T_z$. Parallel transport is a g-isometry, and along a geodesic the velocity is parallel, so $\gamma'(t) = P_{z\to\gamma(t)}\gamma'(0)$ and $g_{\gamma(t)}(P_{z\to\gamma(t)}G, \gamma'(t)) = g_z(G, \gamma'(0))$ is constant in t. Integration gives $g_z(G, \log_z \mu_C) = d\, g_z(G, \widehat{\log_z \mu_C})$, the claimed collapse. For a general curve, $\mathrm{d}h(\dot\gamma) = (h \circ \gamma)'$, so $\int_\gamma \mathrm{d}h(\dot\gamma) = h(\gamma_1) - h(\gamma_0)$ and every loop integral vanishes. The uniqueness of geodesics on the Cartan–Hadamard sheet makes the transported path the sub-arc, so no path-dependence enters. ∎

Proof [Theorem 3] Any functional $\int F(t, S[W]\circ\gamma)\,\mathrm{d}t$ depends only on the restriction of the pointwise field $S[W] = \mathrm{d}h(W)$ to γ, so aggregation adds no information beyond that field. The velocity-pairing loop integral is 0 for every metric (Theorem 2), hence curvature-blind, whereas the holonomy of transporting W around a loop is curvature-exact (Theorem 4). For the scope, the map $x \mapsto \sqrt{c}\,x$ is an isometry up to scale onto $\mathbb{L}^n_1$, and every primitive is an even entire function of $\sqrt{c} \times$ (lengths), so $F_c = \sum_k c^k F^{(k)}$ with $F^{(k)}$ homogeneous of degree $2k + w$. If $F_0 \equiv 0$ then the leading term is $cF^{(1)}$, a linear contraction of the Riemann tensor (Theorem 5), and $F_c = \Theta(c\,d^2 \cdot d^w)$ exactly when $F^{(1)} \neq 0$. The universal claim fails: θ^2_{hol} isolates curvature yet has $F^{(1)} \equiv 0$ and leading defect $\Theta(c^2 d^4)$. ∎

Proof [Theorem 4] For a geodesic triangle of curvature $-c$, Gauss–Bonnet gives angle defect $\pi - \sum$ angles $= c \cdot$ Area, and by Ambrose–Singer the holonomy around the loop equals the enclosed integrated curvature, so $|\theta_{\mathrm{hol}}| = c \cdot$ Area. The two-path construction transports around the triangle (o, μ_C, z). The raw angle between $\log_z \mu_C$ and $P_{o\to z}\nu_C$ (with ν_C the concept direction at the origin o, and a bar denoting the spatial part of a point) tends to the Euclidean angle between $\bar\mu - \bar z$ and $\bar\mu - \bar o$ as $c \to 0$, a nonzero baseline, so only the closed loop isolates the $O(c\,d^2)$ term. Numerically the identity holds to relative error $\leq 10^{-11}$. ∎

Proof [Theorem 5] In normal coordinates at z, $g_{ij} = \delta_{ij} - \frac{c}{3}(|x|^2\delta_{ij} - x_i x_j) + O(|x|^4)$. The geodesic equation for the midpoint of μ_A, μ_B solves to $\log_z m = \frac{a+b}{2} - \frac{c}{12}|v|^2 a_\perp + O(c^2 d^5)$. Hence $W_{\mathrm{comp}} = a + b - 2\log_z m = \frac{c}{6}|v|^2 a_\perp = -\frac{1}{6}R(a, v)v$, with R the Riemann curvature tensor, and $S[W_{\mathrm{comp}}] = -\frac{1}{6}\langle \operatorname{grad} h, R(a, v)v\rangle = \Theta(c)$, cubic in d before normalizing by $|v|$. The holonomy contraction is $F^{(1)} = \pm\frac{1}{2}|a \wedge b|$, h-independent. Both capture the single first-order invariant, holonomy against the area bivector and the comparator against $\operatorname{grad} h$. ∎

Proof [Theorem 6] For $Q \in O^+(n, 1)$, $Q^\top J Q = J$ gives $Q^{-1} = JQ^\top J$. Then $\langle Qx, Qy\rangle_{\mathbb{L}} = \langle x, y\rangle_{\mathbb{L}}$ leaves α, d invariant, $\Pi_{Qz}(Qv) = Q\Pi_z(v)$, and $\log_{Qz}(Q\mu) = Q\log_z \mu$. With $h' = h \circ Q^{-1}$, $\nabla h'(Qz) = JQJ\,\nabla h(z)$, so $J\nabla h'(Qz) = QJ\nabla h(z)$ and $\operatorname{grad} h'(Qz) = \Pi_{Qz}(QJ\nabla h) = Q \operatorname{grad} h(z)$. Therefore $S^{\mathbb{H}}(Qz) = \langle Q \operatorname{grad} h, Q\, w_C\rangle_{\mathbb{L}} = \langle \operatorname{grad} h, w_C\rangle_{\mathbb{L}} = S^{\mathbb{H}}(z)$. Two smooth extensions of h agreeing on the sheet differ by $\varphi\,(c\langle x, x\rangle_{\mathbb{L}} + 1)$, whose J-gradient is proportional to z and killed by Π_z, so the score is extension-independent. ∎

Proof [Proposition 7] In spatial coordinates $x = (\sqrt{1/c + \|\bar x\|^2}, \bar x)$ one has $\alpha = 1 + \frac{c}{2}\|\bar z - \bar\mu\|^2 + O(c^2)$, so $d \to \|\bar z - \bar\mu\|$ and the spatial part of $\log_z \mu \to \bar\mu - \bar z$. The induced metric tends to δ, hence $\operatorname{grad} h \to \nabla_{\bar x} h$ and $S^{\mathbb{H}} \to \nabla_{\bar x} h \cdot \widehat{(\bar\mu - \bar z)}$ at rate $O(c)$. Uniqueness of μ_C follows because the sheet is Cartan–Hadamard, on which $d(\cdot, p)^2$ is geodesically strictly convex, so the Fréchet functional has a unique minimizer. ■

Proof [Theorem 8] For $\|f\| = \rho$, $E(f) = \exp_o((0, f)) = \left(\frac{\cosh\hat{\rho}}{\sqrt{c}}, \frac{\sinh\hat{\rho}}{\sqrt{c}}\hat{f}\right)$ with $\hat{\rho} = \sqrt{c}\rho$, so $d(o, E(f)) = \rho$ and every embedding lies on $S_\rho(o)$. The parametrization by $u \in S^{n-1}$ gives induced metric $R^2|\mathrm{d}u|^2$ with $R = \sinh(\hat{\rho})/\sqrt{c}$, an isometry to $S^{n-1}(R)$ of intrinsic curvature $+c/\sinh^2\hat{\rho}$. The radial coordinate is constant, hence carries zero variance. For $z, \mu_C \in S_\rho(o)$ at angle Δ, $\log_z \mu_C$ has radial component $-\cos\beta(\Delta)\, d$ with β deterministic in Δ, so under the radial extension of h the score is $S^{\mathbb{H}}_{C,k} = \frac{\sin\beta(\Delta)}{R} D_u h$, a positive multiple of the angular derivative. Thus $\mathrm{TCAV}^{\mathbb{H}} = \mathrm{TCAV}$ and is independent of c, and the holonomy is $\theta_{\mathrm{hol}} = \pi - \Delta - 2\beta(\Delta)$, a deterministic function of Δ with zero conditional variance. ■

Appendix C. A released hyperbolic model and an inherited radial signal

The controlled tree of Section 7 sets each node's geodesic radius to its hierarchy depth, so the sharper question is whether a hyperbolic model trained on real data, under its own objective, organizes its radius the same way, and a released hyperbolic model answers this directly. BioCLIP-HC (Tao et al., 2026) is a released hyperbolic vision–language model that lifts the output of a ViT-B/16 CLIP encoder to the Lorentz model, and it ships beside a Euclidean counterpart trained on the same data, which isolates what the hyperbolic geometry contributes on a real model. Its published construction is $x = \exp_o(u)$ at fixed curvature $c = 1$, with the encoder output taken directly as the tangent vector u at the origin and no intervening normalization, following MERU (Desai et al., 2023). The released weights load into a standard ViT-B/16, the encoder output scaled by the model's learned global factor gives u, and that factor and the fixed curvature are positive constants that leave every rank correlation unchanged, so the reconstruction reproduces the published embedding geometry for the rank tests here. A subtree of the animal taxonomy with 70 nodes and depths 1 through 7 fixes the ground truth, each node embedded from its taxonomic name, so a node's taxonomic rank is its known hierarchy depth.

The text encoder, shared by both models, sets the radial order. Its correlation with taxonomic depth is nearly the same in the hyperbolic model and the Euclidean-space model, near -0.8 for the full taxonomic string and still present when each name is embedded alone (Table 2), and its sign is opposite to the entailment convention that would place general concepts near the origin.

On the hyperbolic checkpoint the loop holonomy tracks taxonomic depth, with Spearman between -0.47 and -0.86 across the three text forms (Table 2). The sign is negative because this model's radius shrinks with depth, opposite to the depth-increasing radius of the tree in Section 7, so it is the magnitude that compares with the $+0.87$ there. On the Euclidean-space model the same correlation is within 0.07 of zero. The signal is specific to the curved geometry. The shared encoder gives both models a similar radius–depth order, and only the hyperbolic checkpoint turns that order into a depth-tracking holonomy, because its radii sit near the scale where $c \cdot \mathrm{Area}$ still grows with distance, while the Euclidean lift pushes the same features out to radii where the area saturates and the correlation washes out. The holonomy reads a radius the encoder already supplied, and the hyperbolic loss

Table 2: Rank correlations on the BioCLIP-HC taxonomy probe (Spearman, 70 nodes, 1144 concept–node pairs, non-thin triangles). The radius–depth order is nearly identical in the hyperbolic model and the Euclidean-space model, so it comes from the shared encoder. Holonomy tracks taxonomic depth on the hyperbolic checkpoint (−0.47 to −0.86). On the Euclidean-space model it is within 0.07 of zero, and neither model tracks fine graph separation.

	radius vs. depth		holonomy vs. depth		holonomy vs. sep.
text form	hyperbolic	Euclidean	hyperbolic	Euclidean	hyperbolic
taxonomic lineage	−0.80	−0.79	−0.86	−0.01	−0.01
name only	−0.57	−0.62	−0.47	+0.07	+0.03
templated name	−0.73	−0.75	−0.62	+0.08	−0.00

adds no hierarchy of its own. The depth it tracks is coarse, and the finer pairwise separation, the taxonomic shortest-path distance between the two concepts of a pair, stays within 0.03 of zero on the hyperbolic model in every text form. That separation null is not an artifact of near-collinear points, since almost every triangle is non-thin (over 99% of the 1144 pairs) and the encoder does not arrange concepts by pairwise taxonomic distance. A from-scratch objective with an entailment term places generality on the radius by training, and it would sharpen this coarse depth signal into one that also resolves separation.

Appendix D. Baseline comparison of attribution probes

On the synthetic tree of Section 7, the same embedding as Table 1, every probe in the family is scored the same way, by the rank correlation of its magnitude with combined hierarchy depth over all concept–node pairs, across curvature and five seeds (Table 3). The loop holonomy tracks depth most strongly at every curvature and rises with c, matching the $n = 8$ native arm of Table 1. The loop-free comparator tracks depth as well, more weakly, which confirms that it reaches the same first-order Riemann invariant as the holonomy without a loop (Theorem 5). The single-concept scores, the geodesic-local s_{local} and the Euclidean s_{euclid}, correlate far more weakly and carry no zero-baseline curvature term. The first-order temporal probe has the opposite sign because its magnitude $\sqrt{c}\,\coth(\sqrt{c}\,d)\sin\theta$ falls with distance.

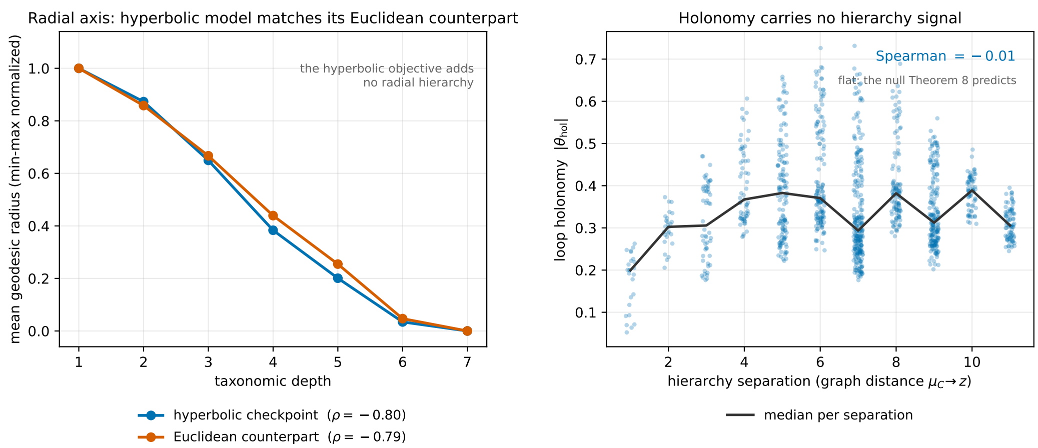

Figure 4: The released BioCLIP-HC checkpoints on the animal taxonomy. Left: mean geodesic radius against taxonomic depth, each curve min-max normalized, for the hyperbolic model and the Euclidean-space model, which trace the same curve, so the radial order comes from the shared encoder. Right: loop holonomy against combined taxonomic depth. Holonomy falls with depth on the hyperbolic checkpoint (Table 2) and stays flat on the Euclidean-space model, so the curved geometry is what turns the encoder's inherited radius into a depth-tracking signal.

Table 3: Attribution probes on the synthetic tree ($b = 3$, depth 4, 121 nodes, $n = 8$, 4641 concept–node pairs). Each cell is the Spearman correlation between the probe magnitude and combined hierarchy depth, mean over five seeds (standard deviation ≤ 0.05). The second-order curvature probes (holonomy and the loop-free comparator) track depth; the single-concept scores are markedly weaker; the first-order temporal magnitude falls with depth.

	second-order		first-order	single-concept	
c	holonomy	comparator	temporal	s_{local}	s_{euclid}
0.25	+0.75	+0.48	−0.10	+0.15	+0.13
0.5	+0.79	+0.51	−0.13	+0.23	+0.16
1	+0.84	+0.58	−0.16	+0.35	+0.20
2	+0.87	+0.62	−0.25	+0.45	+0.26