CURVSPREAD: CRITICAL CONFINEMENT FOR HYPERBOLIC GRAPH SELF-SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dispersion prevents collapse in self-supervised learning, but hyperbolic space has no finite radius at which spreading naturally stops. We show that whether spreading stops depends on the reference measure of the spreading law. A *wrapped* law, specified in tangent coordinates and pushed through the exponential map as in tangent-space isotropy objectives, always inherits a finite radial scale from the tangent law. An *intrinsic* law, specified against hyperbolic volume under linear radial confinement, has a finite equilibrium exactly when the confinement coefficient exceeds the volume-growth rate $(d-1)\sqrt{\kappa}$. The ratio of the two is thus a single critical coordinate across dimensions and curvatures, with the boundary at one. We introduce CURVSPREAD, which adds the exponential-map Jacobian to a local-spacing entropy surrogate and calibrates confinement to this rate, so that the confinement strength is set by the geometry rather than tuned per dataset. We prove that its finite-sample objective, before clipping, has exactly this threshold for every sample size. Across 25 dimension–curvature settings, the operational midpoints of the trained-encoder response are centered on the predicted boundary (median 1.00), while the sharpness of that response varies with geometry. One calibration, selected on three development graphs, is competitive on ten node-classification benchmarks.

1 INTRODUCTION

Self-supervised representation learning balances alignment, which pulls views of the same instance together, against dispersion, which prevents representations from collapsing onto a small region or a few directions. In a common Euclidean formulation, dispersion acts on the unit hypersphere, a compact space on which spreading has a finite extent (Wang & Isola, 2020). Hyperbolic space has no analogous endpoint: the volume of a geodesic ball of radius r in $\mathbb{H}^d_\kappa$ (dimension d, curvature $-\kappa$) grows as $e^{(d-1)\sqrt{\kappa}\,r}$, so an intrinsic objective that keeps rewarding separation has no built-in radial scale at which spreading should stop.

Hyperbolic graph self-supervised methods control this outward pressure in different ways. Hyperbolic contrastive methods (Liu et al., 2021; Yang et al., 2022) spread representations by pairwise repulsion, and the realized scale can then become sensitive to clipping, since finite-precision representations degrade at large radii (Mishne et al., 2023) and motivate clipped models (Guo et al., 2022). Our closest prior, HyperGCL (Zhang et al., 2025), targets the hyperbolic analogue of dimensional collapse (Jing et al., 2022) and replaces uniformity by an isotropy objective on tangent coordinates at the origin, inducing an *outer shell* near the boundary. That construction is wrapped in the sense made precise below: its radial scale is set by the tangent prior. Boundary-seeking behaviour is therefore not our contribution; the open question is what determines whether *intrinsic* spreading admits a finite equilibrium, and at what radial scale.

The two kinds of spreading are not different classes of measures: because $\exp_c^\kappa$ is a diffeomorphism, every absolutely continuous law on $\mathbb{H}^d_\kappa$ can be written either as the pushforward of a tangent density or as a density against Riemannian volume, the isometry-invariant reference measure. The two constructions differ in what is specified. A *wrapped* construction specifies a proper tangent density and maps it through $Z = \exp_c^\kappa(Y)$ (Nagano et al., 2019), so properness and radial scale are inherited. An *intrinsic* construction specifies an energy against Riemannian volume, as in $q_{\beta,c}(z) \propto e^{-\beta d_\kappa(c,z)}$; in tangent coordinates that law carries the Jacobian of the exponential map, so normalizability

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.

depends on confinement versus volume growth. How strong must confinement be when dimension and curvature jointly set the available volume, and where do existing objectives sit relative to the resulting boundary?

The answer is governed by the volume-growth rate $h_\kappa = (d-1)\sqrt{\kappa}$ (Figure 1): under linear confinement of strength β the ratio $\chi = \beta/h_\kappa$ decides existence, with a finite equilibrium for $\chi > 1$ whose mode diverges as $\chi \downarrow 1$ and none for $\chi \le 1$. The dependence enters a learnable objective through the exponential-map Jacobian: a local-spacing estimator on tangent coordinates targets the entropy of Y relative to Lebesgue measure on $T_c\mathbb{H}^d_\kappa$, and adding the mean log-Jacobian changes the reference measure to Riemannian volume, $H_{\mathbb{H}}(Z) = H(Y) + \mathbb{E}\log J_\kappa(\|Y\|_c)$; without it the critical scale does not appear.

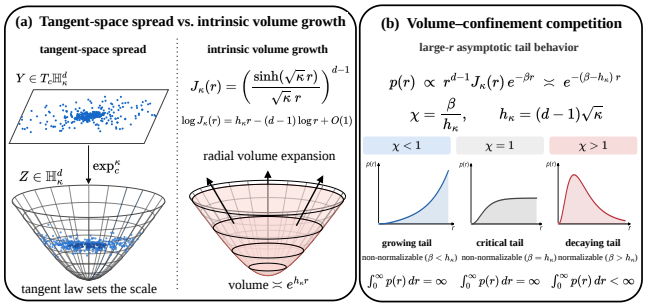


Figure 1: **Tangent spread versus intrinsic volume.** (a) A wrapped law inherits its scale from tangent coordinates; the exponential-map Jacobian has exponential rate h_κ, with $\log J_\kappa(r) = h_\kappa r - (d-1)\log r + O(1)$. (b) Against confinement $e^{-\beta r}$ the radial tail is $e^{-(\beta - h_\kappa)r}$ up to a constant factor, normalizable only for $\chi = \beta/h_\kappa > 1$.

We operationalize this principle in a regularizer with three components, CURVSPREAD (Curvature-Calibrated Representation Spreading): (i) a KoLeo-style k-nearest-neighbor spacing surrogate (Sablayrolles et al., 2019; Oquab et al., 2024); (ii) the exact exponential-map Jacobian, which refers the surrogate to Riemannian volume; and (iii) linear radial confinement with coefficient $\beta = (1+\delta)h_\kappa$, evaluated at a stop-gradient of the current curvature so that $\chi = 1 + \delta$ at every step without a per-dataset search over β. It attaches to a conventional cross-view alignment loss, introduces no new augmentation, critic or encoder, and is scoped to $0 < \kappa_{\min} \le \kappa \le \kappa_{\max}$.

The threshold itself is elementary; what makes it useful is that it survives the passage from the population law to the finite-sample objective, while the statements concern the objective rather than the training trajectory. It also answers where existing objectives sit: pairwise uniformity has no finite stopping scale at all, wrapped objectives such as HyperGCL lie off this axis because their radial scale is fixed by the tangent prior, and a fixed β can cross the boundary as d or κ changes, whereas CURVSPREAD stays at $\chi = 1 + \delta$ even as curvature is learned. Geometry fixes the scale inside the spreading term, not its trade-off with alignment, which is selected once on development graphs.

Direct sampling of the two population laws confirms the wrapped–intrinsic contrast: the wrapped mode grows as $\sigma\sqrt{d-1}$, whereas the intrinsic mode is dimension-free at fixed χ (Section 5.4). The analysis covers intrinsic entropy under linear confinement, not InfoNCE or general pairwise critics, whose outward pressure behaves differently (Appendix F.1). We use training only as a consistency check, on a theory-centered grid in which the spread in raw β is not separate evidence. The contribution is a chain whose links can be checked separately:

- **A critical coordinate.** Unconstrained intrinsic spreading has no finite stopping scale (Proposition 1); under linear confinement the intrinsic law is proper exactly when $\chi = \beta/h_\kappa > 1$ (Proposition 2), a boundary at the same χ for every dimension and curvature.
- **A finite-sample guarantee.** The finite-sample CURVSPREAD objective has exactly this threshold for every sample size, with a configuration-uniform bound that holds at the moving centroid (Proposition 3); the coefficient transfers, but the equilibrium radius does not.
- **Consistency in trained encoders.** Under clipping and a finite horizon, which the theory does not cover, operational response midpoints cluster near $\chi = 1$ across 25 dimension–curvature settings, while their sharpness varies with geometry (Sections 5.2–5.3).
- **Downstream evidence.** One calibration, used unchanged, is competitive on ten benchmarks and the validation optimum on a held-out graph; removing the Jacobian costs about two thirds of the spreading gain and per-dataset tuning of β about half (Sections 5.5–5.6).

2 Hyperbolic Space and Two Ways to Specify Spread

We work on the d-dimensional hyperbolic space $\mathbb{H}^d_\kappa$ of constant sectional curvature $-\kappa$, with $d \geq 2$ and $\kappa > 0$. Let d_κ denote the geodesic distance, and let $\exp^\kappa_c$ and $\log^\kappa_c$ denote the exponential and logarithmic maps at a reference point c, with $\|\cdot\|_c$ the norm induced by the metric at c. In geodesic polar coordinates (r, ω) around c, the exponential map preserves radial length and the Riemannian volume element is explicit:

$$d_\kappa(c, \exp^\kappa_c(y)) = \|y\|_c, \qquad \mathrm{dvol}_\kappa = \left(\frac{\sinh(\sqrt{\kappa}\, r)}{\sqrt{\kappa}}\right)^{d-1} dr\, d\omega. \tag{1}$$

Its radial factor grows as $\exp((d-1)\sqrt{\kappa}\, r)$; we call $h_\kappa = (d-1)\sqrt{\kappa}$ the volume-growth rate.

Two constructions. A wrapped law is specified by a proper density q_T with respect to Lebesgue measure on $T_c\mathbb{H}^d_\kappa$: draw $Y \sim q_T$ and set $Z = \exp^\kappa_c(Y)$. Its density with respect to Riemannian volume is

$$q_{\mathrm{wrap}}(z) = \frac{q_T(\log^\kappa_c z)}{J_\kappa(\|\log^\kappa_c z\|_c)}, \qquad J_\kappa(r) = \left(\frac{\sinh(\sqrt{\kappa}\, r)}{\sqrt{\kappa}\, r}\right)^{d-1}. \tag{2}$$

Since $\mathrm{dvol}_\kappa = J_\kappa(\|y\|_c)\, dy$ under $z = \exp^\kappa_c(y)$, the factor J_κ^{-1} offsets the volume element, so properness and the radial law are both inherited from q_T: $d_\kappa(c, Z) = \|Y\|_c$, and an isotropic Gaussian of scale σ has radial mode $\sigma\sqrt{d-1}$ for every κ. The outer-shell isotropy of HyperGCL (Zhang et al., 2025) is a wrapped construction in this sense. An intrinsic, or volume-specified, law instead specifies its energy directly with respect to dvol_κ; the radial family relevant here is

$$q_{\beta,c}(z) = \frac{1}{Z_{\beta,c}} \exp\left[-\beta d_\kappa(c, z)\right], \qquad Z_{\beta,c} = \int_{\mathbb{H}^d_\kappa} e^{-\beta d_\kappa(c,z)}\, \mathrm{dvol}_\kappa(z). \tag{3}$$

Same measures, different specification. The two constructions describe the same absolutely continuous laws but differ in the reference measure against which the energy is specified: an energy E against dvol_κ is the tangent energy $E \circ \exp^\kappa_c - \log J_\kappa(\|\cdot\|_c)$ against Lebesgue measure. For $q_{\beta,c}$ the Jacobian therefore enters the tangent density, and for $\beta \leq h_\kappa$ no law of that shape exists in any coordinates, whereas the tangent law $\propto e^{-\beta\|y\|_c}\, dy$ is proper for every $\beta > 0$ (Appendix F.3). The reference measures also differ geometrically: dvol_κ is isometry-invariant, whereas the pushforward of tangent Lebesgue measure depends on the base point c. Only the volume-specified construction is subject to the threshold of Section 3.

3 The Critical Intrinsic Regime

Two results govern intrinsic spreading on $\mathbb{H}^d_\kappa$: unconstrained uniformity has no finite scale, whereas intrinsic entropy under linear confinement has a threshold at h_κ. Proofs are in Appendices F.1 and F.2.

Unconstrained intrinsic spreading. For $t > 0$ and a Borel probability measure μ on $\mathbb{H}^d_\kappa$, let

$$U_t(\mu) = \log I_t(\mu), \qquad I_t(\mu) = \iint \exp\left(-t\, d_\kappa(x, y)^2\right) d\mu(x)\, d\mu(y), \tag{4}$$

the intrinsic counterpart of the hyperspherical uniformity term (Wang & Isola, 2020).

Proposition 1 (Escape of unconstrained intrinsic spreading)**.** *Over probability measures on* $\mathbb{H}^d_\kappa$, $\inf_\mu U_t(\mu) = -\infty$, *and the infimum is not attained:* $U_t(\mu) > -\infty$ *for every* μ.

The infimum is approached by spreading normalized volume over balls of growing radius (Appendix F.1). The escape is a volume-of-space phenomenon: the same dilation applies in Euclidean space, and hyperbolic geometry changes only the threshold below. The pairwise term selects no finite scale, so any observed scale comes from other losses, architecture, optimization, or numerical boundaries.

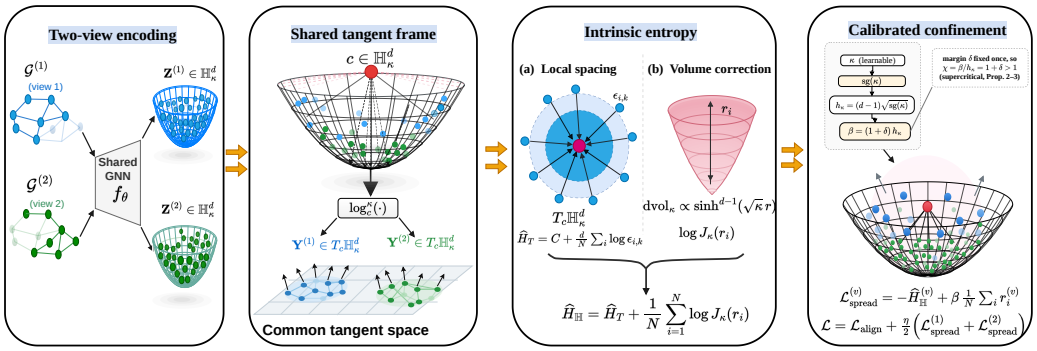


Figure 2: **CURVSPREAD.** Two augmented views $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}$ are encoded by a shared GNN f_θ into hyperbolic representations $Z^{(1)}, Z^{(2)}$, mapped by $\log_c^\kappa$ into one tangent frame at the Lorentz centroid c with radii preserved. Local-spacing entropy from nearest-neighbor distances $\epsilon_{i,k}$ (a) is combined with the exponential-map Jacobian J_κ (b), so the surrogate $\widehat{H}_\mathbb{H}$ targets intrinsic Riemannian entropy rather than tangent-space entropy. Linear radial confinement is calibrated through $\kappa \to \mathrm{sg}(\kappa) \to h_\kappa \to \beta = (1+\delta)h_\kappa$, and the resulting regularizer is added to the alignment loss.

Intrinsic entropy with linear confinement. For a reference point c and $\beta > 0$, define, over probability densities ρ with respect to dvol_κ for which both integrals are finite,

$$F_{\beta,c}(\rho) = \int \rho \log \rho \, \mathrm{dvol}_\kappa + \beta \int d_\kappa(c,z)\, \rho(z)\, \mathrm{dvol}_\kappa(z). \tag{5}$$

Proposition 2 (Critical intrinsic confinement)**.** *Let $d \geq 2$, and let $Z_{\beta,c}$ and $F_{\beta,c}$ be as in (3) and (5). Then $Z_{\beta,c} < \infty$ if and only if $\beta > h_\kappa$. For $\beta > h_\kappa$, $q_{\beta,c}$ is the unique minimizer of $F_{\beta,c}$, with minimum $-\log Z_{\beta,c}$; for $\beta \leq h_\kappa$, $F_{\beta,c}$ is unbounded below.*

Whenever $Z_{\beta,c} < \infty$, the Gibbs identity $F_{\beta,c}(\rho) = D_{\mathrm{KL}}(\rho \,\|\, q_{\beta,c}) - \log Z_{\beta,c}$ identifies $q_{\beta,c}$ as the unique minimizer, and the polar integrand $e^{-\beta r} \sinh^{d-1}(\sqrt{\kappa}\, r) \asymp e^{-(\beta - h_\kappa)r}$ as $r \to \infty$ gives $Z_{\beta,c} < \infty$ exactly when $\beta > h_\kappa$; the case $\beta \leq h_\kappa$ is treated in Appendix F.2.

The critical scale depends on both dimension and curvature, so a fixed β may cross the boundary as either changes. We therefore normalize by the volume-growth rate:

$$\chi = \frac{\beta}{h_\kappa} = \frac{\beta}{(d-1)\sqrt{\kappa}}. \tag{6}$$

Then $\chi > 1$ is the finite-equilibrium regime, where the radial density $p(r) \propto e^{-\beta r} \sinh^{d-1}(\sqrt{\kappa}\, r)$ has mode $r^\star = \kappa^{-1/2} \operatorname{arccoth}(\chi)$, diverging as $\chi \downarrow 1$ like $(2\sqrt{\kappa})^{-1} \log(2/(\chi - 1))$; for $\chi \leq 1$ the density is not normalizable and no mode exists (Figure 1b). These are properties of the equilibrium of $F_{\beta,c}$, not of any training trajectory; they concern intrinsic entropy under linear confinement, not InfoNCE or general pairwise critics (Appendix F.1).

4 CURVSPREAD

CURVSPREAD is a spreading regularizer for two-view hyperbolic graph self-supervision. It replaces pairwise repulsion by a finite-sample surrogate of intrinsic Riemannian entropy, confined linearly at a scale calibrated to the volume-growth rate (Figure 2).

4.1 INTRINSIC ENTROPY AND CRITICAL SCALING

A shared graph encoder maps two augmented views of the input graph to $z_i^{(v)} = \exp_o^\kappa(h_i^{(v)})$ with $h_i^{(v)} = f_\theta(\mathcal{G}^{(v)})_i \in \mathbb{R}^d$, $v \in \{1,2\}$, $i = 1, \ldots, N$, with no projector head and no normalization of h, since radial scale is what the regularizer controls; clipping (Section 5.1) is applied after the lift as a numerical safeguard. Both views are expressed in one tangent frame at their common Lorentz centroid

c. Following the Kozachenko–Leonenko estimator (Kozachenko & Leonenko, 1987) underlying KoLeo-style spreading (Sablayrolles et al., 2019; Oquab et al., 2024), we use

$$y_i^{(v)} = \log_c^\kappa\left(z_i^{(v)}\right), \quad r_i^{(v)} = \|y_i^{(v)}\|_c, \quad \widehat{H}_T^{(v)} = C + \frac{d}{N}\sum_{i=1}^{N} \log \epsilon_{i,k}^{(v)}, \tag{7}$$

where $\epsilon_{i,k}^{(v)}$ is the distance from $y_i^{(v)}$ to its k-th nearest neighbor within the view and C depends only on N, k and d. Its population target is $H(Y)$ relative to Lebesgue measure on $T_c\mathbb{H}^d_\kappa$, not the Riemannian entropy of Z. The two entropies of the same law are related by $H_{\mathbb{H}}(Z) = H(Y) + \mathbb{E}[\log J_\kappa(R)]$ with $R = \|Y\|_c$, which yields the intrinsic surrogate

$$\widehat{H}_{\mathbb{H}}^{(v)} = \widehat{H}_T^{(v)} + \frac{1}{N}\sum_{i=1}^{N} \log J_\kappa\left(r_i^{(v)}\right). \tag{8}$$

Three objects. We keep apart (A) the population free energy $F_{\beta,c}$ of Proposition 2; (B) the finite-sample spreading functional of one view at fixed β, κ and reference point c,

$$\mathcal{L}_\beta(Y) = -\widehat{H}_{\mathbb{H}}(Y) + \frac{\beta}{N}\sum_{i=1}^{N} r_i = -\widehat{H}_T(Y) + \frac{1}{N}\sum_{i=1}^{N}\left[\beta r_i - \log J_\kappa(r_i)\right]; \tag{9}$$

and (C) training, in which $\mathcal{L}_\beta$ enters Eq. (13) with alignment, a moving centroid $c(Z)$, an encoder, stochastic gradients and learnable curvature. CURVSPREAD is not a discretization of $F_{\beta,c}$, so the critical coefficient of (B) must be shown for (B) itself.

Proposition 3 (Finite-sample critical coefficient)**.** *Let $d \geq 2$, $\kappa > 0$, $1 \leq k < N$ and $\epsilon_{\text{num}} \geq 0$; let $\widehat{H}_T$ in (9) use $\log(\epsilon_{i,k} + \epsilon_{\text{num}})$, with $\mathcal{L}_\beta = +\infty$ when some $\epsilon_{i,k} + \epsilon_{\text{num}} = 0$.*

(i) If $\beta > h_\kappa$, then for every configuration $Y \in (\mathbb{R}^d)^N$

$$\mathcal{L}_\beta(Y) \;\geq\; B_\beta := -C + d - d\log\frac{2dN}{\beta - h_\kappa} - \frac{(\beta - h_\kappa)\,\epsilon_{\text{num}}}{N},$$

$\mathcal{L}_\beta(Y) \to \infty$ as $\max_i r_i \to \infty$, and $\mathcal{L}_\beta$ attains its minimum.

(ii) If $\beta \leq h_\kappa$, then $\mathcal{L}_\beta(RY) \to -\infty$ as $R \to \infty$ for every configuration of distinct points; hence $\inf \mathcal{L}_\beta = -\infty$ and no minimizer exists.

(iii) At fixed κ and $\eta > 0$, let $\Phi_\beta(Z) = \mathcal{L}_{\text{align}}(Z) + \frac{\eta}{2}\sum_v \mathcal{L}_\beta(Y^{(v)})$ be the objective of Eq. (13) as a function of the $2N$ embeddings, with each $Y^{(v)}$ taken in the tangent space at the Lorentz centroid $c(Z)$ of both views. If $\beta > h_\kappa$, then $\Phi_\beta \geq \eta B_\beta$ and Φ_β attains its minimum; if $\beta \leq h_\kappa$, then $\inf \Phi_\beta = -\infty$.

Scope of Proposition 3. The proposition is global over finite configurations. Its proof (Appendix F.5) uses only $0 \leq \log J_\kappa(r) \leq h_\kappa r$ and $\epsilon_{i,k} \leq 2\max_j r_j$, never the accuracy of the KoLeo estimate, so the threshold is exactly h_κ for every N and k, whatever the estimator bias. What depends on N and k is the explicit bound B_β, which degrades only logarithmically as $\beta \downarrow h_\kappa$, through the term $-d\log(2dN/(\beta - h_\kappa))$; the margin δ of Section 4.2 therefore keeps the objective a fixed distance, in χ, from the regime where this bound is lost. As a statement about the functional it covers anisotropic, clustered, low-rank and nearly collapsed configurations alike, and the bound in (i) is uniform in the reference point, which carries it to the moving centroid in (iii). The dimension d is that of the ambient hyperbolic reference volume; effective-dimension qualifications are discussed in Appendix D.2. Proposition 3 says nothing about (C): it concerns the objective as a function of the embeddings, not the encoder, the stochastic-gradient trajectory or the point training reaches. We therefore test only where the operational training response lies relative to $\chi = 1$ (Section 5.2), without treating the observed response as a theorem-level phase transition.

The dilation family. Under $y_i \mapsto Ry_i$ at a fixed reference point and curvature, every spacing scales as $\epsilon_{i,k} \mapsto R\epsilon_{i,k}$, so for $\epsilon_{\text{num}} = 0$

$$\widehat{H}_T(RY) = \widehat{H}_T(Y) + d\log R \tag{10}$$

exactly, whatever the estimator bias. With $\bar{r} = N^{-1}\sum_i r_i$ and all $r_i > 0$, the large-radius form of the Jacobian then gives, with c_d independent of R,

$$\mathcal{L}_\beta(RY) = (\beta - h_\kappa)\, R\bar{r} - \log R + c_d + o(1), \qquad R \to \infty. \tag{11}$$

The linear term changes sign exactly at $\beta = h_\kappa$. Eq. (11) is an intuitive one-dimensional section of Proposition 3, not its proof, and it does not determine the finite-sample equilibrium radius; the derivation is in Appendix F.5. For $\beta > h_\kappa$ its leading terms have the stationary point $R^\star = 1/((\beta - h_\kappa)\bar{r})$, which diverges as $\beta \downarrow h_\kappa$. At $\beta = h_\kappa$ the linear term vanishes and the residual $-\log R$ still decreases without bound, so the boundary case $\chi = 1$ is itself subcritical: the objective escapes, but only logarithmically.

The coefficient transfers; the radius does not. Proposition 3 transfers the critical coefficient from (A) to (B), not the population equilibrium radius. The population mode $r^\star = \kappa^{-1/2}\mathrm{arccoth}(\chi)$ is d-independent and only logarithmic in $\delta = \chi - 1$, whereas the leading dilation scale $R^\star\bar{r} = 1/(h_\kappa\delta)$ matches the exponential tail scale of $q_{\beta,c}$; the two are different quantities (Appendix D.2). The radius reached in training is therefore measured, not theoretically predicted.

What the theory predicts for the experiments. Bound (i) holds for every realized subset of $M < N$ nodes, and, under learnable curvature, for the frozen objective of every step with a bound uniform over $[\kappa_{\min}, \kappa_{\max}]$, so neither subsampling nor moving curvature erodes the supercritical side (Appendix D.2). Clipping confines radii about the origin to a compact set, on which the objective attains its minimum for every β; below the threshold the proposition then predicts no finite interior scale, but not that minimizers lie on the clipping boundary. The observable we test is therefore boundary proximity in trained encoders, $Q_{0.95}(r^{(o)})/R_{\mathrm{clip}}$, rather than a predicted radius.

4.2 Geometry-Calibrated Objective

Proposition 2 fixes the leading scale of the confinement coefficient, and CurvSpread sets

$$\beta = (1+\delta)(d-1)\sqrt{\mathrm{sg}(\kappa)}, \qquad \delta > 0, \tag{12}$$

with sg the stop-gradient operator. The coefficient is evaluated at the current curvature, so $\chi = 1 + \delta$ holds exactly in the forward objective, for fixed and learnable curvature alike, and every step's spreading functional satisfies Proposition 3(i), with a lower bound uniform over $\kappa \in [\kappa_{\min}, \kappa_{\max}]$ (Appendix F.5). The stop-gradient is needed because β depends on κ: without it the confinement term $\beta(\kappa)\,\bar{r}$ would push gradient into κ through the coefficient itself, a feedback path absent from Proposition 2, where β is a parameter rather than a function of the state. It blocks only that path; curvature still receives gradient through the embeddings, the logarithmic map, the radii and the Jacobian. At step t the loss is thus a well-defined objective with coefficient frozen at β_t and the update is its exact gradient, but the family moves with κ_t: the scheme is stepwise and does not optimize one time-independent objective, and no convergence statement follows. With fixed curvature the distinction disappears, since β is a constant and the objective is one fixed surrogate. The margin δ is one global constant, fixed once rather than tuned per dataset, and curvature is learnable within $[\kappa_{\min}, \kappa_{\max}]$ (Appendix G.1). Geometry fixes the scale internal to the spreading term, not the trade-off η below, which is selected once. The spreading term is computed for each view and added to a conventional cross-view hyperbolic alignment loss, for instance $\mathcal{L}_{\mathrm{align}} = N^{-1}\sum_i d_\kappa(z_i^{(1)}, z_i^{(2)})^2$,

$$\mathcal{L}^{(v)}_{\mathrm{spread}} = -\widehat{H}^{(v)}_{\mathbb{H}} + \beta\frac{1}{N}\sum_{i=1}^{N} r_i^{(v)}, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{align}} + \frac{\eta}{2}\left(\mathcal{L}^{(1)}_{\mathrm{spread}} + \mathcal{L}^{(2)}_{\mathrm{spread}}\right). \tag{13}$$

The alignment term does not confine the radius: it vanishes whenever $z_i^{(1)} = z_i^{(2)}$, however large $d_\kappa(c, z_i)$ is, and at fixed κ it neither removes nor shifts the threshold (Proposition 3(iii)). The shared center c is a gauge choice through which gradients flow (Appendix E). Whenever $N > M$, the spreading term and its center are estimated on one random node subset S of size M shared by both views (Algorithm 1).

Why linear confinement. Quadratic confinement βr^2 is coercive for every coefficient, because a Gaussian tail dominates exponential volume growth; it therefore has no threshold and nothing

for geometry to calibrate. Linear confinement is the critical asymptotic order: it competes with volume growth at the same exponential rate, which is why its coefficient must be set relative to h_κ. The resulting law $q_{\beta,c}$ is also the maximum-entropy intrinsic law under a mean-radius constraint and is heavier-tailed than its quadratic counterpart; whether that heavier tail helps hierarchical representations is an empirical question, and the quadratic variant is compared in Appendix I.2.

5 Experiments

5.1 Protocol

Two evaluation tracks are used. Track A covers eight standard graphs (Yang et al., 2016; Shchur et al., 2018; Hu et al., 2020), with ten predetermined paired runs on the seven small and medium graphs and the official OGB split with five model seeds on `ogbn-arxiv`; our reproductions keep each method's official architecture but share splits, linear probe and validation budget, with literature-reported cells listed in Appendix H.3. Track B compares with the nearest prior, HyperGCL (Zhang et al., 2025), on five graphs under its published splits and against its author-reported scores, and is therefore descriptive rather than paired. The controlled training runs of Sections 5.2–5.4 share one encoder, augmentation, optimizer, schedule and seed policy. Throughout, the downstream metric is node-classification accuracy of a frozen encoder, model selection reads validation splits only, $d = 64$ in Track A and the auxiliary controlled arms, and CurvSpread uses one global configuration selected once on Cora, Amazon Photo and Coauthor CS and frozen for every entry (Appendix H.3). Clipping is a Poincaré projection with margin ϵ and radius $R_{\text{clip}} = \kappa^{-1/2} \log((2-\epsilon)/\epsilon)$ about the origin (Appendix E); the full protocol and numerical audits are in Appendices H.2 and G.2.

5.2 Training Response Relative to the Critical Boundary

Propositions 1–3 characterize regimes of the objective, not the encoder trajectory. We therefore use training only as an operational consistency check. The controlled Pairwise-U, wrapped-Gaussian and intrinsic-linear arms, together with clipping and precision diagnostics, are specified in Appendix H.2; they serve as auxiliary controls rather than as evidence for a training-time phase transition. The main training diagnostic is the terminal origin-referenced boundary proximity $Q_{0.95}(r^{(o)})/R_{\text{clip}}$.

Operational midpoints in χ. Proposition 3 places the boundary of the unclipped objective at $\chi = 1$, whereas training adds an encoder, finite optimization and clipping. We therefore summarize each terminal response curve by an operational plateau-midpoint location $\hat{\chi}_c$ (Appendix H.2) and ask where that descriptor falls relative to $\chi = 1$ across a 5×5 grid of d from 16 to 256 and κ from 0.1 to 2. The grid is theory-centered, $\beta = \chi h_\kappa$, with $\chi \in [0.7, 1.5]$ in steps of 0.05; interpolated midpoint locations can fall between grid values. A midpoint is only a location descriptor: smooth or weak responses can have one without exhibiting a sharp change point. Across the 25 cells $\hat{\chi}_c$ has median 1.00, interquartile range $[0.98, 1.03]$, and range $[0.95, 1.07]$, with no visible trend in d or κ (Table 9). These are point estimates; the bootstrap procedure is specified in Appendix H.2. In raw β the corresponding locations span a 78-fold range because h_κ varies from 4.7 to 361 on this theory-centered grid; this spread is therefore not separate evidence for the normalization.

5.3 Geometry-Normalized Response Profiles

On $(d, \kappa) = (16, 0.1)$, $(64, 1)$ and $(256, 2)$, with $h_\kappa = 4.7$, 63 and 361, we trace the response $Q_{0.95}(r^{(o)})/R_{\text{clip}}$ over $\chi \in \{0.80, 0.85, \ldots, 1.20\}$ under the protocol of the 5×5 grid (Disease, seeds 301–303). The curves do not coincide in χ, and their shape remains geometry-dependent. The strongest response change occurs at the largest h_κ and is concentrated near $\chi = 1$: the response leaves the clipping boundary after $\chi = 0.90$ and falls to 0.92, 0.59 and 0.43 at $\chi = 0.95$, 1.00 and 1.05. At $h_\kappa = 63$ it declines gradually and levels off beyond $\chi \approx 1.05$; at $h_\kappa = 4.7$ it is smooth (0.265 to 0.235) with no sharp change point, so the estimate of Table 9 there marks a plateau-midpoint crossing rather than a sharp transition. The center-referenced response $Q_{0.95}(r)/R_{\text{clip}}$, which clipping does not pin, shows the same pattern in all three cells (Table 10): at $h_\kappa = 361$ its largest drop is also between $\chi = 0.95$ and 1.00 (0.57 to 0.31), and elsewhere it declines gradually.

Table 1: Frozen-encoder node-classification accuracy (%), mean ± s.d.; Geo.: E Euclidean, H hyperbolic. CURVSPREAD follows Track A (Section 5.1); protocols and result sources are given in Appendix H.3. Shading marks the best (bold) and second-best mean in each column.

Method	Geo.	Cora	CiteSeer	PubMed	Comp.	Photo	CS	Physics	arXiv
GRACE	E	81.88 ± 0.84	71.13 ± 0.42	80.88 ± 0.13	86.48 ± 0.24	92.20 ± 0.16	92.90 ± 0.27	95.25 ± 0.26	71.51 ± 0.11
BGRL	E	81.86 ± 0.32	72.10 ± 0.31	80.65 ± 0.42	89.21 ± 0.47	92.28 ± 0.44	92.73 ± 0.41	95.31 ± 0.26	71.64 ± 0.12
GraphMAE	E	85.45 ± 0.40	72.48 ± 0.77	85.74 ± 0.14	88.04 ± 0.61	92.73 ± 0.17	93.47 ± 0.04	96.13 ± 0.03	71.86 ± 0.00
GraphECL	E	84.25 ± 0.05	73.15 ± 0.41	82.21 ± 0.05	88.67 ± 0.23	94.22 ± 0.11	93.77 ± 0.72	95.16 ± 0.27	71.75 ± 0.22
S3GCL	E	88.47 ± 0.39	76.31 ± 1.67	87.89 ± 1.23	88.45 ± 1.98	94.31 ± 0.83	92.55 ± 0.89	95.01 ± 0.19	71.36 ± 0.60
GCL-JAM	E	85.54 ± 0.53	73.29 ± 0.39	85.46 ± 0.46	90.69 ± 0.28	94.35 ± 0.21	92.81 ± 0.69	95.01 ± 0.28	71.65 ± 0.12
GALE	E	85.36 ± 0.15	74.58 ± 0.18	85.06 ± 0.27	90.60 ± 0.19	94.51 ± 0.34	93.46 ± 0.28	95.91 ± 0.38	71.16 ± 0.28
EDA-GCL	E	85.63 ± 0.70	75.71 ± 0.88	86.50 ± 0.39	89.37 ± 0.23	93.08 ± 0.46	92.96 ± 0.09	95.70 ± 0.09	71.05 ± 0.12
Dual-Kernel GCCL	E	83.77 ± 1.37	72.68 ± 1.19	82.56 ± 1.85	89.87 ± 0.13	93.86 ± 0.15	94.68 ± 0.14	96.12 ± 0.17	**72.20 ± 0.87**
HGCL	H	82.37 ± 0.47	72.11 ± 0.64	83.14 ± 0.58	88.07 ± 0.20	93.15 ± 0.11	93.08 ± 0.46	95.27 ± 0.22	71.35 ± 0.37
HyperGCL-Normal	H	84.77 ± 0.46	74.43 ± 0.35	85.56 ± 0.21	89.57 ± 0.13	93.77 ± 0.72	94.31 ± 0.83	93.87 ± 0.34	71.14 ± 0.41
HyperGCL-Laplace	H	84.67 ± 0.50	74.37 ± 0.42	85.49 ± 0.22	89.66 ± 0.23	94.42 ± 0.19	94.56 ± 0.83	93.98 ± 0.22	71.34 ± 0.78
CURVSPREAD	H	**88.82 ± 0.12**	**76.86 ± 0.38**	**88.47 ± 0.27**	**91.49 ± 0.28**	**95.16 ± 0.27**	**95.27 ± 0.22**	**96.42 ± 0.05**	71.63 ± 0.12

5.4 WRAPPED VERSUS INTRINSIC RADIAL SCALE

The two constructions of Section 2 predict different population radial laws: the wrapped mode $\sigma\sqrt{d-1}$ grows with dimension and ignores curvature, while the intrinsic mode $\kappa^{-1/2}\,\mathrm{arccoth}(\chi)$ is dimension-free at fixed χ. Because the surrogate inherits the critical coefficient but not the radius (Section 4.1), we sample the population laws directly and report the radial law realized by the trained arms as a measurement, not as a test of Proposition 2. Both procedures are specified in Appendix H.2.

In Figure 3 the sampled modes of the two population laws match their predictions to within 0.02 at every d: the wrapped mode grows as $\sigma\sqrt{d-1}$, from 0.97 to 3.99, while the intrinsic mode stays at 1.20. The trained $d = 64$ overlays are illustrative sanity checks rather than estimates of the population laws: the wrapped control gives 1.82 for seed 301, while the intrinsic arm gives 1.12 as the mean over seeds 301–303, 8% and 7% below their respective population modes.

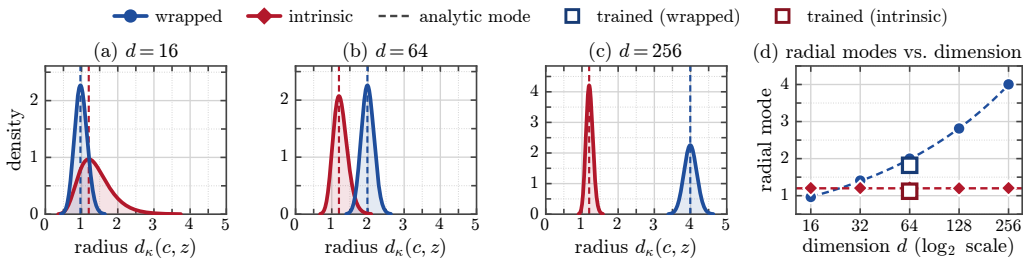


Figure 3: **Wrapped versus intrinsic radial scaling** ($\kappa = 1$, $\sigma = 0.25$, $\chi = 1.2$). Solid curves show the population radial densities, and vertical dashed lines mark their analytic modes, $r^\star_{\mathrm{wrap}} = \sigma\sqrt{d-1}$ and $r^\star_{\mathrm{intrinsic}} = \kappa^{-1/2}\,\mathrm{arccoth}(\chi)$. (a)–(c) Population laws for $d \in \{16, 64, 256\}$. Open squares indicate the trained-encoder modes at $d = 64$. (d) Analytic mode scaling with dimension: the wrapped mode grows as $\sqrt{d-1}$, whereas the intrinsic mode is independent of d at fixed χ.

5.5 DOWNSTREAM REPRESENTATION QUALITY

Benchmark comparison. Table 1 places CURVSPREAD alongside GRACE (Zhu et al., 2020), BGRL (Thakoor et al., 2022), GraphMAE (Hou et al., 2022), GraphECL (Xiao et al., 2024), S3GCL (Wan et al., 2024), GCL-JAM (Yang et al., 2025), GALE (Wang et al., 2025b), EDA-GCL (Chen et al., 2026a), Dual-Kernel GCCL (Chen et al., 2026b), HGCL (Liu et al., 2021) and HyperGCL (Zhang et al., 2025). Some baseline cells are literature-reported under their published protocols and the rest are our Track A reproductions (Appendix H.3 lists which), so the table establishes competitiveness rather than a fully paired ranking. In particular, HyperGCL's Cora/CiteSeer/PubMed entries are literature-reported while its remaining Track A entries are reproduced by us, so we do not claim a single-stack paired comparison against HyperGCL. The wrapped control of Section 5.2 is a mechanistic tangent-specified reference, not a HyperGCL reimplementation. On `ogbn-arxiv`,

CURVSPREAD obtains 71.63, below the column best, 72.20, and we claim no advantage there; it is also the most aggressively subsampled regime ($M/N \approx 0.012$), but whether subsampling explains the weaker result is not established.

Comparison with the nearest prior. Under the published protocol of HyperGCL, whose Normal and Laplace variants are wrapped (Table 7, Appendix I.3), CURVSPREAD has the highest reported mean on all five datasets: by 0.5–0.6 points on the citation graphs and by 0.2 and 1.0 on Disease and Airport. These are not paired runs under one implementation stack, and the Disease margin is within the reported standard deviations, so we treat the comparison as descriptive rather than as a significance claim.

5.6 DEPENDENCE ON THE REGULARIZER

Figure 4 decomposes the method under one encoder, augmentation scheme and alignment term, varying one factor at a time from the frozen configuration and thus, unlike Table 1, without architectural confounding; Pairwise-U and the wrapped Gaussian are auxiliary mechanistic controls (Appendix H.2), not accuracy baselines. The ordering of the variants is the same on all three graphs. *Spreading.* Adding the spreading term to alignment is worth 2.92 points on Cora, 3.10 on Disease and 3.16 on Airport. *Riemannian-volume correction.* Removing the Jacobian costs 2.02, 2.20 and 1.96 points, about two thirds of the spreading gain. *Calibration.* Replacing the calibrated coefficient by a β tuned per dataset costs 1.42, 1.40 and 1.36 points, roughly half of the spreading gain, even though the tuning grid contains $\chi = 1.25$, adjacent to the calibrated $\chi = 1.1$; a constant coefficient does not follow the curvature learned during training, whereas the calibrated one does. Calibration also transfers to a held-out graph: on Disease, which took no part in the development search, the validation optimum over the full (δ, η) and (k, M) grids coincides with the configuration selected on the development graphs (Appendix I.4). *Learnable curvature.* Fixed curvature retains about half of the spreading gain on all three graphs (1.30 of 2.92 points on Cora); learnable curvature supplies the rest. *Stop-gradient.* Coupling β directly to curvature is empirically worse than not learning curvature at all: without the stop-gradient, accuracy falls to 86.60 on Cora, 93.60 on Disease and 93.50 on Airport, 0.60, 0.70 and 0.60 below fixed curvature. This is consistent with the feedback path $\kappa \to \beta$ that Proposition 2 does not contain, since there β is a parameter rather than a function of the state.

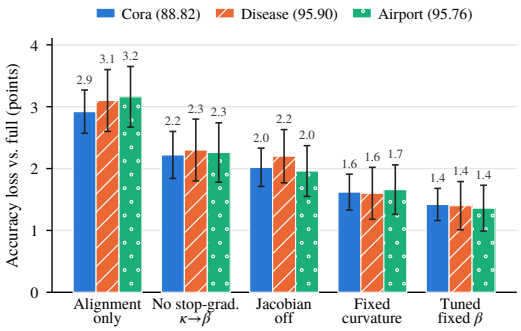


Figure 4: **Controlled ablation.** Bars report the decrease in mean accuracy relative to the full model when one component is modified, while whiskers show ± 1 standard deviation of the corresponding variant accuracy across ten runs. Cora uses Track A; Disease and Airport use Track B splits, with $d = 64$. Full results are in Table 6.

6 CONCLUSION

We separated two ways of specifying spread in hyperbolic space that represent the same laws but differ in the energy's reference measure. Wrapped constructions inherit their radial scale from a tangent law, whereas intrinsic ones, specified against Riemannian volume, admit a finite equilibrium under linear confinement exactly when $\beta > h_\kappa$, and $\chi = \beta/h_\kappa$ places them on one axis across dimensions and curvatures. CURVSPREAD carries this into training via the exponential-map Jacobian and a geometry-set confinement coefficient, and its finite-sample functional provably has the same critical coefficient. In trained encoders, operational plateau-midpoint locations cluster near $\chi \approx 1$ over a $16\times$ range of d and $20\times$ of κ, while the response can be sharp at large h_κ and gradual at small h_κ; this tests consistency with the proved objective boundary rather than proving a training-time phase transition. The theory establishes neither the equilibrium radius nor the training dynamics; a mean-field analysis of the minimizers is the next step.

Ethics Statement

This work studies representation geometry for self-supervised learning on standard public citation, co-purchase, co-authorship, and disease- and airport-network benchmarks. It introduces no new data collection, involves no human subjects, and we are not aware of application-specific risks beyond those already present in graph representation learning generally.

Reproducibility Statement

Propositions 1–3 are proved in Appendices F.1, F.2 and F.5, with the change of variables in Appendix F.4. The method is specified by Eqs. (7)–(13), with Lorentz formulas in Appendix E, the algorithm in Appendix G.1, and numerical evaluations and audits in Appendix G.2. The experimental protocol, operational midpoint estimator, selection rules, exclusion criteria, and validation-only hyperparameter search are specified in Appendices H.2 and H.3. These definitions, numerical conventions, and reported hyperparameter grids form the reproducibility record for the present submission.

AI Use Statement

Generative AI tools were used to assist with language editing and proofreading; the authors reviewed the final manuscript and take responsibility for its content.

References

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Chunchun Chen, Xing Wei, Jiayi Yang, Chenrun Wang, Yiwei Fu, Yuxing Zhang, Xin Sun, Rui Fan, and Wei Ye. Edge self-adversarial augmentation enhances graph contrastive learning against neighborhood inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20005–20013, 2026a. doi: 10.1609/aaai.v40i24.39085.

Xiang Chen, Kun Yue, Wenjie Liu, Zhenyu Zhang, and Liang Duan. Dual-kernel graph community contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14583–14591, 2026b. doi: 10.1609/aaai.v40i17.38476.

Ying Chen, Lianming Zhang, Jiusheng Li, and Pingping Dong. Hyperbolic graph representation learning: Methods, applications and challenges—a survey. *Neurocomputing*, 652:131044, 2025. doi: 10.1016/j.neucom.2025.131044.

Hervé Fournier, Anas Ismail, and Antoine Vigneron. Computing the gromov hyperbolicity of a discrete metric space. *Information Processing Letters*, 115(6–8):576–579, 2015. doi: 10.1016/j.ipl.2015.02.002.

Yali Fu, Jindong Li, Jiahong Liu, Qianli Xing, Qi Wang, and Irwin King. HC-GLAD: Dual hyperbolic contrastive learning for unsupervised graph-level anomaly detection. *Neural Networks*, 202:109009, 2026. doi: 10.1016/j.neunet.2026.109009.

Mina Ghadimi Atigh, Martin Keller-Ressel, and Pascal Mettes. Hyperbolic Busemann learning with ideal prototypes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 103–115, 2021.

Yunhui Guo, Xudong Wang, Yubei Chen, and Stella X. Yu. Clipped hyperbolic classifiers are super-hyperbolic classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, 2022. doi: 10.1109/CVPR52688.2022.00010.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 594–604, 2022. doi: 10.1145/3534678.3539321.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders. In *NIPS Workshop on Bayesian Deep Learning*, 2016. arXiv:1611.07308.

L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, 23(2):95–101, 1987.

Jianqing Liang, Zhiqiang Li, Xinkai Wei, Yuan Liu, and Zhiqiang Wang. ML^2-GCL: Manifold learning inspired lightweight graph contrastive learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37222–37236, 2025.

Jiahong Liu, Menglin Yang, Min Zhou, Shanshan Feng, and Philippe Fournier-Viger. Enhancing hyperbolic graph embeddings via contrastive learning. In *NeurIPS 2021 Workshop on Self-Supervised Learning: Theory and Practice*, 2021. arXiv:2201.08554.

Chang Lu, Chandan K. Reddy, and Yue Ning. Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE Transactions on Cybernetics*, 53(4):2124–2136, 2023. doi: 10.1109/TCYB.2021.3109881.

Emile Mathieu, Charline Le Lan, Chris J. Maddison, Ryota Tomioka, and Yee Whye Teh. Continuous hierarchical representations with Poincaré variational auto-encoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Péter Mernyei and Cătălina Cangea. Wiki-CS: A Wikipedia-based benchmark for graph neural networks. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+)*, 2020. arXiv:2007.02901.

Gal Mishne, Zhengchao Wan, Yusu Wang, and Sheng Yang. The numerical stability of hyperbolic representation learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24925–24949, 2023.

Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4693–4702, 2019.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 14, pp. 849–856, 2001.

Chaoxi Niu, Guansong Pang, and Ling Chen. Affinity uncertainty-based hard negative mining in graph contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9): 11681–11691, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. Adversarially regularized graph autoencoder for graph embedding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2609–2615, 2018. doi: 10.24963/ijcai.2018/362.

Jongmin Park, Seunghoon Han, Won-Yong Shin, and Sungsu Lim. Metapath-based hyperbolic contrastive learning for heterogeneous graph embedding. *arXiv preprint arXiv:2506.16754*, 2025.

Xavier Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006. doi: 10.1007/s10851-006-6228-4.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 701–710, 2014. doi: 10.1145/2623330.2623732.

Zhida Qin, Wentao Cheng, Wenxing Ding, and Gangyi Ding. Hyperbolic graph contrastive learning for collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering*, 37(3):1255–1267, 2025. doi: 10.1109/TKDE.2024.3522960.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *International Conference on Learning Representations (ICLR)*, 2019.

Shruti Saxena, Roshni Chakraborty, and Joydeep Chandra. HCNA: Hyperbolic contrastive learning framework for self-supervised network alignment. *Information Processing & Management*, 59(5):103021, 2022. doi: 10.1016/j.ipm.2022.103021.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. In *Relational Representation Learning Workshop, NeurIPS 2018*, 2018.

Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *International Conference on Learning Representations (ICLR)*, 2021.

Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. In *International Conference on Learning Representations (ICLR)*, 2020.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L. Dyer, Rémi Munos, Petar Veličković, and Michal Valko. Large-scale representation learning on graphs via bootstrapping. In *International Conference on Learning Representations (ICLR)*, 2022.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations (ICLR)*, 2019.

Guancheng Wan, Yijun Tian, Wenke Huang, Nitesh V. Chawla, and Mang Ye. S3GCL: Spectral, swift, spatial graph contrastive learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49973–49990, 2024.

Siman Wang, Xu Chen, Yongjun Jing, and Shuyang Wang. Hyperbolic geometric graph representation learning: A survey. *Computer Science Review*, 58:100800, 2025a. doi: 10.1016/j.cosrev.2025.100800.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939, 2020.

Yejiang Wang, Yuhai Zhao, Zhengkui Wang, Ling Li, Jiapu Wang, Fangting Li, Miaomiao Huang, Shirui Pan, and Xingwei Wang. Equivalence is all: A unified view for self-supervised graph learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65776–65789, 2025b.

Jun Xia, Lirong Wu, Ge Wang, Jintao Chen, and Stan Z. Li. ProGCL: Rethinking hard negative mining in graph contrastive learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24332–24346, 2022.

Teng Xiao, Huaisheng Zhu, Zhiwei Zhang, Zhimeng Guo, Charu C. Aggarwal, Suhang Wang, and Vasant G. Honavar. Efficient contrastive learning for fast and accurate inference on graphs. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54363–54381, 2024.

Xiaoyu Xu, Guansong Pang, Di Wu, and Mingsheng Shang. Joint hyperbolic and Euclidean geometry contrastive graph neural networks. *Information Sciences*, 609:799–815, 2022. doi: 10.1016/j.ins.2022.07.060.

Haoran Yang, Hongxu Chen, Shirui Pan, Lin Li, Philip S. Yu, and Guandong Xu. Dual space graph contrastive learning. In *Proceedings of the ACM Web Conference 2022 (WWW)*, pp. 1238–1247, 2022. doi: 10.1145/3485447.3512211.

Liang Yang, Zhenna Li, Jiaming Zhuo, Jing Liu, Ziyi Ma, Chuan Wang, Zhen Wang, and Xiaochun Cao. Graph contrastive learning with joint spectral augmentation of attribute and topology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21983–21991, 2025. doi: 10.1609/aaai.v39i20.35506.

Menglin Yang, Min Zhou, Rex Ying, Yankai Chen, and Irwin King. Hyperbolic representation learning: Revisiting and advancing. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39639–39659, 2023.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48, 2016.

Yiding Zhang, Xiao Wang, Chuan Shi, Xunqiang Jiang, and Yanfang Ye. Hyperbolic graph attention network. *IEEE Transactions on Big Data*, 8(6):1690–1701, 2022a. doi: 10.1109/TBDATA.2021.3081431.

Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, and Irwin King. COSTA: Covariance-preserving feature augmentation for graph contrastive learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2524–2534, 2022b. doi: 10.1145/3534678.3539425.

Yifei Zhang, Hao Zhu, Menglin Yang, Jiahong Liu, Rex Ying, Irwin King, and Piotr Koniusz. Understanding and mitigating hyperbolic dimensional collapse in graph contrastive learning". In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1984–1995, 2025. doi: 10.1145/3690624.3709249.

Han Zhao, Xu Yang, Zhenru Wang, Erkun Yang, and Cheng Deng. Graph debiased contrastive learning with joint representation clustering. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3434–3440, 2021. doi: 10.24963/ijcai.2021/473.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+)*, 2020.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021 (WWW)*, pp. 2069–2080, 2021. doi: 10.1145/3442381.3449802.

A Use of Large Language Models

Large language models were used during the preparation of this manuscript for (i) language editing and improving the clarity and concision of sentences; (ii) checking the internal consistency of the manuscript, including notation, cross-references, and agreement between text and tables; (iii) assistance with mathematical derivations and proof checking; and (iv) assistance in interpreting and organizing existing analyses and results. The authors reviewed the manuscript and remain fully responsible for all mathematical statements, proofs, numerical values, and scientific claims. No experimental measurement or numerical result reported in this paper was fabricated or generated by a large language model as evidence.

B Notation

Table 2 lists the main symbols. The margin δ of Eq. (12) is distinct from the Gromov quantities $\hat{\delta}_b$, δ_{rel} used only for dataset selection.

Table 2: Main notation.

$\mathbb{H}^d_\kappa$	hyperbolic space, curvature $-\kappa$	h_κ	volume-growth rate $(d-1)\sqrt{\kappa}$
d	representation dimension	β	linear confinement coefficient
c	Lorentz centroid of both views	χ	critical coordinate β/h_κ
d_κ	geodesic distance	δ	calibration margin, $\beta = (1+\delta)h_\kappa$
$\exp_c^\kappa$, $\log_c^\kappa$	exponential / log map at c	$\epsilon_{i,k}$	k-th nearest-neighbor distance
$J_\kappa(r)$	exponential-map Jacobian	$\widehat{H}_T$, $\widehat{H}_{\mathbb{H}}$	tangent / intrinsic entropy estimate
$q_{\beta,c}$	intrinsic Gibbs density	η	weight of the spreading term
$r^\star$	equilibrium radial mode	R_{clip}, $Q_{0.95}$	clipping radius; 95th-pct. radius

C Related Work

C.1 Non-collapse mechanisms in graph self-supervision

Self-supervised learning on graphs must prevent representations from concentrating, and the field has converged on several distinct mechanisms for doing so. Deep Graph Infomax (Veličković et al., 2019) maximizes mutual information between node representations and a graph-level summary, contrasting against a corrupted view; the corruption supplies the negative signal. GRACE (Zhu et al., 2020) and GCA (Zhu et al., 2021) instead contrast two augmented views node-by-node, pulling together the two copies of a node and pushing apart different nodes, so that spread comes from explicit pairwise repulsion. BGRL (Thakoor et al., 2022) removes negatives altogether, relying on an asymmetric predictor and a target network in the manner of bootstrapped image representations; COSTA (Zhang et al., 2022b) augments in feature space while preserving covariance structure; GraphMAE (Hou et al., 2022) abandons contrast entirely and reconstructs masked features with a scaled cosine error. VICReg (Bardes et al., 2022) makes the anti-collapse term explicit rather than implicit, regularizing the variance and covariance of the embedding coordinates directly, and is the closest Euclidean analogue of the view taken here, where spreading is an additive term with its own scale. Recent node-level methods add further routes: joint spectral augmentation of attribute and topology (Yang et al., 2025), spectral and spatial contrast (Wan et al., 2024), equivalence-based unification of self-supervised objectives (Wang et al., 2025b), edge self-adversarial augmentation (Chen et al., 2026a), dual-kernel community contrast (Chen et al., 2026b), and efficient contrastive learning for fast inference (Xiao et al., 2024). Jing et al. (2022) analyzes how contrastive objectives collapse along a subset of dimensions even when they do not collapse to a point, a failure mode that the hyperbolic methods below revisit in curved space.

What unites all of these is that their anti-collapse mechanism acts in Euclidean or ℓ_2-normalized coordinates. On the unit hypersphere, dispersion has a finite geometric extent and a uniform reference measure (Wang & Isola, 2020), so the question of how strongly to spread has a natural answer. None of these objectives contains the Riemannian volume element, and none therefore meets the integrability boundary that Section 3 describes.

C.2 Hyperbolic graph representation learning

Hyperbolic encoders exist because tree-like graphs embed into negatively curved space with far less distortion than into Euclidean space of the same dimension. HGCN (Chami et al., 2019) formulates graph convolution on the hyperboloid with learnable curvature, HGAT (Zhang et al., 2022a) adds attention, Shimizu et al. (2021) gives a general layer formalism with fewer parameters, and HIE (Yang et al., 2023) revisits how hierarchy is encoded in the radial coordinate. Mixed-curvature variational autoencoders (Skopek et al., 2020), which extend hyperbolic ones (Mathieu et al., 2019), learn the curvature rather than fixing it and are the precedent for the learnable curvature of Section 4.2. Numerical behavior is a recurring practical constraint: Mishne et al. (2023) documents how finite-precision representations degrade at large radii, and Guo et al. (2022) bounds the representation radius directly, a hard constraint where CurvSpread uses a soft one.

C.3 Spreading objectives in hyperbolic space

The work closest to ours studies what happens when a contrastive or spreading objective is transported to negative curvature. HGCL (Liu et al., 2021) contrasts using hyperbolic distances within one space and DSGC (Yang et al., 2022) across Euclidean and hyperbolic ones; Xu et al. (2022) mixes the two geometries layer-wise; Saxena et al. (2022) applies a hyperbolic contrastive objective to network alignment. Hyperbolic self-supervision has also been used beyond homogeneous node classification: for temporal health-event prediction (Lu et al., 2023), heterogeneous graphs with metapath-based contrast (Park et al., 2025), collaborative filtering (Qin et al., 2025), and graph-level anomaly detection with a dual hyperbolic contrast (Fu et al., 2026); recent surveys map this space (Wang et al., 2025a; Chen et al., 2025). In each case the repulsion is pairwise, or the tangent prior is fixed in advance, and the realized representation scale is an outcome of training rather than a designed quantity.

The nearest prior is HyperGCL (Zhang et al., 2025). It begins from the observation that maximizing hyperbolic uniformity drives representations toward the ideal boundary, and responds by replacing uniformity with an isotropy target on tangent-space coordinates at the origin, in Normal and Laplace variants, producing an *outer shell* of near-uniform angular density. In the terms of Section 2 this is a wrapped construction: the target is specified in tangent coordinates at the origin, and its radial scale is inherited from that prior — $\sigma\sqrt{d-1}$ for an isotropic Gaussian — rather than set by hyperbolic volume growth. Its pushforward has a density against dvol_κ and could be rewritten as an intrinsic Gibbs law with energy $-\log q_T + \log J_\kappa$. The point is therefore not that such a method cannot express intrinsic laws. It is that, because a proper tangent law is specified first, normalizability is inherited and there is no confinement coefficient for Proposition 2 to constrain. The distinction is one of specification principle, and it matters because it decides whether a geometry-dependent threshold arises at all. Boundary-based objectives such as hyperbolic Busemann learning (Ghadimi Atigh et al., 2021) prescribe ideal-point prototypes, so their radial behavior is likewise set by the loss design rather than by volume growth.

CurvSpread differs in specifying its entropy and confinement against dvol_κ rather than against a tangent chart, which makes normalizability a geometry-dependent question and yields the coordinate χ. Written in tangent coordinates it is a tangent objective with radial potential $\beta r - \log J_\kappa(r)$; what is specific to it is that this potential is the one induced by the invariant reference measure, and that its confinement coefficient is calibrated to the resulting threshold h_κ (Propositions 2 and 3). The estimator it uses is the Kozachenko–Leonenko nearest-neighbor entropy (Kozachenko & Leonenko, 1987) underlying KoLeo-style spreading (Sablayrolles et al., 2019) and DINOv2 (Oquab et al., 2024), but applied without ℓ_2 normalization, since the radial coordinate is part of the geometry being controlled, and corrected by the exponential-map Jacobian so the surrogate targets Riemannian rather than tangent entropy. Intrinsic Riemannian statistics (Pennec, 2006) treat densities with respect to the volume measure in the same spirit, though not for representation spreading.

D Limitations and Scope of the Theory

D.1 Limitations

The analysis assumes constant negative curvature, and Proposition 1 reflects infinite volume rather than curvature: the same dilation construction applies in Euclidean space. The threshold $\beta > h_\kappa$

is the elementary radial normalizability condition induced by the volume element of Eq. (2). The contribution is not the threshold itself but the identification of which specification principle is subject to it (Section 2), the coordinate χ, its finite-sample counterpart (Proposition 3), and the calibrated objective. The wrapped–intrinsic distinction concerns specification, not representable measures: every absolutely continuous law can be written in either form.

The finite-sample functional provably shares the critical coefficient $\beta = h_\kappa$ with the population free energy (Proposition 3) but not its equilibrium: we do not locate its minimizers, give no mean-field analysis connecting them to $q_{\beta,c}$ as $N \to \infty$, and the radius training realizes is measured rather than predicted (Section 4.1). Proposition 3 is a statement about the objective as a function of the embeddings; it does not address the encoder parameterization or the stochastic-gradient trajectory. Appendix D.2 states these gaps and the experiments that probe them.

Geometry fixes the confinement scale internal to the spreading term. It does not fix the trade-off η of Eq. (13), which is still selected by search, so CURVSPREAD removes one hyperparameter rather than removing tuning. With learnable curvature the coefficient is recomputed each step from a stop-gradient of the current curvature, which keeps $\chi = 1 + \delta$ exactly in the forward pass but means the procedure is not descent on a single time-independent objective.

Whenever $N > M$ the spreading term is estimated on a random subset of size M; with $M^\star = 2048$ this includes every Track A graph. `ogbn-arxiv` is the most aggressively subsampled regime in the current evaluation and the one Track A column in which CURVSPREAD is below the highest entry; whether subsampling explains that result is not established, and we do not claim that the transition behavior is unaffected by aggressive subsampling.

D.2 THEORETICAL ASSUMPTIONS AND APPLICABILITY

Propositions 1 and 2 concern population measures on a homogeneous space. Proposition 3 concerns the finite-sample objective as a function of the embeddings. CURVSPREAD is trained by stochastic gradients through an encoder on a data-dependent frame. We collect here what is established at each of the three levels of Section 4.1, and what is not.

What is established at each level. *(A) Population.* Proved: $Z_{\beta,c} < \infty$ iff $\beta > h_\kappa$, the Gibbs density $q_{\beta,c}$ is the unique minimizer of $F_{\beta,c}$, and its radial mode is $\kappa^{-1/2}\,\mathrm{arccoth}(\chi)$ (Proposition 2, Eq. (21)). *(B) Finite-sample functional.* For every N and k, the functional is bounded below, coercive and attains a minimum when $\beta > h_\kappa$; the lower bound is uniform over the reference point and holds for every finite configuration. When $\beta \le h_\kappa$, Proposition 3(ii) gives an escaping dilation for every configuration of distinct points, so the global infimum is $-\infty$. At fixed κ the same threshold holds after adding the alignment term as stated in Proposition 3(iii). Derived only on the dilation family: the expansion (25) and the scale $R^\star$. Not established: where the minimizers lie or what radius they have, and whether empirical minimizers approach $q_{\beta,c}$ as $N \to \infty$; the dilation scale tracks the tail scale $1/(h_\kappa\delta)$, not the mode, and the two differ by more than an order of magnitude at large d: at $d = 256$, $\kappa = 1$, $\chi = 1.1$ the population mode is 1.52 and the population mean 1.54, whereas $1/(h_\kappa\delta) = 0.039$ (Section 4.1). *(C) Training.* Nothing is proved beyond what holds for the objective at each step. Proposition 3 makes no statement about which configurations the encoder can realize, about convergence of stochastic gradients, or about the point training reaches. The response profiles observed in training (Section 5.2) are empirical findings compared with the objective-level boundary, not consequences of the proposition. Section 5.4 therefore tests the population law by direct sampling and measures the learned radius separately, and does not present the second as a test of the first.

Moving centroid. In training the center is the Lorentz centroid $c = c(Z)$ of the current embeddings, so it moves and receives gradient. The lower bound of Proposition 3(i) holds at every reference point, hence at $c(Z)$. The subcritical sequence of part (iii) consists of centroid-fixed configurations. The threshold of the objective is therefore unaffected by the centroid. What the centroid does affect is the gradient: the trajectory is not that of a fixed-frame objective, and the dilation expansion (25) holds only at a fixed reference point or for centroid-fixed configurations. We do not currently separate the moving center's contribution to the trajectory from that of the spreading term, and record this as an open item.

Subsampling, clipping and curvature. Whenever $N > M$ the functional is evaluated on a random subset of size M. Proposition 3(i) and (ii) hold for every realized subset with N replaced by M, part (i) at any reference point, including the subset centroid; the subcritical half of part (iii) is proved only for the full-set objective with the full-set centroid. Clipping confines radii about o to $[0, R_{\mathrm{clip}}]$, which makes the domain compact, so the clipped objective attains its minimum for every β. The threshold is a property of the unclipped objective. Whether subcritical minimizers of the clipped objective lie on the clipping boundary is not established, which is why boundary proximity is measured rather than predicted (Section 5.2). With learnable curvature and $\beta = (1+\delta)(d-1)\sqrt{\mathrm{sg}(\kappa)}$, Proposition 3(i) applies to the frozen objective of every step, with a lower bound uniform over $[\kappa_{\min}, \kappa_{\max}]$ (Section 4.2). It says nothing about the sequence of iterates.

Estimator bias in high dimension. The Kozachenko–Leonenko estimator is biased at $d = 256$ with $M \leq 4096$, and nearest-neighbor distances concentrate in that regime. Proposition 3 does not use the estimator's accuracy: it uses only the functional form of $\widehat{H}_T$ and the bound $\epsilon_{i,k} \leq 2 \max_j r_j$, so bias cannot move the threshold. The dilation slope $\partial \widehat{H}_T(RY)/\partial \log R = d$ is exact for $\epsilon_{\mathrm{num}} = 0$; Appendix I.5 checks its implementation over the full (d, M) range. How the bias shapes the minimizer, as opposed to the threshold, remains an empirical matter.

Ambient versus effective dimension. Three notions of dimension are kept apart: the ambient dimension d of $\mathbb{H}^d_\kappa$, the dimension m of a set on which a population law may be supported, and empirical effective dimension such as numerical rank. The threshold $h_\kappa = (d-1)\sqrt{\kappa}$ belongs to the ambient-volume objective and involves only the first; no effective-dimension estimate enters Proposition 3, which holds for configurations of any rank, its subcritical part requiring only distinct points. Proposition 2 ranges over laws absolutely continuous with respect to dvol_κ on $\mathbb{H}^d_\kappa$ and so excludes lower-dimensional support by construction. Interpreting $\widehat{H}_T$ as an estimator of d-dimensional differential entropy requires a suitable full-dimensional population density; for a singular lower-dimensional population the usual d-dimensional interpretation does not apply. Not established: whether minimizers of $\mathcal{L}_\beta$, or trained representations, occupy all d tangent directions. The spacing term depends only on local distances, the theorem does not exclude low-rank minimizers, and we do not claim that dimensional collapse is prevented.

Scope of the threshold. Proposition 2 applies to intrinsic entropy under *linear* radial confinement. It does not describe InfoNCE or general pairwise critics, whose asymptotic spreading pressure differs (Appendix F.1), and it does not constrain wrapped constructions, which carry no confinement coefficient of this form and inherit properness from the tangent law. The generalization to Cartan–Hadamard spaces in Appendix F.2 uses only the asymptotic volume-growth rate and should be read as an indication of what the argument depends on rather than a sharp theorem, since subexponential factors can matter at $\beta = h$.

Interaction with the alignment term. At fixed κ, Proposition 3(iii) shows that adding the alignment term leaves the threshold unchanged, because $\mathcal{L}_{\mathrm{align}} \geq 0$ and it vanishes on identical views. In the combined objective the realized radius also depends on η, which geometry does not fix and which is selected by search (Section 4.2). The controlled arms probe the spreading objective in the regime where the theory speaks, while Section 5.2 reports only the operational training response relative to the objective-level boundary.

E Lorentz Model and Geometric Formulas

We use the Lorentz (hyperboloid) model throughout, and convert to the Poincaré ball only for the clipping audit. Work in the ambient space $\mathbb{R}^{d+1}$ with the Minkowski form $\langle x, y\rangle_L = -x_0 y_0 + \sum_{i=1}^d x_i y_i$. The hyperbolic space of curvature $-\kappa$ is the upper sheet

$$\mathbb{H}^d_\kappa = \left\{ x \in \mathbb{R}^{d+1} : \langle x, x\rangle_L = -\tfrac{1}{\kappa},\ x_0 > 0 \right\}, \quad o = \left(\tfrac{1}{\sqrt{\kappa}}, 0, \ldots, 0\right), \tag{14}$$

with tangent space $T_x\mathbb{H}^d_\kappa = \{v : \langle x, v\rangle_L = 0\}$ and the Riemannian metric induced by $\langle\cdot,\cdot\rangle_L$. The geodesic distance is

$$d_\kappa(x, y) = \frac{1}{\sqrt{\kappa}} \operatorname{arccosh}(-\kappa \langle x, y\rangle_L). \tag{15}$$

For $x \in \mathbb{H}^d_\kappa$ and $v \in T_x\mathbb{H}^d_\kappa$ with $\|v\|_L = \sqrt{\langle v, v\rangle_L}$,

$$\exp^\kappa_x(v) = \cosh(\sqrt{\kappa}\|v\|_L)\, x + \frac{\sinh(\sqrt{\kappa}\|v\|_L)}{\sqrt{\kappa}\|v\|_L}\, v, \quad \log^\kappa_x(y) = \frac{d_\kappa(x,y)}{\sqrt{\langle u, u\rangle_L}}\, u, \tag{16}$$

with $u = y + \kappa\langle x, y\rangle_L\, x$ the component of y Minkowski-orthogonal to x. The exponential map preserves radial length, $d_\kappa(x, \exp^\kappa_x(v)) = \|v\|_L$, which is the identity used throughout the main text.

Common center and tangent frame. The shared center c is the Lorentz centroid of the union of the two views. For representations $\{z_i\}$ with Minkowski mean $m = \frac{1}{n}\sum_i z_i$,

$$c = \frac{m}{\sqrt{\kappa}\sqrt{-\langle m, m\rangle_L}}, \tag{17}$$

which lies on $\mathbb{H}^d_\kappa$. Tangent coordinates at c are obtained by parallel transport of $\log^\kappa_c(z_i)$ to the origin,

$$\mathrm{PT}_{x\to y}(v) = v + \frac{\kappa\langle y, v\rangle_L}{1 - \kappa\langle x, y\rangle_L}(x + y), \tag{18}$$

followed by projection onto a fixed orthonormal frame of $T_o\mathbb{H}^d_\kappa \cong \mathbb{R}^d$. The spacings $\epsilon_{i,k}$ of (7) are Euclidean k-th nearest-neighbor distances in these coordinates, with the point itself excluded; since parallel transport is an isometry, they equal the norms $\|y_i - y_j\|_c$ in $T_c\mathbb{H}^d_\kappa$, so the spacing term is computed at c, not at o, and uses tangent rather than geodesic distances. The spreading objective depends on the embeddings only through the tangent spacings and the radii $r_i = d_\kappa(c, z_i)$, both of which are invariant under a joint isometry of the two views; the loss is therefore gauge-invariant under isometries, and re-centering to the origin at each step is a numerical convenience rather than a modeling choice. Gauge invariance under isometries does not by itself remove the data-dependence of c: because c is a function of the embeddings, gradients flow through it, which is the departure from the fixed-frame condition of the dilation argument recorded in Appendix D.2.

Poincaré conversion and the clipping radius. The diffeomorphism to the Poincaré ball of radius $1/\sqrt{\kappa}$ is $p = \frac{1}{\sqrt{\kappa}}\frac{(x_1,\dots,x_d)}{x_0 + 1/\sqrt{\kappa}}$. A radial clamp of the tangent vector at the origin to length R maps to the Poincaré norm $\|p\| = \kappa^{-1/2}\tanh(\sqrt{\kappa}R/2)$, so the projection $\|p\| \le (1-\epsilon)/\sqrt{\kappa}$ corresponds to

$$R_{\mathrm{clip}} = \frac{2}{\sqrt{\kappa}}\,\mathrm{artanh}(1-\epsilon) = \frac{1}{\sqrt{\kappa}}\log\frac{2-\epsilon}{\epsilon}. \tag{19}$$

The factor 2 is the recurring source of convention error; Appendix G.2 reports a direct numerical check that the realized clamp radius equals (19) and not $\kappa^{-1/2}\mathrm{artanh}(1-\epsilon)$.

Notation for the three small quantities. Three symbols that are easy to conflate are kept distinct: the k-nearest-neighbor spacing $\epsilon_{i,k}$ used by the entropy surrogate, the clipping margin ϵ of (19), and the numerical clamp ϵ_{num} inside $\log(\epsilon_{i,k} + \epsilon_{\mathrm{num}})$.

F Proofs and Derivations

F.1 Proof of Proposition 1: escape of unconstrained intrinsic spreading

Proof of Proposition 1. Fix $t > 0$. Let $C_t = \int_{\mathbb{H}^d_\kappa} e^{-t\, d_\kappa(o,z)^2}\, \mathrm{dvol}_\kappa(z)$. In geodesic polar coordinates about o,

$$C_t = \omega_{d-1}\int_0^\infty e^{-tr^2}\left(\frac{\sinh(\sqrt{\kappa}\, r)}{\sqrt{\kappa}}\right)^{d-1} dr < \infty,$$

because the Gaussian factor e^{-tr^2} dominates the $e^{(d-1)\sqrt{\kappa}\, r}$ growth of the volume element; here ω_{d-1} is the surface area of the unit sphere S^{d-1}. Let μ_R be the normalized Riemannian volume on the geodesic ball $B(o, R)$. Then

$$\begin{aligned} I_t(\mu_R) &= \frac{1}{\mathrm{Vol}(B_R)^2}\iint_{B_R\times B_R} e^{-t\, d_\kappa(x,y)^2}\, \mathrm{dvol}_\kappa(x)\, \mathrm{dvol}_\kappa(y) \\ &\le \frac{1}{\mathrm{Vol}(B_R)^2}\int_{B_R} C_t\, \mathrm{dvol}_\kappa(x) = \frac{C_t}{\mathrm{Vol}(B_R)}, \end{aligned}$$

where the inequality extends the inner integral from B_R to all of $\mathbb{H}^d_\kappa$ and uses homogeneity to bound it by C_t. Since $\mathrm{Vol}(B_R) \to \infty$, $I_t(\mu_R) \to 0$ and $U_t(\mu_R) = \log I_t(\mu_R) \to -\infty$, so $\inf_\mu U_t(\mu) = -\infty$. The construction dilates the support rather than translating a fixed cloud, consistent with the isometry invariance of U_t. Because the integrand $e^{-t\, d_\kappa(x,y)^2}$ is strictly positive, $I_t(\mu) > 0$ and hence $U_t(\mu) > -\infty$ for every μ, so the infimum is never attained.

For a configuration of N points with empirical measure $\hat{\mu} = \frac{1}{N}\sum_i \delta_{z_i}$ and self-pairs included, $I_t(\hat{\mu}) = \frac{1}{N^2}\sum_{i,j} e^{-t d_{ij}^2} \geq \frac{1}{N^2}\sum_i e^0 = \frac{1}{N}$, so $U_t(\hat{\mu}) \geq -\log N$; the bound is approached as every off-diagonal distance $d_{ij} \to \infty$ but is never attained by a finite configuration. Excluding self-pairs replaces the diagonal contribution by 0, and the same dilation argument drives $U_t \to -\infty$. □

Scope: the linear-geodesic-energy critic. The escape statement is about the squared-distance uniformity kernel. A pairwise log-sum-exp critic with energy linear in geodesic distance, $-d_\kappa(z_i, z_j)/\tau$, behaves differently. When a fixed angular configuration is placed at radius R, two generic points satisfy $d_\kappa(z_i, z_j) \approx 2R$ up to an angle-dependent constant, so the outward gradient pressure from the critic is $\approx 2/\tau$, independent of d and κ. Proposition 2 governs intrinsic-entropy objectives and is not claimed to describe this critic or InfoNCE-style objectives in general.

F.2 Proof of Proposition 2: critical intrinsic confinement

Proof of Proposition 2. In geodesic polar coordinates about c,

$$Z_{\beta,c} = \omega_{d-1}\int_0^\infty e^{-\beta r}\left(\frac{\sinh(\sqrt{\kappa}\, r)}{\sqrt{\kappa}}\right)^{d-1} dr. \tag{20}$$

As $r \to 0$ the integrand is $\Theta(r^{d-1})$ and integrable. As $r \to \infty$, $\sinh(\sqrt{\kappa}\, r) = \frac{1}{2}e^{\sqrt{\kappa}\, r}(1 - e^{-2\sqrt{\kappa}\, r})$, so the integrand is $\Theta(e^{-(\beta - h_\kappa)r})$ with $h_\kappa = (d-1)\sqrt{\kappa}$. The integral converges at infinity iff $\beta - h_\kappa > 0$. Hence $Z_{\beta,c} < \infty$ if and only if $\beta > h_\kappa$.

Assume $\beta > h_\kappa$, so $q_{\beta,c} = Z_{\beta,c}^{-1} e^{-\beta d_\kappa(c,\cdot)}$ is a well-defined density. For any density ρ with finite $F_{\beta,c}(\rho)$,

$$\begin{aligned} D_{\mathrm{KL}}(\rho \,\|\, q_{\beta,c}) &= \int \rho \log \frac{\rho}{q_{\beta,c}}\, \mathrm{dvol}_\kappa \\ &= \int \rho \log \rho \, \mathrm{dvol}_\kappa - \int \rho\, (-\beta d_\kappa(c,\cdot) - \log Z_{\beta,c})\, \mathrm{dvol}_\kappa \\ &= F_{\beta,c}(\rho) + \log Z_{\beta,c}. \end{aligned}$$

Therefore $F_{\beta,c}(\rho) = D_{\mathrm{KL}}(\rho \,\|\, q_{\beta,c}) - \log Z_{\beta,c} \geq -\log Z_{\beta,c}$, with equality iff $\rho = q_{\beta,c}$ almost everywhere. So $q_{\beta,c}$ is the unique minimizer and the minimum value is $-\log Z_{\beta,c}$.

Assume $\beta \leq h_\kappa$, so $Z_{\beta,c} = \infty$. Let $q_R = Z_R^{-1} e^{-\beta d_\kappa(c,\cdot)} \mathbf{1}_{B(c,R)}$ with $Z_R = \int_{B(c,R)} e^{-\beta d_\kappa(c,\cdot)}\, \mathrm{dvol}_\kappa \to \infty$. A direct computation gives

$$\begin{aligned} F_{\beta,c}(q_R) &= \int q_R \log q_R + \beta \int d_\kappa(c,\cdot)\, q_R \\ &= (-\beta\, \mathbb{E}_{q_R}[d_\kappa(c,\cdot)] - \log Z_R) + \beta\, \mathbb{E}_{q_R}[d_\kappa(c,\cdot)] = -\log Z_R \to -\infty, \end{aligned}$$

so $F_{\beta,c}$ is unbounded below. □

Radial mode and near-critical divergence. For $\beta > h_\kappa$ the radial density is $p(r) \propto e^{-\beta r}\sinh^{d-1}(\sqrt{\kappa}\, r)$. Setting $\frac{d}{dr}\left(-\beta r + (d-1)\log\sinh(\sqrt{\kappa}\, r)\right) = -\beta + (d-1)\sqrt{\kappa}\coth(\sqrt{\kappa}\, r) = 0$ gives $\coth(\sqrt{\kappa}\, r^\star) = \chi$, i.e.

$$r^\star = \frac{1}{\sqrt{\kappa}}\,\mathrm{arccoth}(\chi), \qquad \chi = \frac{\beta}{(d-1)\sqrt{\kappa}} > 1. \tag{21}$$

Since $\mathrm{arccoth}(\chi) = \frac{1}{2}\log\frac{\chi+1}{\chi-1}$, as $\chi \downarrow 1$ we have $r^\star \sim (2\sqrt{\kappa})^{-1}\log(2/(\chi-1)) \to \infty$. For $\chi \leq 1$ no interior stationary point exists and no equilibrium radial mode is defined.

General Cartan–Hadamard spaces. The strict inequalities use only the exponential rate of radial volume growth. On a Cartan–Hadamard manifold whose radial volume density about c has exponential rate h, i.e. its logarithm is $hr + o(r)$, the same polar estimate gives $Z_\beta < \infty$ for $\beta > h$ and $Z_\beta = \infty$ for $\beta < h$. At $\beta = h$ the exponential rates cancel and convergence is decided by the subexponential factors, so no general statement holds there without further assumptions. For $\mathbb{H}^d_\kappa$ the explicit factor $\sinh^{d-1}(\sqrt{\kappa}\, r)$ makes the integrand of order one at $\beta = h_\kappa$, so $Z_{h_\kappa, c} = \infty$ and the boundary is exactly that of Proposition 2. The Euclidean limit $\kappa \to 0$ has $h = 0$: every $\beta > 0$ is confining and no threshold appears, which is consistent with restricting the method to the strictly hyperbolic range $0 < \kappa_{\min} \le \kappa \le \kappa_{\max}$. No dynamical claim about the trajectory at $\chi = 1$ is made; the main text reports that case descriptively.

F.3 Wrapped and intrinsic specifications

This appendix records the identities behind Section 2. Throughout, $z = \exp_c^\kappa(y)$ and $\mathrm{dvol}_\kappa = J_\kappa(\|y\|_c)\, dy$.

(a) *Every wrapped law has a Riemannian density.* The pushforward of a proper q_T has density $q_T(\log_c^\kappa z)/J_\kappa(\|\log_c^\kappa z\|_c)$ against dvol_κ, and $\int q_{\mathrm{wrap}}\, \mathrm{dvol}_\kappa = \int q_T\, dy = 1$. Properness is inherited from q_T, and no further condition arises.

(b) *Every law with a Riemannian density is wrapped from its pullback.* If $q_{\mathbb{H}}$ is a proper density against dvol_κ, then $q_{\mathrm{pull}} = (q_{\mathbb{H}} \circ \exp_c^\kappa)\, J_\kappa$ is a proper tangent density whose pushforward is $q_{\mathbb{H}}$. Together, (a) and (b) show that the two constructions describe the same set of absolutely continuous laws.

(c) *What differs is where the shape is specified.* A shape f specified against dvol_κ is normalizable iff $\int f(\exp_c^\kappa y)\, J_\kappa(\|y\|_c)\, dy < \infty$. For $f = e^{-\beta d_\kappa(c,\cdot)}$ this is the condition $\beta > h_\kappa$ of Proposition 2, and for $\beta \le h_\kappa$ no law of that shape exists in either coordinate system. The same shape specified against Lebesgue measure on $T_c\mathbb{H}^d_\kappa$ is normalizable for every $\beta > 0$. The reference measure dvol_κ is isometry-invariant; $(\exp_c^\kappa)_\# dy = J_\kappa(d_\kappa(c,\cdot))^{-1}\mathrm{dvol}_\kappa$ depends on c.

(d) *The radial law of a wrapped construction is inherited.* $d_\kappa(c, Z) = \|Y\|_c$, so an isotropic Gaussian of scale σ gives the radial density $\propto r^{d-1}e^{-r^2/(2\sigma^2)}$, with mode $\sigma\sqrt{d-1}$ independent of κ. HyperGCL's Normal and Laplace tangent-isotropy targets are wrapped constructions in this sense. They carry no confinement coefficient of the form in Eq. (3), so Proposition 2 places no constraint on them. This is a statement about how they are specified, not about which measures they could represent.

F.4 Hyperbolic entropy under change of variables

With $Z = \exp_c^\kappa(Y)$ and $R = \|Y\|$, change variables $z = \exp_c^\kappa(y)$, under which $\mathrm{dvol}_\kappa(z) = J_\kappa(\|y\|)\, dy$:

$$
\begin{aligned}
H_{\mathbb{H}}(Z) &= -\int q_{\mathrm{wrap}}(z) \log q_{\mathrm{wrap}}(z)\, \mathrm{dvol}_\kappa(z) = -\int \frac{q_T(y)}{J_\kappa(\|y\|)} \log \frac{q_T(y)}{J_\kappa(\|y\|)} J_\kappa(\|y\|)\, dy \\
&= -\int q_T(y)\, (\log q_T(y) - \log J_\kappa(\|y\|))\, dy = H(Y) + \mathbb{E}\left[\log J_\kappa(R)\right]. \qquad (22)
\end{aligned}
$$

The identity holds for every absolutely continuous law of Z, since by Appendix F.3(b) every such law is the pushforward of its pullback; its two sides are entropies of one law relative to dvol_κ and to Lebesgue measure on $T_c\mathbb{H}^d_\kappa$. The finite-sample surrogate replaces $H(Y)$ by the local-spacing estimator $\widehat{H}_T$ and the expectation by the empirical mean, giving $\widehat{H}_{\mathbb{H}} = \widehat{H}_T + \frac{1}{N}\sum_i \log J_\kappa(r_i)$. Dropping the second term leaves $H(Y)$, the entropy relative to $(\exp_c^\kappa)_\# dy = J_\kappa^{-1}\mathrm{dvol}_\kappa$, whose maximization under a linear radial penalty yields the tangent law $\propto e^{-\beta\|y\|_c}\, dy$, proper for every $\beta > 0$. The Jacobian term is the only place hyperbolic volume growth enters, and hence the only source of the threshold.

F.5 THE k-NEAREST-NEIGHBOR SURROGATE AND PROOF OF PROPOSITION 3

The Kozachenko–Leonenko estimator of differential entropy from k-nearest-neighbor distances is

$$\widehat{H}_T = \psi(N) - \psi(k) + \log V_d + \frac{d}{N}\sum_{i=1}^{N} \log \epsilon_{i,k}, \qquad V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}, \tag{23}$$

with ψ the digamma function and $\epsilon_{i,k}$ the distance from y_i to its k-th nearest neighbor in the tangent frame. The constant $C = \psi(N) - \psi(k) + \log V_d$ depends only on N, k, and d. Under a dilation $y_i \mapsto Ry_i$ every spacing scales as $\epsilon_{i,k} \mapsto R\epsilon_{i,k}$, so, for $\epsilon_{\text{num}} = 0$ and whatever the estimator's bias,

$$\widehat{H}_T(RY) = \widehat{H}_T(Y) + d \log R. \tag{24}$$

Combined with the large-radius expansion $\log J_\kappa(r) = (d-1)\big[\sqrt{\kappa}\, r - \log(\sqrt{\kappa}\, r) - \log 2\big] + o(1)$, the mean Jacobian term satisfies $\frac{1}{N}\sum_i \log J_\kappa(Rr_i) = h_\kappa R\bar{r} - (d-1)\log R + c_d + o(1)$ with c_d independent of R, so

$$\mathcal{L}_\beta(RY) = -\widehat{H}_{\mathbb{H}}(RY) + \beta R\bar{r} = (\beta - h_\kappa)R\bar{r} - \log R + c_d' + o(1). \tag{25}$$

The stationary point of the leading terms of (25) is $R^\star = 1/((\beta - h_\kappa)\bar{r})$, which diverges as $\beta \downarrow h_\kappa$. It tracks the tail scale of $q_{\beta,c}$ rather than its mode, as discussed with a numerical example in Appendix D.2. With the numerical clamp of Appendix G.1, (24) holds up to $O(\epsilon_{\text{num}}/\epsilon_{i,k})$. Sensitivity to k is reported in Appendix I.4. The expansion (25) concerns one family of configurations; Proposition 3 is its global counterpart and is proved next.

Two elementary facts. (F1) Let $g(x) = \log(\sinh x/x)$, so $\log J_\kappa(r) = (d-1)\, g(\sqrt{\kappa}\, r)$ and $g(0) = 0$. For $x > 0$,

$$g(x) = x - \log(2x) + \log\big(1 - e^{-2x}\big), \qquad 0 \le g(x) \le x,$$

the upper bound because $\sinh x \le x\cosh x \le xe^x$. Hence $0 \le \log J_\kappa(r) \le h_\kappa r$ for all $r \ge 0$. (F2) With $\rho = \max_j r_j$, the triangle inequality in the tangent frame gives $\epsilon_{i,k} \le \max_{j\neq i} \|y_i - y_j\| \le 2\rho$.

Proof of Proposition 3. *(i)* Let $\beta > h_\kappa$. By (F1), $\beta r_i - \log J_\kappa(r_i) \ge (\beta - h_\kappa)r_i \ge 0$, so the second sum in (9) is at least $(\beta - h_\kappa)\rho/N$. By (F2), $\log(\epsilon_{i,k} + \epsilon_{\text{num}}) \le \log(2\rho + \epsilon_{\text{num}}) \le \log(2(\rho + \epsilon_{\text{num}}))$. Therefore, with $s = \rho + \epsilon_{\text{num}}$,

$$\mathcal{L}_\beta(Y) \;\ge\; -C + \frac{(\beta - h_\kappa)\,\rho}{N} - d\log(2\rho + \epsilon_{\text{num}}) \;\ge\; -C + \varphi(s) - \frac{(\beta - h_\kappa)\,\epsilon_{\text{num}}}{N},$$
$$\varphi(s) = \frac{(\beta - h_\kappa)s}{N} - d\log(2s).$$

On $s > 0$, φ is minimized at $s = dN/(\beta - h_\kappa)$ with value $d - d\log(2dN/(\beta - h_\kappa))$, which gives $\mathcal{L}_\beta \ge B_\beta$. The middle expression tends to $+\infty$ as $\rho \to \infty$, which gives coercivity. Each $\epsilon_{i,k}$ is a continuous function of Y, being an order statistic of continuous functions, so $\mathcal{L}_\beta : (\mathbb{R}^d)^N \to (-\infty, +\infty]$ is lower semicontinuous. Its sublevel sets are closed and, by coercivity, bounded, hence compact, and they are nonempty for any level above the (finite) value at a configuration of distinct points. A lower semicontinuous function attains its minimum on a nonempty compact sublevel set.

(ii) Let $\beta \le h_\kappa$ and let Y consist of distinct points, so every $\epsilon_{i,k} > 0$ and at most one point, counted by $n_0 \in \{0, 1\}$, lies at the origin. For $R > 0$, $\log(R\epsilon_{i,k} + \epsilon_{\text{num}}) \ge \log R + \log \epsilon_{i,k}$. By the identity in (F1), for $r_i > 0$,

$$-\log J_\kappa(Rr_i) = -(d-1)\sqrt{\kappa}\, Rr_i + (d-1)\log R$$
$$+ (d-1)\log(2\sqrt{\kappa}\, r_i) - (d-1)\log\Big(1 - e^{-2\sqrt{\kappa}\, Rr_i}\Big),$$

and $\log J_\kappa(0) = 0$. Summing,

$$\mathcal{L}_\beta(RY) \;\le\; (\beta - h_\kappa)R\bar{r} - \Big(d - \tfrac{(N-n_0)(d-1)}{N}\Big)\log R + K_Y + \epsilon_R,$$

where K_Y does not depend on R and $\epsilon_R = -\frac{d-1}{N}\sum_{r_i>0}\log(1-e^{-2\sqrt{\kappa}\,Rr_i}) \to 0$. The coefficient of $\log R$ is at least 1 and $(\beta - h_\kappa)R\bar{r} \le 0$, so $\mathcal{L}_\beta(RY) \to -\infty$.

(iii) Invariance. Let Λ be an isometry of $\mathbb{H}^d_\kappa$, i.e. a linear map of $\mathbb{R}^{d+1}$ preserving $\langle\cdot,\cdot\rangle_L$ and the upper sheet. The Minkowski mean transforms as $m \mapsto \Lambda m$ with $\langle \Lambda m, \Lambda m\rangle_L = \langle m, m\rangle_L$, so $c(\Lambda Z) = \Lambda\, c(Z)$ by (17), and $\log^\kappa_{\Lambda c}(\Lambda z) = \Lambda \log^\kappa_c(z)$ by (16). The functional $\mathcal{L}_\beta$ depends on the tangent vectors only through their norms and pairwise distances, so it is independent of the choice of orthonormal frame and invariant under Λ; $\mathcal{L}_{\text{align}}$ is invariant because d_κ is. Hence $\Phi_\beta(\Lambda Z) = \Phi_\beta(Z)$.

Supercritical case. The bound B_β in (i) does not depend on the reference point, so it holds for each view at $c(Z)$; with $\mathcal{L}_{\text{align}} \ge 0$ this gives $\Phi_\beta \ge \eta B_\beta$. For attainment, take a minimizing sequence Z_n and, by transitivity and invariance, assume $c(Z_n) = o$. Since $\Phi_\beta(Z_n) \le a$ for some a and the other view contributes at least $\frac{\eta}{2}B_\beta$, each view satisfies $\mathcal{L}_\beta(Y_n^{(v)}) \le 2a/\eta - B_\beta$, so by coercivity in (i) all radii $d_\kappa(o, z_{n,i})$ are bounded. A subsequence converges to some Z^*. The centroid is continuous because m is a nonzero future timelike vector, so $c(Z^*) = o$. The map $Z \mapsto Y^{(v)}(Z)$ is continuous, so Φ_β is lower semicontinuous and $\Phi_\beta(Z^*) \le \liminf_n \Phi_\beta(Z_n) = \inf \Phi_\beta$.

Subcritical case. Let $Y \subset T_o\mathbb{H}^d_\kappa \cong \mathbb{R}^d$ be a centrally symmetric set of N distinct points: pairs $\pm y_j$, together with 0 if N is odd. Set $z_i^{(1)} = z_i^{(2)} = \exp^\kappa_o(Ry_i)$. Since $\exp^\kappa_o(v) + \exp^\kappa_o(-v) = 2\cosh(\sqrt{\kappa}\|v\|)\, o$, the Minkowski mean is a positive multiple of o, so $c(Z) = o$ for every R. The tangent coordinates are then Ry_i and $\mathcal{L}_{\text{align}} = 0$, so $\Phi_\beta = \eta\,\mathcal{L}_\beta(RY) \to -\infty$ by (ii). □

Remarks. *Without the Jacobian.* For $\mathcal{L}^T_\beta = -\widehat{H}_T + \beta\bar{r}$ the proof of (i) goes through with $\beta - h_\kappa$ replaced by β, so the tangent-space objective is bounded below and attains its minimum for every $\beta > 0$; the threshold enters only through the Jacobian, as in the population statement of Appendix F.4. *Clipping.* If radii about o are constrained to $[0, R_{\text{clip}}]$, the domain is compact and the objective attains its minimum for every β; Proposition 3 is a statement about the unclipped objective. *Learnable curvature.* With $\beta = (1+\delta)h_\kappa$, the gap $\beta - h_\kappa = \delta(d-1)\sqrt{\kappa}$ ranges over the compact interval $[\delta(d-1)\sqrt{\kappa_{\min}},\, \delta(d-1)\sqrt{\kappa_{\max}}] \subset (0,\infty)$ as κ ranges over $[\kappa_{\min}, \kappa_{\max}]$. Since B_β depends continuously on $\beta - h_\kappa$, it has a finite minimum over that interval, and every frozen per-step objective of the stop-gradient scheme is bounded below by this one constant. This is a property of each objective, not of the iterates. *Low-dimensional configurations.* Neither (F1) nor (F2) involves the span of Y: (F1) is a property of the ambient Jacobian and (F2) is the triangle inequality, so Proposition 3 holds verbatim for configurations of any rank, with (ii) requiring only distinct points. With d replaced by $m \ge 2$ throughout ($\widehat{H}_T$, C and J_κ), the same proof gives the threshold $(m-1)\sqrt{\kappa}$; that is the functional of a different, m-dimensional model, not a correction to the present one. *What the proposition does not give.* It does not locate or characterize the minimizers, does not relate them to $q_{\beta,c}$, and does not concern the encoder parameterization or the optimization trajectory.

G Algorithm and Numerical Implementation

G.1 Algorithm and complexity

Algorithm 1 gives one optimization step. In all experiments the numerical clamp is $\epsilon_{\text{num}} = 10^{-8}$. Whenever $N > M$, the spacing term, the Jacobian term, the mean radius and the center are computed on one random node subset S of size M, drawn at each step and shared by both views; otherwise S is the full node set. With the selected $M^\star = 2048$ this applies to every Track A graph, and `ogbn-arxiv`, where M/N is smallest, is the most aggressively subsampled. Curvature is parameterized as $\kappa = \kappa_{\min} + (\kappa_{\max} - \kappa_{\min})\,\text{sigmoid}(\theta_\kappa)$ with $\kappa_{\min} = 0.1$, $\kappa_{\max} = 2$ and θ_κ initialized so that $\kappa = 1$, as in mixed-curvature latent spaces (Skopek et al., 2020); this keeps κ in the strictly hyperbolic range of Proposition 2 and bounds h_κ above and below throughout training. Smoothing the coefficient instead, $\bar{\kappa} = \text{sg}(\text{EMA}(\kappa))$, gives $\chi = (1+\delta)\sqrt{\bar{\kappa}/\kappa}$, which drops below 1 whenever $\kappa > (1+\delta)^2\bar{\kappa}$; we therefore use the instantaneous form of Eq. (12) and keep the EMA as an ablation. The per-step cost beyond the encoder is $O(M^2 d)$ for the exact pairwise spacings on the subset, or $O(Md\log M)$ with an approximate neighbor index, plus $O(Md)$ for the radii, Jacobian, and confinement terms. Subsampling applies only to the spreading term: the alignment loss is computed over all N paired nodes at $O(Nd)$, so the spreading regularizer adds no term that scales with N beyond what alignment already costs.

Algorithm 1 CURVSPREAD step (one minibatch)

input views $G^{(1)}, G^{(2)}$; encoder f_θ; curvature κ; margin δ; weight η; neighbors k; subset size M
$Z^{(v)} \leftarrow f_\theta(G^{(v)}) \in \mathbb{H}^d_\kappa$ for $v \in \{1, 2\}$
$S \leftarrow \{1, \dots, N\}$ if $N \le M$, else a uniformly random subset of $\{1, \dots, N\}$ with $|S| = M$ ▷ shared by both views
$c \leftarrow \mathrm{Centroid}(Z^{(1)}_S \cup Z^{(2)}_S)$ ▷ Eq. (17)
$\beta \leftarrow (1 + \delta)(d - 1)\sqrt{\mathrm{sg}(\kappa)}$ ▷ stop-gradient on the coefficient only
for $v \in \{1, 2\}$ **do**
  $y_i \leftarrow \log^\kappa_c(z^{(v)}_i)$ in the tangent frame; $r_i \leftarrow \|y_i\|$ for $i \in S$
  $\epsilon_{i,k} \leftarrow k$-th nearest tangent spacing within $\{y_j\}_{j \in S}$; $\widehat{H}^{(v)}_T \leftarrow C + \frac{d}{|S|} \sum_{i \in S} \log(\epsilon_{i,k} + \epsilon_{\mathrm{num}})$, with $C = C(|S|, k, d)$
  $\widehat{H}^{(v)}_{\mathbb{H}} \leftarrow \widehat{H}^{(v)}_T + \frac{1}{|S|} \sum_{i \in S} \log J_\kappa(r_i)$
  $\mathcal{L}^{(v)}_{\mathrm{spread}} \leftarrow -\widehat{H}^{(v)}_{\mathbb{H}} + \beta \frac{1}{|S|} \sum_{i \in S} r_i$
end for
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{align}} + \frac{\eta}{2}(\mathcal{L}^{(1)}_{\mathrm{spread}} + \mathcal{L}^{(2)}_{\mathrm{spread}})$; update θ, κ by one optimizer step on $\mathcal{L}$

G.2 NUMERICAL STABILITY

Stable $\log \mathrm{sinhc}$**.** The Jacobian uses $\log J_\kappa(r) = (d-1) g(\sqrt{\kappa}\, r)$ with $g(x) = \log(\sinh x / x)$. It is evaluated by three branches for accuracy across scales: a five-term Taylor series $g(x) = \frac{x^2}{6} - \frac{x^4}{180} + \cdots$ for small x (threshold 0.1 in FP64, 0.5 in FP32); the direct form for moderate x; and the log-domain form $g(x) = x - \log(2x) + \log(1 - e^{-2x})$ for $x \ge 20$, which avoids overflow of $\sinh$. Against a 50-digit `mpmath` reference the implementation is finite and accurate over $r \in [10^{-6}, 10^3]$, $\kappa \in [10^{-3}, 10]$, $d \le 1024$ in both precisions (worst relative error 3.5×10^{-7} in FP32, 9.0×10^{-14} in FP64).

Distance form. At the radii reached under the clipping ladder the textbook arccosh inner-product form and the Minkowski-difference form lose all significant digits for nearby points even in FP64. The cancellation-free polar form

$$d_\kappa(x, y) = \frac{2}{\sqrt{\kappa}} \operatorname{arcsinh} \sqrt{\sinh^2\Big(\tfrac{\sqrt{\kappa}(a-b)}{2}\Big) + \sinh(\sqrt{\kappa}\, a) \sinh(\sqrt{\kappa}\, b) \tfrac{\|u-w\|^2}{4}}, \tag{26}$$

with (a, u) and (b, w) the polar radius and unit direction of x and y, is used everywhere and stays within ten times the input-resolution floor across the tested regime.

Clipping-convention audit. An audit over $\kappa \in \{0.1, 0.25, 0.5, 1, 2\} \times \epsilon \in \{10^{-2}, \dots, 10^{-7}\} \times \{\mathrm{FP32}, \mathrm{FP64}\}$ confirms that the realized Lorentz clamp radius equals R_{clip} of (19): the ratio realized$/R_{\mathrm{clip}} = 1.0000$, realized$/[\kappa^{-1/2} \mathrm{artanh}(1-\epsilon)] = 2.0000$, and the Poincaré projection mask agrees with the Lorentz clamp; the worst relative deviation of the realized radius is 2.2×10^{-7} (FP32) and 4.0×10^{-16} (FP64). An FP32 Poincaré-norm implementation saturates $\epsilon = 10^{-7}$ at a realized 1.19×10^{-7}; the Lorentz clamp is unaffected. Boundary quantities ($Q_{0.95}$, boundary occupancy and clipping frequency) are measured from the origin, about which clipping acts; center-referenced radii are logged alongside as spread diagnostics. Training-run precision and clipping stress tests include FP64 reruns of the boundary-seeking and critical arms at the extreme ϵ settings.

H Experimental Details

H.1 Datasets

Table 3 lists the ten graphs used. The three citation graphs are the Planetoid benchmarks (Yang et al., 2016); the co-purchase and co-authorship graphs are those of Shchur et al. (2018); `ogbn-arxiv` is from the Open Graph Benchmark (Hu et al., 2020); Disease and Airport follow the preprocessing of Chami et al. (2019). All edge counts use one convention, undirected pairs with self-loops removed, which is why Airport has 18,630 edges rather than the 18,631 sometimes reported. `ogbn-arxiv` is a directed citation graph, symmetrized before training.

Table 3: Dataset statistics. Edges are undirected pairs after removing self-loops; the track and split used for each graph are listed on the right.

Dataset	#Nodes	#Edges	#Cl.	#Feat.	Track	Split
Cora	2,708	5,278	7	1,433	A, B	10/10/80†
CiteSeer	3,327	4,552	6	3,703	A, B	10/10/80†
PubMed	19,717	44,324	3	500	A, B	10/10/80†
Amazon Computers	13,752	245,861	10	767	A	10/10/80
Amazon Photo	7,650	119,081	8	745	A	10/10/80
Coauthor CS	18,333	81,894	15	6,805	A	10/10/80
Coauthor Physics	34,493	247,962	5	8,415	A	10/10/80
`ogbn-arxiv`	169,343	1,157,799	40	128	A	official
Disease	1,044	1,043	2	1,000	B	30/10/60
Airport	3,188	18,630	4	4	B	70/15/15

The split column follows the two tracks of Section 5.1. On Track A every graph except `ogbn-arxiv` uses ten predetermined stratified 10/10/80 splits, and `ogbn-arxiv` uses the official OGB split. On Track B the citation graphs (†) use the Planetoid protocol with twenty labeled nodes per class, and Disease and Airport use the 30/10/60 and 70/15/15 splits of Chami et al. (2019).

Disease has $m = n - 1$ and is connected, so it is a tree; this is the fact behind the pre-specified choice of primary datasets in Appendix H.2. Airport contains cycles. The largest-connected-component statistics used there differ slightly from the full-graph counts above for Airport (3,158 nodes and 18,605 edges).

H.2 Protocol and pre-specification

Track definitions. Track A covers Cora, CiteSeer, and PubMed (Yang et al., 2016), Amazon Computers, Amazon Photo, Coauthor CS, and Coauthor Physics (Shchur et al., 2018), and `ogbn-arxiv` (Hu et al., 2020). On the seven small and medium graphs every method is evaluated on exactly ten predetermined run identifiers, each fixing one stratified 10/10/80 train/validation/test split seed and one model seed, and accuracy is reported as mean and standard deviation over the ten paired runs. On `ogbn-arxiv` we use the official OGB split with five fixed model seeds. Encoders are frozen after self-supervised training, a single linear-probe implementation is shared within the track, model selection uses validation splits only, and the test split is never read during selection.

Track B compares CurvSpread with its nearest prior, HyperGCL (Zhang et al., 2025), on Cora, CiteSeer, PubMed, Disease, and Airport. We use the dataset splits reported by HyperGCL: the Planetoid splits with twenty training nodes per class for the citation graphs, 30/10/60 for Disease, and 70/15/15 for Airport; the latter two follow the preprocessing of Chami et al. (2019). The metric is node-classification accuracy. HyperGCL's published scores are identified as literature-reported values and compared descriptively with our independently evaluated results, rather than presented as paired reproductions. Results obtained under incompatible splits are excluded rather than pooled.

Common settings. The controlled training experiments of Sections 5.2–5.4 use a common encoder, augmentation, curvature, initialization, optimizer, schedule and seed policy whenever trained arms are compared; direct population-law sampling (Section 5.4) involves no encoder and is identified separately. The downstream comparison of Section 5.5 is method-level: in our Track A reproductions, methods retain their official architectures while sharing the same splits, frozen-encoder evaluation, linear probe and validation budget, and literature-reported entries, identified cell by cell in Appendix H.3, serve as contextual references rather than members of the controlled comparison. The representation dimension is $d = 64$ in Track A and in the auxiliary controlled arms; in Track B it follows the HyperGCL evaluation setting. CurvSpread uses one global configuration selected by the predeclared two-stage validation search of Appendix H.3 on Cora, Amazon Photo and Coauthor CS, and its selected hyperparameters are fixed for all entries in Tables 1 and 7. Clipping projects

Poincaré coordinates to $\|p\| \le (1-\epsilon)/\sqrt{\kappa}$, corresponding to a radius $R_{\text{clip}} = \kappa^{-1/2}\log((2-\epsilon)/\epsilon)$ about the origin (Appendix E); numerical, clipping-convention and precision audits are reported in Appendix G.2.

Pre-specification record. Before any representation-learning run, the controlled datasets were specified and characterized by the normalized criterion $\delta_{\text{rel}} = \hat{\delta}_b(G)/\text{diam}(G)$, with $\hat{\delta}_b$ the sampled four-point Gromov procedure (Fournier et al., 2015) (maximum over samples) on the largest connected component, undirected, unweighted, self-loops removed, with exact breadth-first distances; the sampling budget is 2×10^5 quadruples per seed over seeds $\{0, 1, 2, 3, 4\}$. The two candidates give Disease ($n_{\text{lcc}} = 1044$, $m_{\text{lcc}} = 1043$, diameter 10, $\hat{\delta}_b = 0.0$, $\delta_{\text{rel}} = 0.0$) and Airport ($n_{\text{lcc}} = 3158$, $m_{\text{lcc}} = 18605$, diameter 12, $\hat{\delta}_b = 1.5$, $\delta_{\text{rel}} = 0.125$). Disease is a tree and Airport contains cycles, and these statistics document that structural contrast. The record designates Disease as the primary controlled dataset and Airport as the complementary one, with cycles. These dataset roles and reading rules were fixed before the corresponding representation-learning runs; we therefore describe them as pre-specified rather than as third-party pre-registered.

Auxiliary controlled arms. The five arms are Pairwise-U ($t = 1$; Eq. (4) at the empirical measure of the N representations, so the double sum includes the N self-pairs and $U_t \ge -\log N$), the wrapped Gaussian control $\mathcal{L}_{\text{wrap}} = -\widehat{H}_T + \frac{1}{2\sigma^2 N}\sum_i r_i^2$ ($\sigma = 0.25$), and the intrinsic-linear objective at $\chi \in \{0.8, 1.0, 1.2\}$. Held fixed across arms: a two-layer GCN encoder (hidden 256, PReLU, Glorot init), GRACE-style augmentation (edge drop 0.2/0.4, feature mask 0.3/0.4), Adam at learning rate 10^{-3} and weight decay 10^{-5} for 2000 full-batch epochs, $d = 64$, $\kappa = 1$, FP32, seeds 301–305, and default clipping $\epsilon = 10^{-5}$. A clipping-sensitivity diagnostic varies only $\epsilon \in \{10^{-2}, 10^{-3}, \ldots, 10^{-7}\}$; radius summaries and terminal drift are recorded, with FP64/Lorentz checks at the extreme settings (Appendix G.2). These controls are auxiliary diagnostics of scale and numerical boundaries; the main-text training claim is based on the operational midpoint grid and selected response profiles, not on a fitted training-time law.

Radial-law constants. The primary configuration of Section 5.4 fixes $\sigma = 0.25$, $\rho_{\max} = 0.50$, $\epsilon = 10^{-5}$, and $\chi = 1.2$. The smallest clipping radius over $\kappa \in \{0.1, 0.25, 0.5, 1, 2\}$ is $R_{\text{clip,min}} = \log((2-10^{-5})/10^{-5})/\sqrt{2} = 8.6310$, and the safety check $\sigma\sqrt{255}/R_{\text{clip,min}} = 0.4625 \le 0.50$ passes; σ is not changed after any result is inspected.

Operational midpoint estimator. The β grid of each (d, κ) cell is placed by the theory, $\beta = \chi\, h_\kappa$ for $\chi \in \{0.70, 0.75, \ldots, 1.50\}$, so the estimator locates a response midpoint relative to the predicted boundary rather than searching for the normalization. The diagnostic is the terminal $Q_{0.95}(r^{(o)})/R_{\text{clip}}$ over the last 5% of epochs, fixed before the grid is examined. For each (d, κ) the seed-mean curve over the β grid is smoothed by a three-point median filter with edge replication; the low- and high-β reference levels are the means over the two smallest and two largest β values, and $\hat{\beta}_c$ is the linearly interpolated crossing of their midpoint. Uncertainty is defined by a 1000-replicate bootstrap over seeds. A plateau gap below 0.05 is treated as a weak response for interpretation, and a grid-edge crossing is likewise treated cautiously; the numerical midpoint location is retained for comparability but is not by itself evidence of a sharp transition. The reported coordinate is $\hat{\chi}_c = \hat{\beta}_c/((d-1)\sqrt{\kappa})$.

Decision language. Each experiment has a pre-specified reading. The auxiliary controlled arms record boundary and scale diagnostics but are not used to fit a training-time law. The midpoint grid reports the cell-wise locations $\hat{\chi}_c$ and their median, interquartile range and range over the 25 cells, while response strength and sharpness are interpreted separately from the response curves. The radial-law comparison sets empirical modes against both analytic predictions. Downstream tables report mean and standard deviation; supervised references are never ranked with self-supervised methods. Runs are excluded only under the predeclared implementation-failure criteria (non-finite loss or gradient, out-of-memory, or a verified code defect fixed and rerun under the same configuration); valid but unfavorable runs are never excluded.

Radial-law measurements. *Population laws.* Sampling the two radial densities directly, with no encoder or estimator in the loop, we check $\sqrt{\kappa}\,\hat{r}_{\text{mode}} = \text{arccoth}(\chi)$ and $r^\star_{\text{wrap}} = \sigma\sqrt{d-1}$ as d and κ vary. This tests Proposition 2 and nothing else, and is the only place where the two analytic

radial laws are compared against their own populations. *Learned surrogate.* Separately, we measure the radial law that the trained arms actually realize. To test the wrapped law itself, σ is held fixed for every d and κ, with the safety check of the radial-law constants above, verified in advance, so the predicted wrapped mode stays clear of the smallest clipping radius. We compare the empirical modes of the wrapped control and the intrinsic $\chi = 1.2$ arm against the two population laws as d and κ vary, reporting the gap rather than asserting equality. CURVSPREAD appears as an additional marker series and not as a test of the population formula, because it is trained with the full pipeline rather than as an isolated spreading objective. *Supercritical mode.* For cells with $\chi > 1$ we test the radial-mode formula directly by comparing $\sqrt{\kappa}\,\hat{r}_{\mathrm{mode}}$ with $\operatorname{arccoth}(\chi)$, where $\hat{r}_{\mathrm{mode}}$ is estimated by a one-dimensional kernel density estimate with one fixed bandwidth rule for every run, and neither the median nor $Q_{0.95}$ is used in its place.

H.3 DEVELOPMENT SEARCH

CURVSPREAD hyperparameters are selected once by a two-stage development search on the validation splits of run identifiers one to three on Cora, Amazon Photo, and Coauthor CS, reading no test data. Stage A fixes $k = 5$, $M = 2048$ and searches $\delta \in \{0.05, 0.1, 0.2, 0.5\} \times \eta \in \{0.1, 0.3, 1, 3\}$; Stage B freezes $(\delta^\star, \eta^\star)$ and searches $k \in \{1, 3, 5, 10\} \times M \in \{1024, 2048, 4096\}$. Ties break toward smaller δ then η (Stage A) and smaller k then M (Stage B). The frozen $(\delta^\star, \eta^\star, k^\star, M^\star)$ is used for every CURVSPREAD entry in Tables 1 and 7. Table 4 records the search: Stage A selects $\delta^\star = 0.1$, $\eta^\star = 1.0$ at 90.07% mean validation accuracy, a unique maximum, so no tie-break fired; with $(\delta^\star, \eta^\star)$ frozen, Stage B returns $k^\star = 5$, $M^\star = 2048$, recovering the values used during Stage A. The full grids are in Appendix I.4.

Table 4: Two-stage development search for $(\delta^\star, \eta^\star, k^\star, M^\star)$ on the validation splits of run identifiers one to three of Cora, Amazon Photo and Coauthor CS; no test data are read.

Stage	Grid	Selection rule	Result
A	$\delta \in \{0.05, 0.10, 0.20, 0.50\} \times \eta \in \{0.1, 0.3, 1.0, 3.0\}$, at $k = 5$, $M = 2048$	max. mean val. acc.; ties $\rightarrow$ smaller δ, then η	$\delta^\star = 0.10$, $\eta^\star = 1.0$
B	$k \in \{1, 3, 5, 10\} \times M \in \{1024, 2048, 4096\}$, at $(\delta^\star, \eta^\star)$	max. mean val. acc.; ties $\rightarrow$ smaller k, then M	$k^\star = 5$, $M^\star = 2048$

In Table 1, every CURVSPREAD cell and the following baseline cells are our Track A reproductions with official code: GraphECL (Comp., CS, Physics); S3GCL (Physics); GCL-JAM (CS, Physics, arXiv); Dual-Kernel GCCL (Comp., arXiv); HGCL (Comp., Photo, CS, Physics, arXiv); HyperGCL-Normal (Comp., Photo, CS, Physics, arXiv); HyperGCL-Laplace (Comp., Photo, CS, Physics, arXiv). Every other baseline cell is the value reported in the cited paper under its own protocol and was not reproduced under the unified Track A protocol; those values are retained as contextual references, and the shading in that table marks the best and second-best value in each column, not a controlled ranking.

I Additional Results

I.1 Link prediction

Setup. We additionally evaluate whether the representations learned by CurvSpread preserve graph structure sufficiently well to support link prediction, providing a complementary evaluation to the node-classification experiments in Section 5.5. We consider six commonly used benchmarks: Cora, CiteSeer, and PubMed (Yang et al., 2016); Amazon-Photo and Amazon-Computers (Shchur et al., 2018); and Wiki-CS (Mernyei & Cangea, 2020). The first three are citation networks, the two Amazon datasets are co-purchase networks, and Wiki-CS is a hyperlink network derived from Wikipedia.

Table 5: Link-prediction AUC (%). Shading marks the best (bold) and second-best value in each column.

Method	Cora	CiteSeer	PubMed	Photo	Comp.	Wiki-CS
Spectral and random-walk embeddings						
Spectral (Ng et al., 2001)	84.6	80.5	84.2	83.6	86.7	81.2
DeepWalk (Perozzi et al., 2014)	83.1	80.5	84.4	83.2	87.2	82.1
Graph autoencoders						
GAE (Kipf & Welling, 2016)	91.0	89.5	96.4	89.9	93.0	83.2
VGAE (Kipf & Welling, 2016)	91.4	90.8	94.4	89.3	92.6	82.5
ARGA (Pan et al., 2018)	92.4	91.9	96.8	91.5	92.5	83.5
ARVGA (Pan et al., 2018)	92.4	92.4	96.5	92.1	93.1	82.6
Graph contrastive learning						
GRACE (Zhu et al., 2020)	90.9	92.1	97.0	90.8	91.6	89.5
GCA (Zhu et al., 2021)	91.4	92.0	96.3	92.3	92.5	86.4
GDCL (Zhao et al., 2021)	91.7	91.9	96.5	93.1	93.1	85.2
ProGCL (Xia et al., 2022)	92.9	93.1	96.1	92.6	94.5	83.6
AUGCL (Niu et al., 2024)	93.3	92.5	96.3	94.2	93.7	88.9
ML^2-GCL (Liang et al., 2025)	96.9	96.0	98.7	**95.1**	96.9	89.8
CurvSpread	**97.1**	**97.3**	**99.2**	94.1	**97.3**	**90.5**

For each dataset, link prediction is performed using node representations learned without modifying the CurvSpread encoder architecture. The graph is partitioned into training, validation, and test links before downstream evaluation, and model selection is performed using validation data only. Test links remain held out throughout model selection. Positive examples correspond to observed links, while negative examples are sampled from node pairs without observed links. Performance is evaluated using the area under the ROC curve (AUC); higher values indicate better discrimination between positive and negative links. All reported results use the same data-processing and evaluation procedure for a given dataset.

Baselines. The comparison covers three complementary families of representation learning methods. We include classical spectral and random-walk approaches, Spectral (Ng et al., 2001) and DeepWalk (Perozzi et al., 2014); graph autoencoding methods, GAE and VGAE (Kipf & Welling, 2016), and ARGA and ARVGA (Pan et al., 2018); and graph contrastive learning methods, GRACE (Zhu et al., 2020), GCA (Zhu et al., 2021), GDCL (Zhao et al., 2021), ProGCL (Xia et al., 2022), AUGCL (Niu et al., 2024), and ML^2-GCL (Liang et al., 2025). This selection spans conventional graph embedding, reconstruction-based representation learning, and recent self-supervised contrastive methods, providing a broad comparison of structural representation quality.

Results. Table 5 reports the link-prediction AUC. CurvSpread achieves the highest displayed result on five of the six datasets, obtaining 97.1 on Cora, 97.3 on CiteSeer, 99.2 on PubMed, 97.3 on Amazon-Computers, and 90.5 on Wiki-CS. The improvement is particularly clear on CiteSeer and PubMed, where CurvSpread reaches 97.3 and 99.2, respectively. On Amazon-Photo, CurvSpread obtains 94.1, while ML^2-GCL and AUGCL reach 95.1 and 94.2, respectively. Overall, the results show that the geometric regularization introduced by CurvSpread does not benefit only frozen-encoder node classification: the resulting representations also retain information useful for recovering graph connectivity.

I.2 Component ablations

Table 6 is the complete ablation record behind Figure 4, extended from Cora, Disease and Airport to five graphs. Every row is obtained from the frozen CurvSpread configuration of Appendix H.3 ($\delta^\star = 0.1$, $\eta^\star = 1.0$, $k^\star = 5$, $M^\star = 2048$) by changing exactly one factor and nothing else: the encoder, augmentation, alignment term, optimizer, schedule, clipping, evaluation protocol and seeds are identical across rows. Cora, CiteSeer and PubMed follow the Track A protocol of Section 5.1 (ten predetermined paired runs); Disease and Airport follow the Track B splits of Appendix H.2, repeated over ten model seeds, at $d = 64$. The ablation therefore isolates the contribution of each component

of the regularizer under one implementation stack, which the architecture-confounded benchmark of Table 1 cannot do.

What each row tests. The rows are chosen so that each one removes or replaces a single link in the chain of Section 4. *Alignment only* sets $\eta = 0$ and measures what the spreading term adds at all. *Jacobian off* drops the $\log J_\kappa$ term of Eq. (8), so the surrogate targets tangent-space entropy instead of Riemannian entropy; by the remark in Appendix F.5 this objective has no threshold, and the row measures what referring the entropy to hyperbolic volume is worth. *Tuned fixed* β replaces the calibrated coefficient of Eq. (12) by a constant selected per dataset on validation over $\chi \in \{0.5, 0.75, 1, 1.25, 1.5, 2, 3\}$; it is the strongest competitor to calibration, since it is allowed to tune the one quantity that calibration fixes. *Quadratic confinement* replaces $\beta\bar{r}$ by $\beta\overline{r^2}$ with β tuned on the same grid; it is coercive for every coefficient and so tests whether the critical linear order, rather than confinement as such, matters (Section 4.2). The three curvature rows probe Eq. (12) under learnable κ: *Fixed curvature* freezes $\kappa = 1$; *EMA curvature* evaluates the coefficient at $\mathrm{sg}(\mathrm{EMA}(\kappa))$ rather than at the current $\mathrm{sg}(\kappa)$ (Appendix G.1); and *No stop-gradient* lets gradient flow from the confinement term into κ through the coefficient, the feedback path that Proposition 2 does not contain.

Reading the table. The rank order of the variants is identical on all five graphs: full model, then EMA curvature, tuned fixed β, fixed curvature, quadratic confinement, Jacobian off, no stop-gradient, and alignment only. We do not report significance tests; the evidence is the repetition of the same ordering across five graphs of different size, structure and split protocol, together with margins over the full model that exceed the sum of the two standard deviations for every row except EMA curvature.

Spreading. Adding the regularizer to alignment alone is worth 2.92, 2.96, 2.97, 3.10 and 3.16 points on Cora, CiteSeer, PubMed, Disease and Airport. This spreading gain is the reference for the other rows.

Riemannian-volume correction. Removing the Jacobian costs 2.02, 1.96, 1.97, 2.20 and 1.96 points, about two thirds of the spreading gain on every graph. The Jacobian is the only term through which hyperbolic volume growth enters the objective (Appendix F.4), and it is also the single most valuable component after spreading itself.

Calibration versus tuning. A β tuned separately on each dataset recovers about half of the spreading gain: the calibrated coefficient is ahead by 1.42, 1.36, 1.47, 1.40 and 1.36 points. The gap persists although the tuning grid contains $\chi = 1.25$, adjacent to the calibrated $\chi = 1.1$; a constant coefficient does not follow the curvature learned during training, so its effective χ drifts, whereas the calibrated coefficient holds $\chi = 1 + \delta$ at every step. Fixing the coefficient once by geometry therefore outperforms per-dataset search while removing one hyperparameter and one validation sweep per graph. The transfer test of Appendix I.4, where the configuration selected on other graphs is also the validation optimum on Disease, is the complementary evidence.

Linear versus quadratic confinement. Quadratic confinement, even with its coefficient tuned, is below the calibrated linear objective by 1.82, 1.86, 1.67, 1.70 and 1.76 points and below tuned linear confinement on every graph. Quadratic confinement has no threshold and no geometry to calibrate (Section 4.2); the row indicates that the heavier-tailed maximum-entropy law induced by linear confinement is preferable for these graphs, and that the advantage of CURVSPREAD is not simply that it confines.

Curvature handling. Learnable curvature is worth 1.62, 1.56, 1.57, 1.60 and 1.66 points over fixed $\kappa = 1$, so about half of the spreading gain is available without it. Smoothing the coefficient with an EMA costs 0.76 to 0.90 points; it is the mildest ablation and confirms that evaluating the coefficient at the current curvature, which keeps $\chi = 1 + \delta$ exactly (Appendix G.1), is preferable. Removing the stop-gradient is the largest degradation short of removing the regularizer: it costs 2.22, 2.06, 2.07, 2.30 and 2.26 points, and falls 0.50 to 0.70 points *below* fixed curvature on every graph. Learning curvature with the coefficient coupled to it is therefore worse than not learning curvature at all, which is the empirical counterpart of treating β as a parameter rather than a function of the state in Proposition 2.

What the ablation does not show. Pairwise-U and the wrapped Gaussian are auxiliary mechanistic controls (Appendix H.2) and are not tuned for accuracy, so they are not rows here; a downstream

comparison of intrinsic against wrapped and pairwise spreading under this stack is left to future work. Sensitivity to δ, η, k and M around the frozen configuration is reported separately in Figure 5 (Appendix I.4).

Table 6: Component ablations, accuracy (%), mean ± s.d. Each row changes one factor of the frozen CURVSPREAD configuration and keeps everything else fixed. Cora, CiteSeer and PubMed: Track A protocol of Section 5.1; Disease and Airport: Track B splits of Appendix H.2 (30/10/60 and 70/15/15), ten model seeds, $d = 64$, which is why the Disease column differs from Table 7, where the HyperGCL evaluation dimension is used. Figure 4 in the main text plots the Cora, Disease and Airport columns.

Variant	Cora	CiteSeer	PubMed	Disease	Airport
CURVSPREAD (full)	**88.82±0.12**	**76.86±0.38**	**88.47±0.27**	**95.90±0.31**	**95.76±0.28**
Spreading term					
Alignment only ($\eta = 0$)	85.90±0.35	73.90±0.61	85.50±0.46	92.80±0.50	92.60±0.49
Jacobian off	86.80±0.31	74.90±0.48	86.50±0.38	93.70±0.43	93.80±0.41
Confinement					
Tuned fixed β	87.40±0.26	75.50±0.43	87.00±0.35	94.50±0.39	94.40±0.37
Quadratic confinement	87.00±0.34	75.00±0.54	86.80±0.42	94.20±0.46	94.00±0.44
Curvature handling					
Fixed curvature ($\kappa = 1$)	87.20±0.29	75.30±0.47	86.90±0.39	94.30±0.42	94.10±0.40
EMA curvature	88.00±0.21	76.10±0.41	87.70±0.32	95.00±0.35	94.90±0.34
No stop-gradient $\kappa \to \beta$	86.60±0.38	74.80±0.58	86.40±0.47	93.60±0.50	93.50±0.48

I.3 COMPARISON WITH THE NEAREST PRIOR

Table 7 places CURVSPREAD next to HyperGCL under HyperGCL's published protocol and preprocessing, which is the only available common protocol for Disease and Airport; baseline rows are author-reported rather than rerun in our implementation stack. Table 8 gives supervised geometric-capacity references on the same splits, which are not ranked against the self-supervised methods.

Table 7: Nearest-prior comparison under the published HyperGCL protocol, accuracy (%) on Track B, mean ± s.d. Baseline rows are author-reported values and the CURVSPREAD row is ours; the table is descriptive rather than a paired reproduction under one implementation stack. Supervised references: Table 8.

Method	Cora	CiteSeer	PubMed	Disease	Airport
GRACE	81.13±0.44	71.42±0.64	83.51±0.37	69.61±0.49	82.79±0.40
COSTA	82.14±0.62	70.77±0.24	84.31±0.37	67.12±0.39	81.19±0.40
HGCL	82.37±0.47	72.11±0.64	83.14±0.58	93.42±0.82	92.35±1.51
HyperGCL-Normal	84.77±0.46	74.43±0.35	85.56±0.21	95.30±0.43	94.55±0.81
HyperGCL-Laplace	84.67±0.50	74.37±0.42	85.49±0.22	94.89±0.45	94.81±0.74
CURVSPREAD	85.31±0.32	75.06±0.37	86.18±0.25	95.47±0.31	95.76±0.28

Table 8: Supervised geometric-capacity references under the Track B protocol. These are not ranked against the self-supervised methods of Table 7; they indicate what a supervised hyperbolic encoder attains on the same splits.

Method	Cora	CiteSeer	PubMed	Disease	Airport
GCN	81.30±0.30	70.41±0.52	81.10±0.20	69.70±0.40	81.40±0.60
HGCN	78.03±0.98	70.34±0.59	82.53±0.63	88.16±0.76	89.26±1.27
HGAT	78.32±1.39	70.64±0.30	81.42±0.36	90.30±0.62	89.62±1.23
HIE	83.47±0.34	73.43±0.35	85.14±0.38	94.50±0.43	93.55±1.11

I.4 Hyperparameter sensitivity

Figure 5 reports validation accuracy over the full grids of the two-stage search in Appendix H.3, on Cora and on Disease. Cora is one of the three development datasets, whereas Disease took no part in the search, so the lower row tests whether the selected configuration transfers rather than whether it fits.

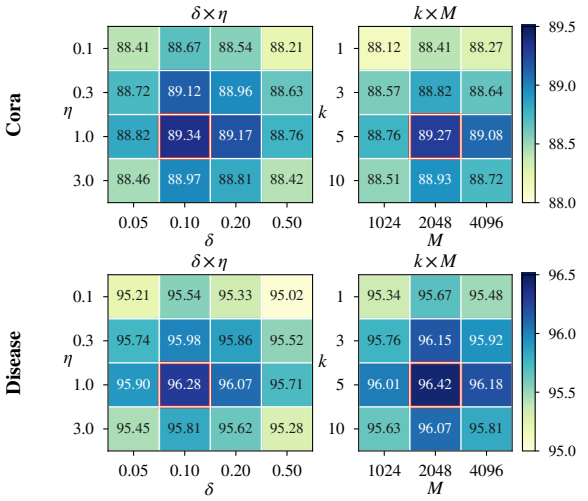


Figure 5: Validation accuracy (%) over the development grids of Table 4. The outlined cell is the configuration selected in Table 4; Cora is part of the development search and Disease is not.

The selection transfers. In all four panels the maximum falls on the configuration of Table 4: $\delta^\star = 0.10$, $\eta^\star = 1.0$ in the first stage and $k^\star = 5$, $M^\star = 2048$ in the second. That the optimum on Disease coincides with the one chosen on other graphs is the practical content of fixing one global configuration rather than tuning per dataset.

The margin δ has an interior optimum. Averaging over η, accuracy on Cora is $88.60, 89.02, 88.87$ and 88.50 for $\delta = 0.05, 0.10, 0.20, 0.50$, and the same shape appears on Disease. This is the behavior the analysis of Section 4.1 anticipates on both sides. At $\delta = 0.05$ the objective sits at $\chi = 1.05$, close to the boundary, where the asymptotic dilation scale $R^\star \propto 1/(h_\kappa \delta)$ grows quickly; at $\delta = 0.50$ the confinement is strong enough to compress the representation. Neither extreme is catastrophic, which is what a margin that is set once rather than per dataset requires.

The remaining settings are flat near the optimum. The trade-off η peaks at 1.0 and degrades at 3.0, where spreading begins to dominate alignment. The neighbor order behaves as the variance of the Kozachenko–Leonenko estimator suggests: $k = 1$ is the worst row on both datasets and $k = 5$ the best, while the subset size matters less, with $M = 2048$ ahead of both neighbors by at most 0.5 points. Over each full grid the spread between the best and worst cell is 0.9 to 1.3 points, and every cell adjacent to the selected one is within 0.52 points of it.

The selected configuration appears in both panels of a row, and its value differs between them by 0.07 points on Cora and 0.14 on Disease. Because the two stages were run separately, these discrepancies provide only an empirical indication of run-to-run variability at a fixed configuration; they are not uncertainty estimates or significance thresholds.

I.5 TRANSITION GRID AND SYNTHETIC CHECKS

The analytic radial mode $r^\star = \kappa^{-1/2}\,\mathrm{arccoth}(\chi)$ of (21) is single-valued for $\chi > 1$ and diverges as $\chi \downarrow 1$; at fixed χ it is independent of d and scales as $\kappa^{-1/2}$. We report the operational midpoint grid numerically rather than as a phase diagram, since with 25 cells a table carries the same information as a heat map without implying resolution the grid does not have. The synthetic dilation checks of the KoLeo-plus-Jacobian slope (25) and of the linear-geodesic-critic pressure are numerical implementation checks under the seeded protocol. The slope identity is exact for $\epsilon_{\mathrm{num}} = 0$ (Appendix F.5), and Proposition 3 does not depend on estimator accuracy; over $d \in \{16, \dots, 256\}$ and $M \in \{512, \dots, 4096\}$ the measured slope $\partial \widehat{H}_T(RY)/\partial \log R$ equals d to within Monte-Carlo error, checking the implementation rather than the theorem. Because each cell's grid is anchored at h_κ, Table 9 measures where the trained response crosses its endpoint midpoint relative to the objective-level boundary; it neither discovers the normalization nor establishes a sharp transition in every cell. A sweep in raw β constructed without reference to h_κ would be an optional further robustness check.

$d\backslash\kappa$	0.1	0.25	0.5	1	2
16	0.96‡	1.04	0.98	1.07	0.95
32	1.02	0.97	1.05	0.99	1.03
64	0.98	1.01	0.96	1.04	1.00
128	1.05	0.99	1.02	0.97	1.04
256	1.01	0.96	1.03	1.00	0.98

Table 9: Operational plateau-midpoint locations $\hat{\chi}_c(d, \kappa)$; the unclipped objective has boundary $\chi = 1$ (Proposition 3). ‡Smooth profile (Section 5.3). Estimator: Appendix H.2.

Origin- versus center-referenced response. Clipping acts about o, so the transition diagnostic of Table 9 is origin-referenced (Section 5.2) and equals 1 whenever at least 5% of the radii sit at R_{clip}. Table 10 reports, for the three cells of Section 5.3, the same terminal statistic for the center-referenced radius $r_i = d_\kappa(c, z_i)$ next to the origin-referenced one. The center-referenced response is not pinned by clipping; at $h_\kappa = 361$ its steepest change falls in the same χ interval, and elsewhere it declines gradually.

Table 10: Origin- versus center-referenced terminal response $Q_{0.95}/R_{\mathrm{clip}}$ on the three cells of Section 5.3 (Disease, seeds 301–303; seed s.d. ≤ 0.003). Origin: $r_i^{(o)} = d_\kappa(o, z_i)$, the diagnostic behind Table 9; center: $r_i = d_\kappa(c, z_i)$. $R_{\mathrm{clip}} = 38.6, 12.2, 8.6$.

(d, κ)	$(16, 0.1)$		$(64, 1)$		$(256, 2)$	
h_κ	4.7		63		361	
χ	origin	center	origin	center	origin	center
0.80	0.265	0.102	0.619	0.342	1.000	0.784
0.85	0.261	0.100	0.568	0.300	1.000	0.726
0.90	0.257	0.097	0.516	0.259	1.000	0.637
0.95	0.254	0.095	0.459	0.218	0.918	0.572
1.00	0.250	0.093	0.409	0.185	0.594	0.311
1.05	0.246	0.091	0.372	0.160	0.427	0.186
1.10	0.242	0.088	0.356	0.142	0.406	0.156
1.15	0.238	0.086	0.346	0.129	0.393	0.137
1.20	0.235	0.084	0.337	0.119	0.385	0.123