UNIFORMITY IS RELATIVE: A MAXIMUM-ENTROPY PRINCIPLE ACROSS REPRESENTATION GEOMETRIES

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental objective of self-supervised learning is to learn representations that preserve semantic consistency while occupying the representation space in a sufficiently dispersed yet non-concentrated manner. Existing notions of uniformity are tied to particular representation spaces, leaving unclear how to define an appropriate reference distribution across different representation spaces. We resolve this by defining uniformity as maximum entropy with respect to intrinsic Riemannian volume under a fixed geodesic scale. This makes uniformity a joint property of geometry and scale, recovering the canonical spherical and Euclidean references while yielding a distinct curved-space target. Meanwhile, tangent-space Gaussian transplantation misses curvature-dependent volume growth, leaving an irreducible Kullback–Leibler divergence gap even after scale retuning. A radial-angular decomposition yields tractable minibatch objectives, and experiments verify the predicted curvature-dependent mismatch in learned geometry, reveal its effects on radial and hierarchical organization, and demonstrate complementary roles for intrinsic uniformity and cross-view alignment.

1 INTRODUCTION

The global organization of representation space is a fundamental concern in self-supervised learning (SSL). Yet representation alignment and dispersion do not determine the appropriate global distribution over a given representation space. This issue becomes particularly important when the representation space has an explicit geometric structure. For example, hyperbolic spaces provide useful inductive biases for hierarchical data (Nickel & Kiela, 2017; Gu et al., 2019), yet in such non-compact spaces, a dispersion objective mismatched to the underlying geometry can drive representations toward the boundary and induce under-utilization of the representation space (Zhang et al., 2025), as illustrated in Figure 1.

Existing work has developed principled notions of uniformity or reference distributions for different representation geometries, but these formulations remain largely geometry-specific. On the hypersphere, contrastive learning motivates the uniform measure (Wang & Isola, 2020; Fang et al., 2024; Nicollier et al., 2026). In Euclidean space, the isotropic Gaussian has likewise been identified as a theoretically motivated embedding distribution (Balestriero & LeCun, 2025), while non-Euclidean representations are commonly constructed by transporting tangent-space Gaussians through the exponential map (Nagano et al., 2019; Mathieu et al., 2019). These approaches, however, are developed within separately prescribed geometries, leaving open whether the appropriate reference distribution can be derived from a common principle as the representation geometry changes.

Motivated by this unification gap, we introduce a variational formulation in which the uniformity reference is jointly determined by representation geometry and scale. Given a geometry and a representation scale, we define the uniformity reference as the maximum-entropy distribution with respect to the intrinsic Riemannian volume (Pennec, 2006; Arvanitidis et al., 2016). The resulting family of distributions provides a unified treatment of both compact and non-compact constant-curvature spaces, recovering standard spherical and Euclidean references while extending them intrinsically to negative curvature (Said et al., 2014; Mathieu et al., 2019).

We further characterize the discrepancy between the intrinsic reference and tangent-space transplantation. The volume Jacobian of the exponential map induces a curvature correction that cannot be

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

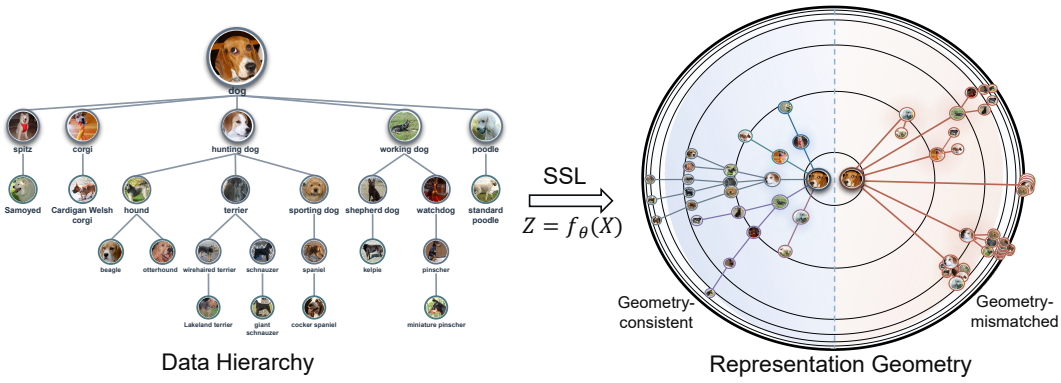


Figure 1: Illustration of how data structure is organized under geometry-dependent uniformity in representation space. For the hierarchical dog taxonomy shown here, a geometry-consistent reference supports coherent radial organization, whereas a mismatched reference distorts the hierarchy in the learned geometry. Black concentric curves denote equal geodesic-radius contours.

removed by scale retuning, yielding a nonzero Kullback–Leibler (KL) divergence gap with a fourth-order near-flat asymptotic. Exploiting isotropy, we exactly decompose the KL divergence into radial and conditional-angular terms and derive a tractable minibatch surrogate. Numerical and learning experiments support the theory: the discrepancy is reflected primarily in radial structure and hierarchical geometry, has only modest effects on downstream discriminability, and vanishes toward the flat limit. Additional from-scratch pretraining and controlled-perturbation experiments further demonstrate that the proposed surrogates are sensitive to distinct distributional failure modes.

Our contributions are summarized as follows:

- We extend the classical alignment–uniformity perspective by formulating uniformity through a variational principle that incorporates both representation geometry and scale, providing a unified notion of representation dispersion across curvatures.
- We demonstrate that tangent-space transplantation induces an irreducible KL gap from the intrinsic reference. This gap persists even after scale retuning, arising from curvature-dependent volume distortion and vanishing only in the flat limit.
- We show that geometry-consistent uniformity shapes the global organization of representation space, improving radial occupancy and hierarchical structure while acting complementarily to cross-view alignment.

2 Related Work

Spherical and Euclidean uniformity. Existing studies have established geometry-specific uniformity references in two canonical representation spaces: the normalized volume measure on the sphere (Wang & Isola, 2020; Fang et al., 2024; Nicollier et al., 2026) and the isotropic Gaussian in Euclidean space (Balestriero & LeCun, 2025). Although these references can be enforced through pairwise potentials, Wasserstein distances, or projected distribution matching, they are formulated separately for pre-specified representation spaces, leaving unclear how the corresponding constructions should vary with curvature and representation scale.

Entropy and distributional regularization. This line of work promotes representation dispersion through entropy maximization, including nearest-neighbor entropy estimation (Sablayrolles et al., 2019), coding-rate and spectral-entropy surrogates (Yu et al., 2020; Liu et al., 2022; Ozsoy et al., 2022; Kim et al., 2023; Yerxa et al., 2023), and latent distribution matching (Mikulasch & Zenke, 2026). Related theoretical studies have further revealed connections between contrastive and non-contrastive objectives through Gram-matrix and spectral structures (Balestriero & LeCun, 2022; Garrido et al., 2023; Zhang et al., 2024), and shown that ideal representations and their implicit distributional priors can constitute important inductive biases (Dubois et al., 2022; Assran et al., 2023).

These works primarily characterize or regularize distributions within a pre-specified representation geometry, rather than deriving the reference distribution from the intrinsic geometry.

Non-Euclidean reference distributions. Work in non-Euclidean spaces has largely focused on transplanting probabilistic models onto prescribed manifolds. A common probabilistic construction pushes a tangent-space Gaussian forward through the exponential map, yielding a wrapped normal distribution (Nagano et al., 2019; Mathieu et al., 2019); related tangent-space constructions and isotropic regularization have also been used in hyperbolic representation learning (Desai et al., 2023; Ge et al., 2023; Zhang et al., 2025). Meanwhile, sliced Wasserstein methods reduce the computational cost of distribution comparison through one-dimensional projections (Bonneel et al., 2015; Kolouri et al., 2019), and have been extended to spherical and hyperbolic spaces, as well as broader classes of manifolds (Bonet et al., 2023a;b; 2025).

3 Preliminaries

3.1 Problem Setup for Self-Supervised Representation Learning

Let $\mathcal{X}$ denote the sample space and p_{data} the data distribution. For each original sample, stochastic data augmentations generate a positive pair $(X, X^+) \sim p_{\mathrm{pos}}$, whose two marginal distributions are both p_{data}. An encoder $f_\theta : \mathcal{X} \to \mathcal{E}$ maps each view to a representation in the representation space $\mathcal{E}$, yielding $Z = f_\theta(X)$ and $Z^+ = f_\theta(X^+)$. The objective of such self-supervised learning (SSL) methods can be summarized by two complementary requirements (Wang & Isola, 2020). Alignment encourages the encoder to map paired views to nearby representations. Let $d_{\mathcal{E}}$ denote a distance or dissimilarity measure on $\mathcal{E}$; the alignment objective can then be written as

$$\mathcal{L}_{\mathrm{align}}(\theta) = \mathbb{E}_{(X,X^+)\sim p_{\mathrm{pos}}}\left[d_{\mathcal{E}}\big(f_\theta(X), f_\theta(X^+)\big)\right]. \tag{1}$$

Uniformity, in contrast, requires the representation distribution to remain sufficiently dispersed to avoid collapse. Let $\mu_\theta = (f_\theta)_{\#}p_{\mathrm{data}}$ denote the marginal representation distribution induced by the encoder. This can be instantiated in different ways, including distribution matching, entropy-based objectives, and second-order statistical constraints; we discuss their connections in Appendix A. In the following, we focus on the explicit-reference-distribution perspective. Given a sufficiently dispersed reference distribution $\nu_{\mathcal{E}}$ that is compatible with the geometry of the representation space and the prescribed scale constraint, the corresponding objective can be written as

$$\mathcal{L}_{\mathrm{unif}}(\theta) = \mathfrak{D}_{\mathcal{E}}(\mu_\theta, \nu_{\mathcal{E}}), \tag{2}$$

where $\mathfrak{D}_{\mathcal{E}}$ denotes an appropriate distributional discrepancy.

3.2 Intrinsic Volume Geometry of Uniformity

Let $\mathcal{E}_{K,d}$ denote a representation space of intrinsic dimension d and constant sectional curvature K, let $o \in \mathcal{E}_{K,d}$ be a reference point, and let $r(z) = \mathrm{dist}(z, o)$ denote the geodesic distance from z to o. Different representation constraints instantiate $\mathcal{E}_{K,d}$ with different geometries: ℓ_2 normalization corresponds to spherical geometry (with $K = 1$ giving the unit sphere), unnormalized representations reside in Euclidean space ($K = 0$), and hyperbolic representations can be modeled using, e.g., the Poincaré ball model ($K < 0$) (Nickel & Kiela, 2017). In geodesic polar coordinates, the Riemannian volume element on $\mathcal{E}_{K,d}$ and its radial shell factor (Lee, 2018) are given by

$$dV_K = s_K(r)^{d-1}\, dr\, d\Omega, \qquad s_K(r) = \begin{cases} \sin(\sqrt{K}r)/\sqrt{K}, & K > 0, \\ r, & K = 0, \\ \sinh(\sqrt{|K|}r)/\sqrt{|K|}, & K < 0. \end{cases} \tag{3}$$

Accordingly, the volume of a geodesic ball of radius R is $V_K(R) = \omega_{d-1}\int_0^R s_K(r)^{d-1}dr$, where $\omega_{d-1} = |\mathbb{S}_{\varkappa}{}^{d-1}|$, $\mathbb{S}_{\varkappa}{}^{d-1}$ denotes the unit $(d-1)$-sphere, and $d\Omega$ is its surface-area element.

Whether the Riemannian volume measure can be normalized into a uniform probability measure depends on whether the total volume of the space is finite. When $K > 0$, the geodesic radius is bounded above by $\pi/\sqrt{K}$ and the total volume satisfies $V_K(\pi/\sqrt{K}) < \infty$. The Riemannian volume measure can therefore be normalized, yielding the uniform probability distribution

on the corresponding hypersphere. In contrast, when $K \leq 0$, the geodesic radius is unbounded and $\lim_{R\to\infty} V_K(R) = \infty$, so no uniform probability distribution over the entire space can be obtained by directly normalizing the Riemannian volume measure. Thus, once the representation geometry is specified, the key question is how to construct the reference distribution $\nu_{\mathcal{E}}$ from the intrinsic geometry of the representation space.

4 Maximum Entropy Across Constant-Curvature Geometries

This section characterizes the intrinsic uniformity reference in constant-curvature representation spaces from an intrinsic maximum-entropy perspective. Under a fixed geodesic scale budget, we seek the maximum-entropy distribution relative to the Riemannian volume measure, yielding a unified reference jointly determined by the curvature K and the scale ρ. We further study its recovery properties across curvatures and its geometric discrepancy from tangent-space constructions.

Let $H(\mu) = -\int_{\mathcal{E}K,d} p(z)\log p(z), dV_K(z)$ denote the entropy of a distribution μ with density $p = d\mu/dV_K$ relative to the Riemannian volume measure dV_K, and let $r(z) := \mathrm{dist}(z, o)$ denote the geodesic distance from z to the reference point o. Given $\rho > 0$, we constrain the dispersion scale by the second moment of the geodesic distance to o. Accordingly, for K and ρ, we define the uniformity reference as the intrinsic maximum-entropy distribution $\nu K, \rho$:

Definition 1 (Intrinsic MaxEnt Reference)**.** Given a curvature K and a scale budget $\rho > 0$, define the intrinsic MaxEnt reference $\nu_{K,\rho}$ on $\mathcal{E}_{K,d}$ as

$$\nu_{K,\rho} = \arg\max_{\mu\in\mathcal{P}(\mathcal{E}_{K,d})} H(\mu) \quad \text{s.t.} \quad \mathbb{E}_\mu[r(z)^2] \leq \rho. \tag{4}$$

A Lagrangian-duality argument (see Appendix B.1) shows that the maximization problem above admits a unique solution, given by a Gibbs density that is isotropic about o:

$$p_{K,\rho}(z) = Z_\lambda^{-1}\exp\big(-\lambda r(z)^2\big), \qquad Z_\lambda := \int_{\mathcal{E}_{K,d}} \exp\big(-\lambda r(\xi)^2\big)\, dV_K(\xi), \qquad \lambda \geq 0. \tag{5}$$

Here, λ is the optimal dual variable jointly determined by K, d, and ρ. For $\lambda > 0$, the density in Eq. 5 is a Riemannian normal distribution, whose maximum-entropy characterization under a fixed Fréchet mean and dispersion is well known (Pennec, 2006; Said et al., 2014); it has also been used as a prior in hyperbolic latent spaces (Mathieu et al., 2019).

Proposition 1 (Canonical recoveries)**.** The intrinsic MaxEnt reference recovers the canonical reference distributions associated with the two standard representation geometries.

(i) For $K > 0$, let σ_K denote the normalized Riemannian volume measure on $\mathcal{E}_{K,d}$, i.e., the uniform distribution on the corresponding hypersphere, and define $\rho_{\mathrm{unif}}(K, d) = \mathbb{E}_{\sigma_K}[r(z)^2]$. If $\rho \geq \rho_{\mathrm{unif}}(K)$, then $\nu_{K,\rho} = \sigma_K$ and $\lambda = 0$, recovering the classical hyperspherical uniformity target (Wang & Isola, 2020).

(ii) For $K = 0$, identifying o with the Euclidean origin and letting I_d denote the $d \times d$ identity matrix, we have:

$$\nu_{0,\rho} = \mathcal{N}\left(0, \frac{\rho}{d} I_d\right), \qquad \lambda = \frac{d}{2\rho}, \tag{6}$$

recovering the isotropic Gaussian reference in Euclidean representation space (Balestriero & LeCun, 2025).

These two recovery results show that the scale constraint plays different roles across geometries: when finite volume already yields a normalizable uniform measure, the scale constraint can become inactive; in non-compact spaces, by contrast, it is necessary for obtaining a normalizable maximum-entropy reference. Thus, this construction places the existing spherical and Euclidean references in SSL under a common variational principle, and extends the same principle to negative-curvature representation spaces as a uniformity reference.

5 Comparison with Tangent-Space Uniformity

Having defined the Intrinsic MaxEnt Reference in the previous section, we now examine its discrepancy from the reference distribution induced by tangent-space Gaussian transplantation. As shown

in Figure 2, the two references are constructed respectively from a transplanted tangent-space Gaussian and the intrinsic volume structure of the manifold. We focus on how curvature alters their radial distributions and whether the resulting discrepancy persists after adjusting the scale of the tangent-space distribution.

The density $p_{K,\rho}$ in Eq. 5 is defined with respect to the Riemannian volume measure dV_K. Substituting Eq. 3 and integrating out the angular variables, the Intrinsic MaxEnt Reference induces the radial density $q_{K,\rho}(r) \propto s_K(r)^{d-1} \exp(-\lambda r^2)$. Existing Gaussian-transplant objectives instead place an isotropic Gaussian in the tangent space $T_o\mathcal{E}_{K,d} \cong \mathbb{R}^d$ and push it forward onto the manifold through the exponential map, yielding the radial density $q_{\tan}(r) \propto r^{d-1} \exp(-\lambda' r^2)$. Their ratio is therefore given by the product of the volume Jacobian of the exponential map and the Gaussian factor (Lee, 2018; Zhang et al., 2025):

$$\frac{q_{K,\rho}(r)}{q_{\tan}(r)} \propto J_K(r) \cdot e^{-(\lambda-\lambda')r^2}, \qquad J_K(r) := \left(\frac{s_K(r)}{r}\right)^{d-1}. \tag{7}$$

Equation 7 separates the geometric factor from the parametric factor and reveals the structure of the gap induced by transplantation. Specifically, taking the difference between the log-densities of the two radial distributions and Taylor-expanding the curvature-dependent Jacobian term around the near-flat regime of the dimensionless curvature–radial-dispersion parameter $\tau := |K|\rho \to 0$ yields:

$$\log J_K(r) = -\frac{d-1}{6} K r^2 + \Theta(K^2 r^4). \tag{8}$$

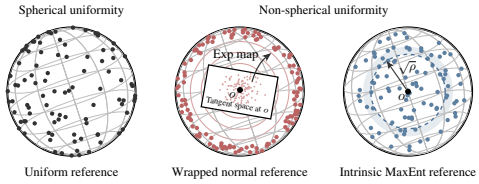


Figure 2: Schematic illustration of uniformity references. left: the spherical uniform reference; middle: the wrapped normal reference obtained by mapping a tangent-space Gaussian through the exponential map; right: the intrinsic geodesic shell under the scale budget ρ.

Equation 8 shows that, under the same second-moment budget, the family of isotropic tangent-space Gaussians can absorb the r^2 term of the Jacobian through its scale parameter; however, when $d \geq 2$ and $K \neq 0$, the r^4 and higher-order terms cannot be represented by this family, resulting in an irreducible radial mismatch. Propositions 2 and 3 below formalize this observation as, respectively, an exact entropy-transfer relation and a non-negativity characterization of the resulting gap.

Proposition 2 (Entropy Transfer)**.** For any $\mu \in \mathcal{P}(\mathcal{E}_{K,d})$ that is absolutely continuous with respect to dV_K, let $H_{\tan}(\mu)$ denote the entropy of the distribution obtained by mapping μ to the tangent space through the logarithmic map $u = \log_o(z)$. Then

$$H(\mu) = H_{\tan}(\mu) + \mathbb{E}_\mu[\log J_K(r)]. \tag{9}$$

Proof. See Appendix B.2. □

Proposition 3 (Intrinsic–Wrapped Distributional Gap)**.** Let $d \geq 2$, and suppose the scale budget yields $\lambda > 0$. Let $q_{\tan}$ and $q_{K,\rho}$ satisfy the same scale constraint, i.e., $\mathbb{E}_{q_{\tan}}[r^2] = \mathbb{E}_{q_{K,\rho}}[r^2] = \rho$. Define

$$\Delta(K, \rho, d) := D_{\mathrm{KL}}(q_{\tan} \| q_{K,\rho})\,. \tag{10}$$

Then $\Delta(K, \rho, d) \geq 0$, with equality if and only if $K = 0$; this gap cannot be eliminated by retuning the tangent-space scale parameter λ'. Furthermore, for fixed $d \geq 2$, as $\tau \to 0$,

$$\Delta(K, \rho, d) = C_d\, \tau^4 + o(\tau^4), \tag{11}$$

where $C_d > 0$ is an explicit constant depending only on the representation dimension d.

Proof. See Appendix B.3. □

Equation 10 is written in radial form. In the idealized setting, however, both reference distributions are isotropic about o and share the same uniform angular distribution, so the radial KL divergence is exactly equal to the KL divergence between the full distributions. Consequently, the discrepancy

between the tangent-space transplant and the intrinsic reference is entirely carried by the radial volume factor, while the angular component introduces no additional curvature mismatch. After matching the same dispersion budget, the resulting distributional gap exhibits fourth-order scaling in τ in the near-flat limit. According to Section 3.2, τ is bounded when $K > 0$, whereas no finite upper bound exists when $K < 0$; the volume mismatch can therefore continue to accumulate as the radial scale increases. This also provides a geometric perspective on boundary drift and dimensional degeneration in hyperbolic representations.

6 Surrogate Objectives via Radial-Angular Decomposition

This section takes the intrinsic MaxEnt reference $\nu_{K,\rho}$ as the target distribution and constructs an empirical uniformity regularizer suitable for minibatch optimization. The ideal distribution-matching criterion is $D_{\mathrm{KL}}(\mu\|\nu_{K,\rho})$, but directly estimating this KL divergence requires high-dimensional density estimation. Exploiting the isotropy of $\nu_{K,\rho}$, we decompose the objective into a radial mismatch and a conditional angular mismatch, and then use one-dimensional W_2^2 as a tractable surrogate for the radial KL term and spherical sliced-W_2^2 as a surrogate for the conditional angular KL term.

In geodesic polar coordinates centered at o, any $z \in \mathcal{E}_{K,d}$ can be written as $z = \exp_o(r\omega)$, where $\omega = \log_o(z)/r \in \mathbb{S}_o^{d-1}$. Let μ_r denote the radial marginal of μ, and let $\mu_{\omega|r}$ denote the conditional distribution of directions given the radius. By the polar decomposition and the chain rule for KL divergence, we obtain:

Proposition 4 (Radial–Angular Decomposition)**.** For any $\mu \in \mathcal{P}(\mathcal{E}_{K,d})$ that admits a density with respect to $\mathrm{d}V_K$, we have

$$D_{\mathrm{KL}}(\mu\|\nu_{K,\rho}) = D_{\mathrm{KL}}(\mu_r\|q_{K,\rho}) + \mathbb{E}_{r\sim\mu_r}\left[D_{\mathrm{KL}}\left(\mu_{\omega|r}\|\,\mathrm{Unif}(\mathbb{S}_o^{d-1})\right)\right]. \tag{12}$$

Proof. See Appendix B.4. □

The two terms on the right-hand side respectively characterize the radial mismatch and the conditional angular mismatch. We use one-dimensional W_2^2 (Peyré & Cuturi, 2019) and spherical sliced-W_2^2 (Bonet et al., 2023a) as tractable surrogates for these two terms, respectively. Specifically, let $Q_{K,\rho}$ denote the quantile function of $q_{K,\rho}$. For the N representations from one augmented view, sort their geodesic radii as $r_{(1)} \le \cdots \le r_{(N)}$. The radial loss is then

$$\mathcal{L}_{\mathrm{rad}} = \frac{1}{N\rho}\sum_{i=1}^{N}\left[r_{(i)} - Q_{K,\rho}\left(\frac{2i-1}{2N}\right)\right]^2. \tag{13}$$

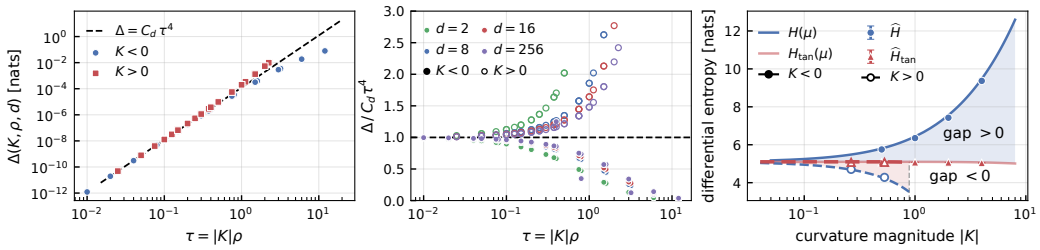
The quantile function $Q_{K,\rho}$ is precomputed offline from $q_{K,\rho}(r) \propto s_K(r)^{d-1}\exp(-\lambda r^2)$.

To approximate the directional distribution conditional on a continuous radius, we partition the representations into B equal-frequency bins I_b according to their empirical radii, and let $n_b = |I_b|$. We sample M slicing directions $e_m \sim \mathrm{Unif}(\mathbb{S}_o^{d-1})$, and let $c_{(j),b,m}$ denote the j-th order statistic of $\{\langle e_m, \omega_i\rangle_o : i \in I_b\}$. By the rotational invariance of the uniform distribution on the sphere, for any fixed e_m, if $\omega \sim \mathrm{Unif}(\mathbb{S}_o^{d-1})$, then $\frac{1+\langle e_m,\omega\rangle_o}{2} \sim \mathrm{Beta}((d-1)/2,(d-1)/2)$ (Mardia & Jupp, 2000; Vershynin, 2018; Nicollier et al., 2026); let $Q_{\mathrm{ang},d}$ denote the quantile function of the raw projection $\langle e_m, \omega\rangle_o$, and the angular loss is then

$$\mathcal{L}_{\mathrm{ang}} = \frac{d}{NM}\sum_{b=1}^{B}\sum_{m=1}^{M}\sum_{j=1}^{n_b}\left[c_{(j),b,m} - Q_{\mathrm{ang},d}\left(\frac{2j-1}{2n_b}\right)\right]^2. \tag{14}$$

The two augmented views share the same set of slicing directions $\{e_m\}_{m=1}^M$ at each optimization step. For the two augmented views of the same sample, a shared encoder produces representations $u_i^{(v)} = f_\theta(x_i^{(v)})$, where $v \in \{1,2\}$. These representations are then mapped to $\mathcal{E}_{K,d}$ through the exponential map centered at o, yielding $z_i^{(v)} = \exp_o(u_i^{(v)})$. The cross-view alignment term between corresponding representations is defined as

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{N}\sum_{i=1}^{N} d_K\left(z_i^{(1)}, z_i^{(2)}\right), \tag{15}$$



(a) (b) (c)

Figure 3: Numerical verification of the distributional gap and entropy transfer. (a) Comparison between $\Delta(K, \rho, d)$ and the fourth-order asymptotic prediction $C_d\tau^4$ at $d = 16$; (b) normalized gap $\Delta/(C_d\tau^4)$ as a function of τ across different dimensions; (c) comparison between the intrinsic entropy $H(\mu)$ and tangent-space entropy $H_{\tan}(\mu)$ at $d = 4, \rho = 3$.

where d_K denotes the unsquared geodesic distance on $\mathcal{E}_{K,d}$. For each augmented view $v \in \{1, 2\}$, we compute $\mathcal{L}_{\text{rad}}^{(v)}$ and $\mathcal{L}_{\text{ang}}^{(v)}$ separately. The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \frac{\alpha}{2} \sum_{v=1}^{2} \left(\mathcal{L}_{\text{rad}}^{(v)} + \mathcal{L}_{\text{ang}}^{(v)} \right), \tag{16}$$

where $\alpha > 0$ controls the strength of the distribution-matching regularizer. The objective consists only of cross-view positive-pair alignment and radial and angular distribution matching; it uses no InfoNCE objective, negative samples, or additional anti-collapse losses. Details of the slicing strategy and numerical tabulation procedure are provided in Appendix C.

7 Experiments

7.1 Numerical Verification of the Intrinsic–Tangent Gap

We first compare the tangent-space transplanted distribution with the intrinsic MaxEnt reference in a controlled synthetic setting. Given (K, ρ, d), we determine the scale parameters of the two distributions under the same second-moment budget and compute their KL gap by numerically integrating the corresponding one-dimensional radial densities. We then jointly vary the curvature K, representation dimension d, and scale budget ρ to examine the scaling behavior of the gap, and compare the intrinsic and tangent-space entropies as functions of curvature for fixed d and ρ.

As shown in Figure 3 (a) and (b), as $\tau \to 0$, Δ agrees with the fourth-order asymptotic prediction $C_d\tau^4$ established in Proposition 3, while the normalized gap $\Delta/(C_d\tau^4)$ approaches 1 across different dimensions and both signs of curvature. As τ increases, the positive- and negative-curvature branches deviate from the local approximation in opposite directions, reflecting the sign dependence of higher-order curvature corrections. Figure 3 (c) further shows that the finite-sample entropy estimates reproduce the sign relation predicted by Proposition 2: $H(\mu) > H_{\tan}(\mu)$ when $K < 0$, whereas the ordering is reversed when $K > 0$. Together, these numerical results verify the asymptotic scaling law of the curvature-induced gap in the near-flat regime and confirm the entropy-transfer behavior induced by the Jacobian volume distortion.

7.2 Downstream Linear Evaluation on Node Classification

To evaluate whether representations learned with the intrinsic MaxEnt reference retain discriminative information for downstream tasks, we conduct transductive node classification on four standard benchmarks commonly used in hyperbolic graph representation learning. All datasets use the self-supervised objective described in Section 6. After pretraining, we freeze the encoder and train a linear node classifier on the encoder representations $\mathbf{U}$.

We report the mean and sample standard deviation of Accuracy, Macro-F1 (Sokolova & Lapalme, 2009), and balanced accuracy (Brodersen et al., 2010) over five random seeds. As shown in Table 1,

Table 1: Node classification performance (%). Accuracy (Acc.), Macro-F1 (M-F1), and balanced accuracy (B-Acc.) are reported as mean ± sample standard deviation over five seeds. Bold values denote the best result in each metric column across all methods.

			Airport			PubMed			CiteSeer			Cora		
Training	Space	Method	Acc.	M-F1	B-Acc.	Acc.	M-F1	B-Acc.	Acc.	M-F1	B-Acc.	Acc.	M-F1	B-Acc.
Supervised	Euclidean	GCN	61.92±0.95	52.79±1.45	56.37±2.02	78.54±0.43	78.30±0.35	78.58±0.64	66.98±0.75	63.65±1.01	63.79±1.06	80.50±1.12	79.32±0.81	80.16±0.60
		GAT	57.57±2.17	34.52±1.17	36.98±1.36	77.10±0.38	76.67±0.36	77.33±0.54	70.46±1.80	65.41±1.65	65.62±1.68	80.50±1.57	79.54±1.22	81.29±0.90
		GraphSAGE	65.10±1.27	56.70±1.14	62.04±0.93	76.46±0.29	75.97±0.38	76.62±0.41	68.68±1.14	64.63±0.60	65.00±0.50	80.18±0.39	79.20±0.20	80.72±0.20
	Hyperbolic	HNN	69.29±2.79	62.15±3.59	64.75±4.51	72.72±1.19	72.82±1.37	72.98±0.54	54.50±2.03	51.74±2.73	52.44±2.40	56.32±2.50	53.42±2.83	54.06±3.13
		HGNN	62.47±2.24	45.46±9.96	47.46±10.21	76.70±1.07	76.10±1.16	76.62±0.67	67.12±0.97	63.80±0.69	64.13±0.89	79.24±0.98	78.31±0.89	79.95±0.91
		HGCN	65.36±0.92	58.27±2.00	62.50±1.89	77.62±0.48	77.47±0.61	77.56±1.15	68.98±1.08	65.05±1.05	65.26±1.12	80.72±0.80	79.64±0.73	80.57±0.70
		HGAT	62.72±1.14	53.61±2.71	59.35±3.13	76.90±1.27	76.54±1.23	77.36±0.79	68.54±0.96	64.60±1.16	65.06±1.19	80.40±0.84	79.52±0.96	81.16±0.45
		HIE	62.85±1.23	52.61±1.97	58.32±1.69	77.58±0.76	77.29±0.74	77.25±0.89	68.94±0.82	65.29±0.64	65.70±0.57	79.88±0.99	78.85±1.24	79.56±0.90
Self-supervised	Euclidean	GRACE	68.62±1.69	64.90±2.41	63.83±2.20	76.94±1.51	76.57±1.13	77.29±1.35	67.20±1.05	62.60±0.82	63.09±0.76	78.74±2.06	76.74±3.18	78.88±3.25
		COSTA	68.08±1.42	64.31±2.00	63.15±2.21	77.36±1.27	76.90±0.90	77.57±1.16	67.16±1.38	62.24±0.96	62.73±0.84	78.70±2.88	76.53±3.58	78.57±3.42
	Hyperbolic	HGCL	67.36±1.16	62.40±2.17	61.61±1.89	77.64±0.80	77.38±0.75	78.06±0.85	68.84±0.96	63.10±2.06	64.08±1.59	76.92±1.18	75.01±1.60	76.86±1.82
		HyperGCL-Laplace	73.85±1.39	71.68±0.99	71.47±1.08	79.50±0.98	78.96±0.98	79.35±0.87	70.34±1.42	66.06±1.63	66.49±1.60	81.54±0.69	80.79±0.61	82.36±0.65
		HyperGCL-Normal	73.01±2.19	71.30±2.34	71.33±2.26	80.56±0.69	**80.11±0.67**	80.18±0.73	69.58±0.86	66.07±0.76	66.60±0.69	81.48±0.60	80.75±0.64	82.23±0.55
		Intrinsic MaxEnt (Ours)	**78.91±1.38**	**78.17±1.56**	**77.11±1.17**	**80.66±0.75**	79.81±0.76	**80.61±0.80**	**71.40±1.50**	**66.40±1.82**	**66.62±1.87**	**82.26±1.28**	**80.96±1.54**	**82.92±1.22**

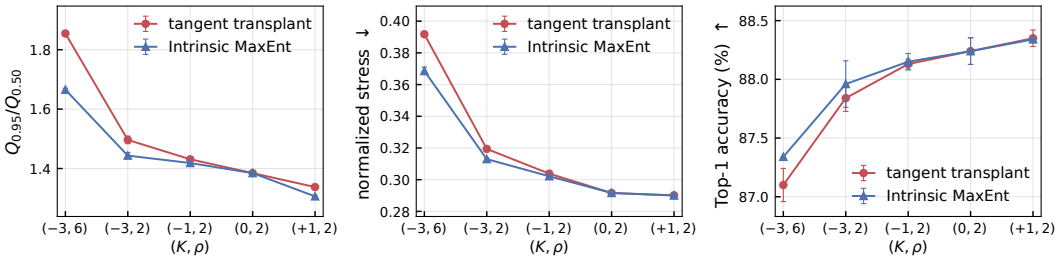


Figure 4: ImageNet-100 matched-reference comparison: (a) radial-tail ratio; (b) hierarchy stress; and (c) linear-probe Top-1 accuracy.

the learned representations achieve the best performance among self-supervised methods on most metrics across the four datasets, indicating that the intrinsic uniformity objective consistently improves the downstream linear separability of the learned representations. Full implementation and evaluation details are provided in Appendix C.2.

7.3 How Does the Gap Manifest in Learned Representations?

We further investigate whether the intrinsic–tangent gap survives representation learning, and if so, how far its effect propagates. We trace this discrepancy along three levels of representation structure: from the radial distribution, to category-level geometry, and finally to downstream discriminability.

We use the ImageNet-100 subset of Ming et al. (2022) together with the WordNet hierarchy (Miller, 1995) adopted in HypStructure (Sinha et al., 2024). Starting from the same supervised pretrained ResNet-34 (Deng et al., 2009; He et al., 2016), we perform label-free adaptation by updating only the final residual block and the linear projection head. For each fixed (K, ρ), the tangent-transplant and intrinsic-MaxEnt runs share the same initialization, data, augmentations, objective, and optimization configuration, and differ only in the radial reference. Class labels and hierarchy information are not used during adaptation and are introduced only afterward for evaluation. The analysis follows the progression below:

Reference mismatch (Tangent vs. Intrinsic) $\longrightarrow$ **Radial structure** ($Q_{0.95}/Q_{0.50}$) $\longrightarrow$ **Hierarchy** (Normalized stress) $\longrightarrow$ **Linear separability** (Top-1 accuracy)

This chain allows us to trace how far the intrinsic–tangent reference gap propagates through the learned representation. **(i)** We first measure whether the reference mismatch remains visible in the radial profile of the manifold representations $\mathbf{Z}$, using $Q_{0.95}/Q_{0.50}$ on a held-out set of 10,000 ImageNet-100 samples. **(ii)** We then test whether this radial difference is accompanied by a change in category-level organization by computing normalized stress between pairwise class-prototype distances in $\mathbf{Z}$ and the corresponding WordNet tree distances. **(iii)** Finally, we freeze the encoder representations $\mathbf{U}$ and train a linear probe, using ImageNet-100 Top-1 accuracy to determine whether the geometric differences also translate into downstream discriminability.

As shown in Figure 4, at $(K, \rho) = (-3, 6)$, where the intrinsic–tangent reference gap is pronounced, intrinsic MaxEnt yields a substantially lower $Q_{0.95}/Q_{0.50}$ than the tangent transplant (1.665 versus 1.855), showing that the reference difference is not absorbed during adaptation but remains clearly visible in the learned radial structure. Moving downstream, the normalized stress decreases from 0.392 to 0.369, indicating that the class-prototype geometry induced by intrinsic MaxEnt is more consistent with the WordNet hierarchy, whereas the improvement in linear-probe Top-1 accuracy remains comparatively modest. As the negative-curvature effect weakens, the differences between the two references in hierarchical geometry and downstream performance also diminish. In particular, at the matched control $K = 0$, the two references produce nearly identical results, consistent with Proposition 3. Full implementation details are provided in Appendix C.3. To further test whether these observations depend on pretrained initialization, we conduct from-scratch SSL pretraining with a randomly initialized encoder, as detailed in Appendix D.

7.4 Interaction Between Alignment and Intrinsic MaxEnt During Training

We next examine how the alignment and intrinsic MaxEnt objectives shape the learned geometry during optimization. Starting from the common ImageNet-100 adaptation setup, we fix the primary configuration at $(K, \rho, d) = (+1, 2, 16)$ with unit weights on both the alignment and intrinsic MaxEnt terms, and vary one factor at a time, including the alignment weight, MaxEnt weight, learning rate, representation dimension, batch size, curvature, and ρ. All other experimental settings are kept fixed. For each of the 15 preregistered configurations, we retain checkpoints at epochs 1, 10, and 20. For each checkpoint, we compute the radial and angular discrepancies, denoted by D_{rad} and D_{ang}, using the same definitions as the radial and angular losses introduced above. The left panel preserves this two-dimensional decomposition, while the right panel additionally reports the normalized effective rank erank $/d$ (Roy & Vetterli, 2007) at epoch 20 as a diagnostic of dimensional collapse.

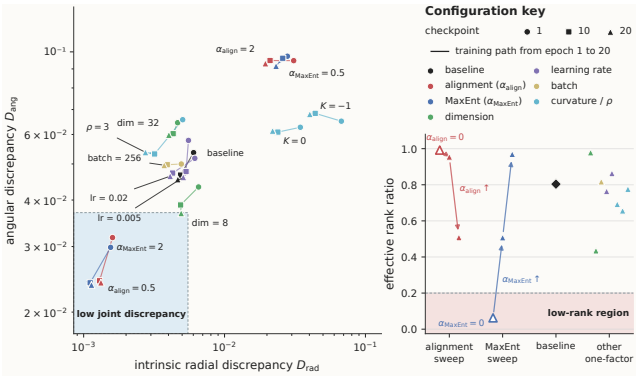


Figure 5: Interaction between alignment and intrinsic MaxEnt under controlled perturbations on ImageNet-100.

Figure 5 shows that alignment and intrinsic MaxEnt regulate complementary aspects of the learned geometry. Removing intrinsic MaxEnt reduces the alignment loss to 0.001 by epoch 20 but increases $D_{\text{rad}} + D_{\text{ang}}$ to 3.382, whereas removing alignment yields the opposite pattern, with an alignment loss of 0.643 and a total geometric discrepancy of 0.003. The equal-weight baseline lies between these extremes at $(0.226, 0.050)$, with intermediate weight settings tracing the corresponding transition. The effective-rank ratio further distinguishes the two endpoints, dropping to 0.063 without intrinsic MaxEnt while remaining at 0.992 without alignment. Other one-factor perturbations remain close to the baseline regime. Overall, these results show that alignment and intrinsic MaxEnt impose complementary rather than redundant constraints on representation geometry.

8 Conclusion

We studied how representation geometry shapes the notion of uniformity through an intrinsic maximum-entropy perspective. Our results show that the curvature-induced mismatch between intrinsic and tangent-space references is not merely a property of the target distributions: when curvature effects are substantial, it remains visible in the learned representations, while disappearing toward the flat limit. Across controlled numerical, graph, and image experiments, intrinsic MaxEnt provides geometry-consistent regularization that complements cross-view alignment. These findings suggest that uniformity cannot be separated from the geometry in which representations are learned. Adaptive curvature selection and extensions to more general representation geometries remain important directions for future work.

AI USE STATEMENT

Generative AI tools served as assistants for background research, grammatical polishing, implementation and editorial work, while the scientific framing, experimental controls, and final scientific and editorial decisions remained with the authors. Portions of the analysis, experimental intervention, and visualization code were developed with AI assistance. The authors reviewed all content and the implementations of the reported experiments, and take full responsibility for the final content of the manuscript and its accompanying research artifacts.

REPRODUCIBILITY STATEMENT

Code and scripts for reproducing our experiments are publicly available at https://anonymous.4open.science/r/anonymous_ICLR_2027/. All datasets are public, and implementation details, hyperparameters, and evaluation protocols are provided in the appendix.

REFERENCES

Georgios Arvanitidis, Lars K Hansen, and Sø ren Hauberg. A locally adaptive normal distribution. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29. Curran Associates, Inc., 2016.

Mido Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 26671–26685, 2022.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

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

Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh-Tan Pham. Spherical sliced-Wasserstein. In *11th International Conference on Learning Representations (ICLR)*, 2023a.

Clément Bonet, Laetitia Chapel, Lucas Drumetz, and Nicolas Courty. Hyperbolic sliced-Wasserstein via geodesic and horospherical projections. In *Topological, Algebraic and Geometric Learning Workshops (TAG-ML)*, pp. 334–370. PMLR, 2023b.

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein distances and flows on Cartan–Hadamard manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124. IEEE, 2010. doi: 10.1109/ICPR.2010.764.

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.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning (ICML)*, pp. 7694–7731. PMLR, 2023.

Yann Dubois, Stefano Ermon, Tatsunori B Hashimoto, and Percy S Liang. Improving self-supervised learning by characterizing idealized representations. In *Advances in neural information processing systems (NeurIPS)*, volume 35, pp. 11279–11296, 2022.

Xianghong Fang, Jian Li, Qiang Sun, and Wang Benyou. Rethinking the uniformity metric in self-supervised learning. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 8359–8380, 2024.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 5350–5360, 2018.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6840–6849, 2023.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *7th International Conference on Learning Representations (ICLR)*, 2019.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 1024–1034, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Jaeill Kim, Suhyun Kang, Duhun Hwang, Jungwook Shin, and Wonjong Rhee. Vne: An effective method for improving deep representation by manipulating eigenvalue distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3799–3810, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964. doi: 10.1007/BF02289565.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer Cham, 2018.

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

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 8228–8239, 2019.

Xin Liu, Zhongdao Wang, Yali Li, and Shengjin Wang. Self-supervised learning via maximum entropy coding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, Chichester, 2000.

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)*, 2019.

Fabian A. Mikulasch and Friedemann Zenke. Understanding self-supervised learning via latent distribution matching. In *International Conference on Machine Learning (ICML)*, 2026. Spotlight.

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. doi: 10.52202/068431-2543.

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

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6341–6350, 2017.

Léo Nicollier, Max Dunitz, Marc Pic, Pablo Musé, Enric Meinhardt-Llopis, and Gabriele Facciolo. Sphere-jepa: Spherical prediction with homogeneous embeddings. *arXiv preprint arXiv:2605.26900*, 2026.

Serdar Ozsoy, Shadi Hamdan, Sercan Arik, Deniz Yuret, and Alper Erdogan. Self-supervised learning with an information maximization criterion. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Jose C. Principe. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer, 2010. doi: 10.1007/978-1-4419-1570-2.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

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

Salem Said, Lionel Bombrun, and Yannick Berthoumieu. New riemannian priors on the univariate normal model. *Entropy*, 16(7):4015–4031, 2014.

Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. Learning structured representations with hyperbolic embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-2895.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

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

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. PMLR, 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)*, pp. 40–48, 2016.

Thomas Yerxa, Yilun Kuang, Eero Simoncelli, and SueYeon Chung. Learning efficient coding of natural images with maximum manifold capacity representations. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021.

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.

Yifan Zhang, Zhiquan Tan, Jingqin Yang, Weiran Huang, and Yang Yuan. Matrix information theory for self-supervised learning. In *International Conference on Machine Learning (ICML)*, pp. 59897–59918, 2024.

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.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.

A Characterizing Uniformity from Three Perspectives

Building on the preceding setup, we synthesize three complementary perspectives from the existing literature to characterize representation uniformity:

(a) Geometric reference measure matching. Hyperspherical uniformity requires μ_θ to match the uniform probability measure $\sigma_{\mathbb{S}}$ on the unit sphere $\mathbb{S}_o^{d-1} \subset \mathbb{R}^d$. (Wang & Isola, 2020; Fang et al., 2024). At the distribution level, this target can be quantified by the Kullback–Leibler divergence: $\mathcal{L}_{\text{Geometry}} = D_{\text{KL}}(\mu_\theta \| \sigma_{\mathbb{S}})$, which vanishes exactly when $\mu_\theta = \sigma_{\mathbb{S}}$.

(b) Entropy maximization. With $\sigma_{\mathbb{S}}$ as the reference measure, the KL objective above coincides with the negative relative entropy $-\mathsf{H}_{\sigma_{\mathbb{S}}}(\mu_\theta)$, yielding the exact entropy-maximization objective $\mathcal{L}_{\text{Shannon}} = -\mathsf{H}_{\sigma_{\mathbb{S}}}(\mu_\theta)$. Entropy maximization and information-maximization principles have also been used to promote dispersed and non-degenerate representations in self-supervised learning (Sablayrolles et al., 2019; Liu et al., 2022; Ozsoy et al., 2022). For finite-sample optimization, given a batch $\{z_i\}_{i=1}^N$, a common finite-sample surrogate is the empirical Gaussian information potential, a kernel-based quantity closely connected to quadratic Rényi entropy (Wang & Isola, 2020; Principe, 2010). Under the Gaussian KDE parameterization whose self-convolution kernel is proportional to $\exp(-t\|z_i - z_j\|_2^2)$, the following entropy equivalence holds:

$$\widehat{\mathcal{L}}_{\text{Rényi}} = \log\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(-t\|z_i - z_j\|_2^2\right)\right] = -\mathsf{H}_2(\widehat{p}_t) + C_t, \tag{17}$$

where $t > 0$ is the inverse-bandwidth parameter of the Gaussian kernel, $\widehat{p}_t$ is the corresponding Gaussian kernel density estimate (KDE), and C_t is independent of the representation samples. This construction provides a finite-sample Rényi-entropy surrogate for representation spreading, while the relative-Shannon identity above remains the population-level characterization with respect to $\sigma_{\mathbb{S}}$.

(c) Basiswise equipartition. From a basiswise second-order perspective, redundancy-reduction methods such as VICReg (Bardes et al., 2022) and Barlow Twins (Zbontar et al., 2021) regularize the interactions among representation dimensions. Let Σ denote an appropriate second-order feature matrix, such as a within-view covariance matrix or a cross-view correlation matrix. Their objectives can be abstracted as

$$\mathcal{L}_{\text{Basis}} = \|\text{offdiag}(\Sigma)\|_F^2 + \Phi(\text{diag}(\Sigma)), \tag{18}$$

where the first term suppresses correlations between distinct coordinates, while Φ is a method-dependent diagonal regularizer enforcing non-degenerate marginal scales. More generally, these objectives promote non-degenerate marginal variation together with reduced cross-coordinate dependence, yielding a well-spread second-order spectrum (Garrido et al., 2023; Liu et al., 2022; Zhang et al., 2024). This dimension-contrastive view provides a practical second-order surrogate for representation uniformity without explicitly constraining the embeddings to a hyperspherical manifold.

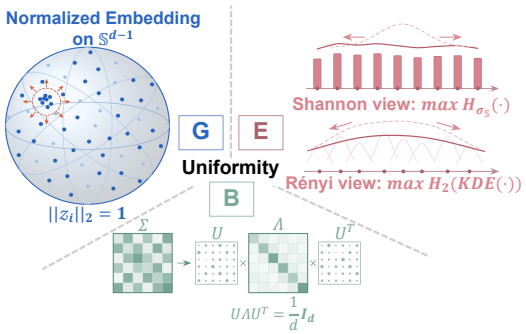


Figure 6: The G.E.B. view of SSL uniformity: **G**eometric reference measure matching, **E**ntropy maximization and **B**asiswise equipartition.

We collectively refer to these three characterizations of uniformity as the G.E.B. view, which is schematically illustrated in Figure 6.

B Proofs and Derivations

B.1 Lagrangian Derivation of the Maximum-Entropy Distribution

Let $H(\mu) = -\int_{\mathcal{E}_{K,d}} p(z)\log p(z)\, dV_K(z)$ denote the differential entropy of a distribution μ with respect to the volume measure dV_K, where $p(z)$ is its probability density. By Definition 1, the density satisfies the normalization constraint $\int p(z)\, dV_K(z) = 1$, while the representation-scale budget

is given by $\int r(z)^2 p(z)\, dV_K(z) \le \rho$. Introducing a Lagrange multiplier $\lambda \ge 0$ for the scale constraint and β for the normalization constraint, we define the following Lagrangian functional Pennec (2006):

$$\mathscr{L}[p] = -\int p \log p \, dV_K - \lambda \left(\int r^2 p \, dV_K - \rho \right) - \beta \left(\int p \, dV_K - 1 \right). \tag{19}$$

Taking the variation with respect to the density $p(z)$ and setting the functional derivative to zero yields the optimality condition:

$$\frac{\delta \mathscr{L}}{\delta p(z)} = -\log p(z) - 1 - \lambda r(z)^2 - \beta = 0. \tag{20}$$

Rearranging and exponentiating gives $p(z) = e^{-1-\beta} e^{-\lambda r(z)^2}$. Enforcing the normalization constraint $\int p(z)\, dV_K(z) = 1$ yields $e^{-1-\beta} = Z_\lambda^{-1}$, where $Z_\lambda = \int_{\mathcal{E}_{K,d}} e^{-\lambda r(\xi)^2}\, dV_K(\xi)$. Therefore, $p(z) = p_{K,\rho}(z)$, which is exactly the Gibbs density in Eq. 5.

B.2 Derivation of Entropy Transfer (Proposition 2)

In geodesic polar coordinates, the volume element in the tangent space is $du = r^{d-1} dr\, d\Omega$, whereas the manifold volume element is given by Eq. 3. The two differ only by an angularly independent radial factor, namely the volume distortion of the exponential map $J_K(r)$:

$$dV_K = J_K(r)\, du. \tag{21}$$

Let $p = d\mu / dV_K$ denote the density of μ on the manifold, and let $\tilde{p}$ denote the density with respect to du of the distribution obtained by pushing μ forward to the tangent space through the logarithmic map $u = \log_o(z)$. The densities of the same distribution with respect to the two volume elements differ only by this distortion factor, i.e., $\tilde{p}(u) = p(z) J_K(r)$. Substituting this relation into the definition of entropy in the tangent space and changing variables back to the manifold gives

$$H_{\tan}(\mu) = -\int \tilde{p} \log \tilde{p}\, du = -\int p \big[\log p + \log J_K(r)\big] dV_K = H(\mu) - \mathbb{E}_\mu\big[\log J_K(r)\big], \tag{22}$$

which, after rearrangement, yields Proposition 2. Thus, the difference between the two entropies is entirely governed by curvature: when $K = 0$, $J_K \equiv 1$ and the two entropies coincide; when $K < 0$, $s_K(r) > r$, so the tangent-space entropy systematically underestimates the intrinsic entropy, while the opposite holds for $K > 0$.

B.3 Proof of the Intrinsic–Wrapped Distributional Gap (Proposition 3)

By the non-negativity of the KL divergence,

$$\Delta(K, \rho, d) = D_{\mathrm{KL}}\left(q_{\tan} \parallel q_{K,\rho}\right) \ge 0. \tag{23}$$

When $K = 0$, we have $J_K(r) \equiv 1$, and the same scale budget gives $\lambda = \lambda' = d/(2\rho)$. It follows from Eq. 7 that the two radial densities coincide exactly, and hence $\Delta(0, \rho, d) = 0$.

When $K \neq 0$, a Taylor expansion of the log-Jacobian factor gives

$$\log J_K(r) = -\frac{d-1}{6} K r^2 - \frac{d-1}{180} K^2 r^4 + O(|K|^3 r^6). \tag{24}$$

Let L_K denote the log-ratio of the two radial densities. By Eq. 7,

$$L_K := \log(q_{K,\rho}/q_{\tan}) = C_K + \log J_K(r) - (\lambda - \lambda') r^2, \tag{25}$$

where C_K is determined by the ratio of the normalization constants of the two radial distributions. Adjusting λ' can only modify the coefficient of the quadratic term in r, but cannot eliminate the quartic term. When $d \ge 2$ and $K \neq 0$, the coefficient of the r^4 term is nonzero, implying $q_{\tan} \neq q_{K,\rho}$, and therefore $\Delta(K, \rho, d) > 0$.

For $K > 0$, the tangent reference is restricted to the injectivity ball $r \in [0, \pi/\sqrt{K}]$ and renormalized. For $K \le 0$, no truncation is required; in this case, $\lambda' = d/(2\rho)$, and r^2 under $q_{\tan}$ follows a Gamma

distribution with shape $d/2$ and scale $2\rho/d$. Hereafter, all expectations, variances, and covariances are taken under $q_{\tan}$ unless otherwise specified. Therefore,

$$\mathbb{E}[r^2] = \rho, \qquad \mathrm{Var}(r^2) = \frac{2\rho^2}{d}, \qquad \mathrm{Cov}(r^2, r^4) = \frac{4(d+2)}{d^2}\rho^3. \tag{26}$$

Let

$$\lambda - \lambda' = c_1 K + c_2 \rho K^2 + O(|K|^3 \rho^2). \tag{27}$$

Substituting Eqs. 24 and 27 into Eq. 25 gives

$$L_K = C_K - \left(c_1 + \frac{d-1}{6}\right) K r^2 - \left(c_2 \rho r^2 + \frac{d-1}{180} r^4\right) K^2 + O\big(|K|^3 (r^6 + \rho^2 r^2)\big). \tag{28}$$

For a normalized exponential perturbation, the first-order change in the mean is determined by the corresponding covariance. To keep $\mathbb{E}[r^2] = \rho$ fixed, the first-order moment condition gives $c_1 = -\frac{d-1}{6}$ and $c_2 \rho \,\mathrm{Var}(r^2) + \frac{d-1}{180} \,\mathrm{Cov}(r^2, r^4) = 0$. Substituting these into Eq. 27 yields

$$\lambda - \lambda' = -\frac{d-1}{6} K - \frac{(d-1)(d+2)}{90d} \rho K^2 + O(|K|^3 \rho^2). \tag{29}$$

Substituting this back into Eq. 25 and absorbing the normalization term gives

$$L_K = K^2 h(r^2) + O\big(|K|^3 (r^6 + \rho^2 r^2 + \rho^3)\big), \tag{30}$$

where

$$h(r^2) = -\frac{d-1}{180} \left(r^4 - \frac{2(d+2)}{d} \rho r^2 + \frac{d+2}{d} \rho^2\right). \tag{31}$$

Using the moment formulas of the Gamma distribution,

$$\mathbb{E}[h(r^2)] = 0, \qquad \mathrm{Var}(h(r^2)) = \frac{8(d-1)^2(d+2)}{180^2 d^3} \rho^4. \tag{32}$$

Since $q_{K,\rho}$ is normalized, $\mathbb{E}[e^{L_K}] = 1$; using the Gamma moments together with tail-decay bounds to control the remainder, a second-order expansion of this identity gives

$$0 = \mathbb{E}[L_K] + \frac{1}{2} \mathbb{E}[L_K^2] + O(\tau^6). \tag{33}$$

On the other hand, squaring Eq. 30 and taking expectations, the cross terms and the fifth- and sixth-order contributions from the squared remainder are absorbed into $o(\tau^4)$, yielding

$$\mathbb{E}[L_K^2] = K^4 \mathbb{E}[h(r^2)^2] + o(\tau^4) = K^4 \,\mathrm{Var}(h(r^2)) + o(\tau^4). \tag{34}$$

Combining Eqs. 25, 32, 33, and 34 gives

$$\Delta(K, \rho, d) = -\mathbb{E}[L_K] = \frac{(d-1)^2(d+2)}{8100 d^3} K^4 \rho^4 + o(\tau^4). \tag{35}$$

Since $(K\rho)^4 = \tau^4$, setting

$$C_d = \frac{(d-1)^2(d+2)}{8100 d^3} > 0, \tag{36}$$

yields Eq. 11.

B.4 Radial–Angular KL Decomposition (Proposition 4)

By Eq. 3, expressing the density p of μ with respect to dV_K in geodesic polar coordinates gives its joint density with respect to $dr\, d\Omega$:

$$\tilde{p}(r, \omega) = p\big(\exp_o(r\omega)\big) s_K(r)^{d-1}. \tag{37}$$

Let

$$\omega_{d-1} = |\mathbb{S}^{d-1}|, \qquad \mathcal{R}_K := \begin{cases} [0, \pi/\sqrt{K}], & K > 0, \\ [0, \infty), & K \le 0. \end{cases} \tag{38}$$

Then $\mathrm{Unif}(\mathbb{S}_o^{d-1})$ has density $u_o(\omega) = \omega_{d-1}^{-1}$ with respect to $d\Omega$. Let $\mu_r(r)$ and $\mu_{\omega|r}(\omega \mid r)$ denote the radial marginal density and the conditional directional density given r, respectively. By marginalization and conditional factorization,

$$\mu_r(r) = \int_{\mathbb{S}_o^{d-1}} \tilde{p}(r,\omega) d\Omega(\omega), \qquad \mu_{\omega|r}(\omega \mid r) = \frac{\tilde{p}(r,\omega)}{\mu_r(r)}. \tag{39}$$

The conditional directional density only needs to be defined for radii satisfying $\mu_r(r) > 0$; its values on the zero-density set do not affect the subsequent integrals. On the other hand, by Eq. 5, the density of $\nu_{K,\rho}$ in geodesic polar coordinates factorizes as

$$\begin{aligned} \tilde{\nu}_{K,\rho}(r,\omega) &= Z_\lambda^{-1} s_K(r)^{d-1} e^{-\lambda r^2} \\ &= \underbrace{\frac{\omega_{d-1}}{Z_\lambda} s_K(r)^{d-1} e^{-\lambda r^2}}_{q_{K,\rho}(r)} \underbrace{\frac{1}{\omega_{d-1}}}_{u_o(\omega)}. \end{aligned} \tag{40}$$

Here, $q_{K,\rho}$ is the radial marginal density of $\nu_{K,\rho}$. By the definition of Z_λ, $\int_{\mathcal{R}_K} q_{K,\rho}(r) dr = 1$. Therefore, in geodesic polar coordinates, $\nu_{K,\rho}$ consists of the radial distribution $q_{K,\rho}$ and the uniform conditional directional distribution $\mathrm{Unif}(\mathbb{S}_o^{d-1})$. Computing the KL divergence with respect to the common base measure $dr\, d\Omega$ then gives

$$\begin{aligned} D_{\mathrm{KL}}(\mu \parallel \nu_{K,\rho}) &= \int_{\mathcal{R}_K} \int_{\mathbb{S}_o^{d-1}} \mu_r(r)\mu_{\omega|r}(\omega \mid r) \log \frac{\mu_r(r)\mu_{\omega|r}(\omega \mid r)}{q_{K,\rho}(r) u_o(\omega)} \, d\Omega(\omega)\, dr \\ &= \int_{\mathcal{R}_K} \int_{\mathbb{S}_o^{d-1}} \mu_r(r)\mu_{\omega|r}(\omega \mid r) \left[\log \frac{\mu_r(r)}{q_{K,\rho}(r)} + \log \frac{\mu_{\omega|r}(\omega \mid r)}{u_o(\omega)} \right] d\Omega(\omega)\, dr \\ &= \int_{\mathcal{R}_K} \mu_r(r) \log \frac{\mu_r(r)}{q_{K,\rho}(r)} \, dr + \int_{\mathcal{R}_K} \mu_r(r) \int_{\mathbb{S}_o^{d-1}} \mu_{\omega|r}(\omega \mid r) \log \frac{\mu_{\omega|r}(\omega \mid r)}{u_o(\omega)} \, d\Omega(\omega)\, dr \\ &= D_{\mathrm{KL}}(\mu_r \parallel q_{K,\rho}) + \mathbb{E}_{r\sim\mu_r}\left[D_{\mathrm{KL}}\big(\mu_{\omega|r} \parallel \mathrm{Unif}(\mathbb{S}_o^{d-1})\big) \right]. \end{aligned} \tag{41}$$

C Experimental Details and Evaluation Protocol

C.1 Numerical construction of the reference quantiles

For each (K, d, ρ) configuration, we construct $q_{K,\rho}(r)$ before training. When the scale constraint is active, we numerically evaluate the second moment and solve $\mathbb{E}_{q_{K,\rho}}[r^2] = \rho$ to determine λ; otherwise, we set $\lambda = 0$. For $K > 0$, the integration domain is $(0, \pi/\sqrt{K})$, whereas for $K < 0$ we truncate the integral at a sufficiently large finite radius. We then precompute the radial target quantiles required during training by numerical integration and CDF inversion. When $K = 0$, the corresponding quantiles are computed directly from $r^2 \sim (\rho/d)\chi_d^2$. The angular reference is obtained directly from a Beta distribution. Let $a = (d-1)/2$. Then $Q_{\mathrm{ang},d}(u) = 2F^{-1}_{\mathrm{Beta}(a,a)}(u) - 1$. For each radial bin, we precompute the midpoint quantiles corresponding to the number of samples in that bin and use them for minibatch angular matching in Eq. 14.

C.2 Additional Details for the Node-Classification Benchmark

Datasets and preprocessing. We conduct node-classification experiments on four public benchmark datasets: Airport (Chami et al., 2019), PubMed, CiteSeer, and Cora (Yang et al., 2016). Airport is an airline network in which nodes represent airports and edges represent flight routes between airports. Cora, CiteSeer, and PubMed are citation networks in which nodes represent papers, edges represent citation relationships, and node labels correspond to different classes or research topics. We adopt the standard transductive node-classification setting. Cora, CiteSeer, and PubMed use the public fixed Planetoid splits, which are kept unchanged across different random-seed runs.

Baselines. We compare against Euclidean graph models including GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2018), and GraphSAGE (Hamilton et al., 2017); hyperbolic models including

HNN (Ganea et al., 2018), HGNN (Liu et al., 2019), HGCN (Chami et al., 2019), HGAT (Zhang et al., 2022a), and HIE (Yang et al., 2023); and self-supervised including GRACE (Zhu et al., 2020), COSTA (Zhang et al., 2022b), HGCL (Liu et al., 2021), and HyperGCL (Zhang et al., 2025).

Implementation settings and repeated runs. All experiments are repeated independently over the same five random seeds. We train for 500 epochs with representation dimension $d = 128$, learning rate 10^{-3}, and weight decay 10^{-5}. The number of slicing directions $M \in \{64, 128\}$ for angular sliced-W_2 matching and the number of equal-frequency radial bins $B \in \{2, 4, 8\}$ are selected for each dataset by grid search. All experiments were conducted on an NVIDIA RTX PRO 6000D GPU.

Pretraining and model architecture. For an input graph with N nodes, let $\mathbf{A} \in \mathbb{R}^{N\times N}$ and $\mathbf{X} \in \mathbb{R}^{N\times F}$ denote its adjacency matrix and node-feature matrix, respectively. At each training step, we generate two augmented views $(\mathbf{A}^{(v)}, \mathbf{X}^{(v)})$, $v \in \{1, 2\}$, from the original graph and encode them using a shared two-layer GCN encoder (Kipf & Welling, 2017) to obtain $\mathbf{U}^{(v)} = f_\theta(\mathbf{X}^{(v)}, \mathbf{A}^{(v)}) \in \mathbb{R}^{N\times d}$. The representations are then mapped to hyperbolic space through the exponential map, yielding $\mathbf{Z}^{(v)} = \exp_o(\mathbf{U}^{(v)})$. Graph-view alignment and both radial and angular distribution matching are computed on the representations $\mathbf{Z}^{(v)}$. After pretraining, the encoder is frozen and a linear node classifier is trained on $\mathbf{U}$. We use the Adam optimizer (Kingma & Ba, 2015) with gradient clipping, and select the remaining hyperparameters using the validation set.

C.3 Additional Details for the ImageNet-100 Experiment

Dataset and model. We use the ImageNet-100 class subset constructed by Ming et al. (2022) and the corresponding class hierarchy adopted by Sinha et al. (2024). The original ImageNet-100 training images are deterministically partitioned into 118,241 images for label-free adaptation on ImageNet-100 and a reference subset of 10,000 images, with 100 images per class. The original validation split of 5,000 images, containing 50 images per class, is kept unchanged and used for downstream evaluation. We initialize from a ResNet-34 (He et al., 2016) pretrained on ImageNet-1K (Deng et al., 2009) under supervision, freeze the stem and layer1–layer3, and update only layer4 and a bias-free linear projection head from 512 to 16 dimensions.

Training and evaluation protocol. Each image produces two independently augmented views using standard ImageNet preprocessing. We train for 20 epochs with SGD, using momentum 0.9, weight decay 10^{-4}, an initial learning rate of 0.01 with cosine decay, batch size 512, gradient clipping at 5, 64 slicing directions, and four radial bins. The alignment and MaxEnt terms are assigned equal weights. For each (K, ρ) setting, the tangent transplant and intrinsic MaxEnt are trained using identical data, augmentations, optimizer settings and paired random seeds; the only component that differs is the radial reference distribution. All five settings and all seeds are evaluated directly, without checkpoint or hyperparameter selection based on validation performance.

For each frozen checkpoint, we first compute the radial diagnostics on the 10,000-image reference subset without accessing its class labels. The ratio $Q_{0.95}/Q_{0.50}$ is computed using linearly interpolated quantiles of the intrinsic radius $r(z)$. Hierarchy diagnostics are computed separately on the unchanged validation split. For each class, we compute the Euclidean mean in projector coordinates, map it to the manifold, and compare the resulting native pairwise distances between class prototypes with the WordNet shortest-path distances. Following the classical multidimensional-scaling stress criterion (Kruskal, 1964), the reported scale-aligned normalized stress is defined as

$$\text{Stress} = \frac{\sqrt{P^{-1}\sum_{i<j}(\gamma d^{\mathbf{z}}_{ij} - t_{ij})^2}}{\frac{1}{P}\sum_{i<j} t_{ij}}, \qquad \gamma = \frac{\sum_{i<j} d^{\mathbf{z}}_{ij} t_{ij}}{\sum_{i<j}(d^{\mathbf{z}}_{ij})^2}, \tag{42}$$

where P is the number of class pairs. Finally, we standardize the frozen backbone features using statistics computed from the adaptation split and train a linear classifier with SGD for 60 epochs, using batch size 4096, learning rate 0.5, momentum 0.9, and no weight decay. Top-1 accuracy is reported on the fixed validation split.

D From-Scratch Self-Supervised Pretraining on STL-10

To rule out potential confounding effects from pretrained initialization, we further conduct from-scratch self-supervised pretraining on STL-10 (Coates et al., 2011) using a randomly initialized

ResNet-18 (He et al., 2016). STL-10 contains 96×96 RGB images from 10 classes. We keep the standard dataset splits fixed and select 10,000 images from the unlabeled data using a fixed random seed as a reference subset for distributional diagnostics. Representations are constructed in spaces with $K \in -1, 0, +1$, the scale budget is fixed to $\rho = 1$, and all models are pretrained for 120 epochs. Unless otherwise specified below, the data augmentation and objective construction follow the image experiment described in Appendix C.3. After pretraining, we extract the learned representations for the analyses below.

We vary the weight coefficient α in Eq. 16 to construct different alignment–MaxEnt trade-offs, including joint, alignment-heavy, and MaxEnt-heavy configurations, together with alignment-only and MaxEnt-only as two boundary cases. For each configuration, we record its optimization trajectory in the $\mathcal{L}_{\text{align}}$–$(\mathcal{L}_{\text{rad}} + \mathcal{L}_{\text{ang}})$ plane and retain checkpoints at epochs 1, 40, 80, and 120. After pretraining, the encoder is frozen and a logistic-regression linear probe is fitted on the clean representations, with accuracy reported on the test split. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 10^{-4}. The initial learning rate is selected from $\{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$ as part of the predefined trajectory configurations, followed by a 5-epoch linear warmup and cosine decay.

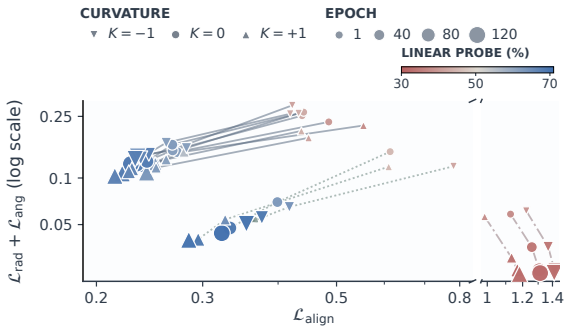


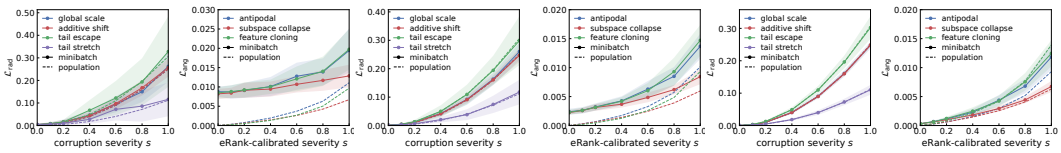
Figure 7: SSL trajectories on STL-10. Axes show alignment loss and intrinsic MaxEnt discrepancy; marker shape, size, and color denote curvature, epoch, and linear-probe accuracy, respectively.

As shown in Figure 7, across all three curvatures, higher linear-probe accuracy generally occurs in regions where both alignment and MaxEnt discrepancy are controlled. Among the 15 trajectories, 12 simultaneously reduce both losses and improve downstream accuracy from the early to the final checkpoint. In contrast, the MaxEnt-only configurations continue to reduce the distributional discrepancy while exhibiting worse view alignment and lower performance. These results indicate that intrinsic MaxEnt and alignment play complementary roles across different curvatures, and that optimizing distribution matching alone is insufficient to obtain discriminative representations.

E Sensitivity of Radial and Angular Losses to Failure Modes

To test the sensitivity of the radial and angular losses, we sample from the intrinsic MaxEnt reference $\nu_{K,\rho}$ at $(K, \rho, d) = (-1, 1, 16)$ without training an encoder. We apply four radial perturbations (global scale, additive shift, tail escape, and tail stretch) and three angular perturbations (antipodal concentration, subspace collapse, and feature cloning). For each angular perturbation, severity $s \in [0, 1]$ is calibrated to the target $\text{erank}(C_\omega) = 16 - 12s$, where $C_\omega = \mathbb{E}[\omega\omega^\top]$ and erank is the effective rank (Roy & Vetterli, 2007). For $N \in \{16, 64, 512\}$, we use $M = 64$ random slices and repeat each setting 50 times.

Figure 8 shows that $\mathcal{L}_{\text{rad}}$ increases with all four radial perturbations, most strongly for tail escape. The three angular perturbations yield different $\mathcal{L}_{\text{ang}}$ values despite their matched target effective rank, so effective rank alone does not determine angular distributional mismatch. As N grows, the minibatch estimates generally move toward the population-reference estimates, while their zero-perturbation bias and variability decrease.



(a) $N = 16$ (b) $N = 64$ (c) $N = 512$

Figure 8: Sensitivity to controlled radial and angular collapse across minibatch sizes.