Hyperbolic Belief Propagation

Zehua Cheng[1] **Wei Dai**[2] **Jiahao Sun**[2]

[1]Department of Computer Science, University of Oxford, Oxford, United Kingdom
[2]FLock.io, London, United Kingdom

Abstract

Belief Propagation (BP) has operated in Euclidean space for four decades, yet the polynomial volume growth of $\mathbb{R}^d$ is fundamentally mismatched to hierarchical graphs: faithful embedding of exponentially branching structure demands $d = \Omega(\log n)$ dimensions and prohibitive $\mathcal{O}(d^3)$ covariance costs, while truncating d corrupts marginal estimates. We introduce Continuous Hyperbolic Belief Propagation (CHBP), formulated natively on the Lorentz hyperboloid $\mathbb{H}^n_K$, whose exponential volume growth eliminates this bottleneck. CHBP parameterises beliefs as Jacobian-corrected Wrapped Normals, approximates message integrals via Gauss–Hermite quadrature, and transports covariance tensors between tangent spaces via closed-form Levi-Civita parallel transport, with a curvature annealing schedule ensuring stable convergence. On hierarchical graphs, CHBP at d=5 reduces marginal KL divergence by $25\times$ versus Euclidean BP at d=50 and achieves up to 91.45% accuracy on real-world taxonomies, outperforming all baselines including Hyperbolic GCNs. On flat topologies, CHBP predictably underperforms Euclidean methods, confirming a topology-specific rather than universal advantage.

1 INTRODUCTION

Probabilistic inference on graphical models is a geometric problem. Belief Propagation (BP) [Pearl, 1988, Kschischang et al., 2001] manipulates distributions whose form is constrained by the ambient geometry—for four decades, Euclidean. This is harmless on lattices, but hierarchical structures (taxonomies, phylogenies, ontologies) have branching topology incompatible with the polynomial volume growth of $\mathbb{R}^d$. A k-ary tree of depth h has $\Theta(k^h)$ leaves, yet a ball in $\mathbb{R}^d$ grows as r^d; bounded-distortion embedding requires $d = \Omega(\log n)$ [Bourgain, 1985], forcing continuous BP into a dilemma: pay $\mathcal{O}(d^3)$ covariance costs per message, or truncate d and corrupt marginals through branch aliasing.

Variational alternatives [Minka, 2001, Wainwright et al., 2003] seek tighter approximations within the same flat geometry—optimising the solver while ignoring the domain. Hyperbolic space $\mathbb{H}^n_K$ ($K < 0$) provides exponential volume growth $\sim e^{(n-1)\sqrt{|K|}\,r}$ that matches hierarchical branching [Sarkar, 2011, Krioukov et al., 2010]; a k-ary tree embeds into $\mathbb{H}^2$ with arbitrarily low distortion. Despite extensive work on hyperbolic embeddings [Nickel and Kiela, 2017, 2018, Ganea et al., 2018], no prior work has addressed the follow-up: if the embedding belongs in hyperbolic space, should the inference algorithm not operate there as well?

We introduce Continuous Hyperbolic Belief Propagation (CHBP), in which beliefs, messages, aggregation, and transport operate natively on the Lorentz hyperboloid. Beliefs are Wrapped Normal distributions $\mathcal{WN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [Nagano et al., 2019] with Jacobian-corrected densities; message integrals are approximated via Gauss–Hermite quadrature in tangent space; covariance tensors are transported between tangent spaces via closed-form Levi-Civita parallel transport, preserving the eigenspectrum; and a curvature annealing schedule deforms the manifold from near-flat to target curvature, stabilising convergence. The design carries formal guarantees (Appendix B): quadrature error bounded by $\mathcal{O}(\|\boldsymbol{\Sigma}\|^Q)$ (Theorem 12), continuous homotopy with fixed-point continuity (Theorem 13), and spectral preservation under transport (Theorem 9).

On hierarchical graphs, CHBP at d=5 reduces marginal KL by $25\times$ versus Euclidean BP at d=50 (2.15 vs 54.25×10^{-2} on 10-ary trees), and achieves up to 91.45% accuracy on real-world taxonomies with the lowest NLL among all baselines, including Hyperbolic GCNs. On flat topologies (grids, Spin Glass), CHBP predictably underperforms TRW-BP—confirming topology-specific rather than universal advantage.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1230–1248.

We summarise our contributions as follows:

1. The first complete formulation of belief propagation on hyperbolic space, integrating Wrapped Normal beliefs, geodesic-kernel potentials, Levi-Civita covariance transport, and curvature annealing into a single algorithm.
2. Formal guarantees including a quadrature error bound ($\mathcal{O}(\|\mathbf{\Sigma}\|^Q)$) and a homotopy continuity result for the annealing schedule.
3. Extensive evaluation across 21 synthetic and 7 real-world graphs, demonstrating that geometric–topological alignment is the dominant factor in continuous inference quality, with transparent failure modes on non-hierarchical structures.

2 RELATED WORK

Belief Propagation and Approximate Inference. Belief Propagation (BP), introduced by Pearl [1988] for tree-structured graphical models and later generalised to factor graphs by Kschischang et al. [2001], remains a foundational algorithm for probabilistic inference. When applied to graphs with cycles, Loopy BP [Yedidia et al., 2003] lacks convergence guarantees but often produces accurate marginals in practice. A rich family of variational alternatives has emerged to address these limitations. Expectation Propagation (EP) [Minka, 2001] approximates intractable factors by iterative moment matching within the exponential family, while Tree-Reweighted BP (TRW-BP) [Wainwright et al., 2003] constructs convex upper bounds on the partition function by reweighting edge contributions. Wainwright and Jordan [2008] provide a comprehensive treatment of the variational perspective, unifying mean-field, belief propagation, and convex relaxation methods under a common framework. More recently, neural extensions of message passing have been explored [Yoon et al., 2018, Satorras and Welling, 2021], parameterising messages with learned functions. However, all of these methods operate exclusively in Euclidean space, inheriting the polynomial volume growth of $\mathbb{R}^d$ and consequently requiring high-dimensional embeddings to faithfully represent hierarchical dependency structures.

Hyperbolic Geometry in Machine Learning. The observation that hyperbolic space provides exponential volume growth—naturally matching the branching structure of hierarchical data—has motivated a growing body of work on hyperbolic embeddings. Nickel and Kiela [2017] pioneered the use of the Poincaré ball model for learning hierarchical representations of symbolic data, achieving superior link prediction and taxonomy reconstruction compared to Euclidean baselines. Nickel and Kiela [2018] subsequently demonstrated that the Lorentz (hyperboloid) model offers substantially improved numerical stability over the Poincaré disk for gradient-based optimisation. The theoretical underpinnings of these approaches rest on the foundational result of Sarkar [2011], who showed that trees can be embedded into the hyperbolic plane with arbitrarily low distortion, and the network-geometric framework of Krioukov et al. [2010], which established the natural relationship between hyperbolic geometry and scale-free network topology. Beyond embeddings, hyperbolic spaces have been adopted for clustering [Monath et al., 2019], natural language processing [Tifrea et al., 2018], and computer vision [Khrulkov et al., 2020]. Despite this breadth, the application of hyperbolic geometry to *probabilistic inference algorithms* such as belief propagation has remained largely unexplored.

Probability Distributions on Riemannian Manifolds. Defining valid probability distributions on curved spaces requires careful treatment of the Riemannian volume form. Pennec [2006] established the intrinsic statistical framework for Riemannian manifolds, introducing Fréchet means and providing the measure-theoretic foundations for density estimation on curved domains. Building on this, Nagano et al. [2019] proposed the Wrapped Normal distribution on hyperbolic space, constructed by pushing forward a tangent-space Gaussian through the exponential map with an explicit Jacobian correction. Mathieu et al. [2019] extended this approach to variational autoencoders, demonstrating that latent-space models benefit from the geometric inductive bias of constant negative curvature. Lou et al. [2020] contributed differentiable Fréchet mean computation, enabling end-to-end gradient-based learning of distributional parameters on manifolds. Complementary work by Skopek et al. [2020] explored mixed-curvature product spaces for richer latent geometries. Our work is the first to integrate the Wrapped Normal distribution into a full message-passing inference algorithm, requiring novel solutions for parallel transport of covariance tensors and numerical quadrature of transition potentials on the manifold.

3 METHODOLOGY

3.1 PROBLEM SETTING AND GEOMETRIC MOTIVATION

We assume familiarity with the basic vocabulary of Riemannian geometry (tangent spaces, exponential and logarithmic maps, parallel transport); readers seeking an introduction are referred to Appendix A.

Consider a pairwise Markov random field over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $i \in \mathcal{V}$ holds a continuous latent variable $\boldsymbol{x}_i \in \mathcal{M}$ and each edge $(i, j) \in \mathcal{E}$ encodes a pairwise dependency. Standard continuous BP [Pearl, 1988, Kschischang et al., 2001] operates in $\mathbb{R}^d$: each node maintains a belief $b_i(\boldsymbol{x}_i)$, and messages $m_{i\to j}(\boldsymbol{x}_j)$ are computed by marginalising over the sender's variable. Because a ball

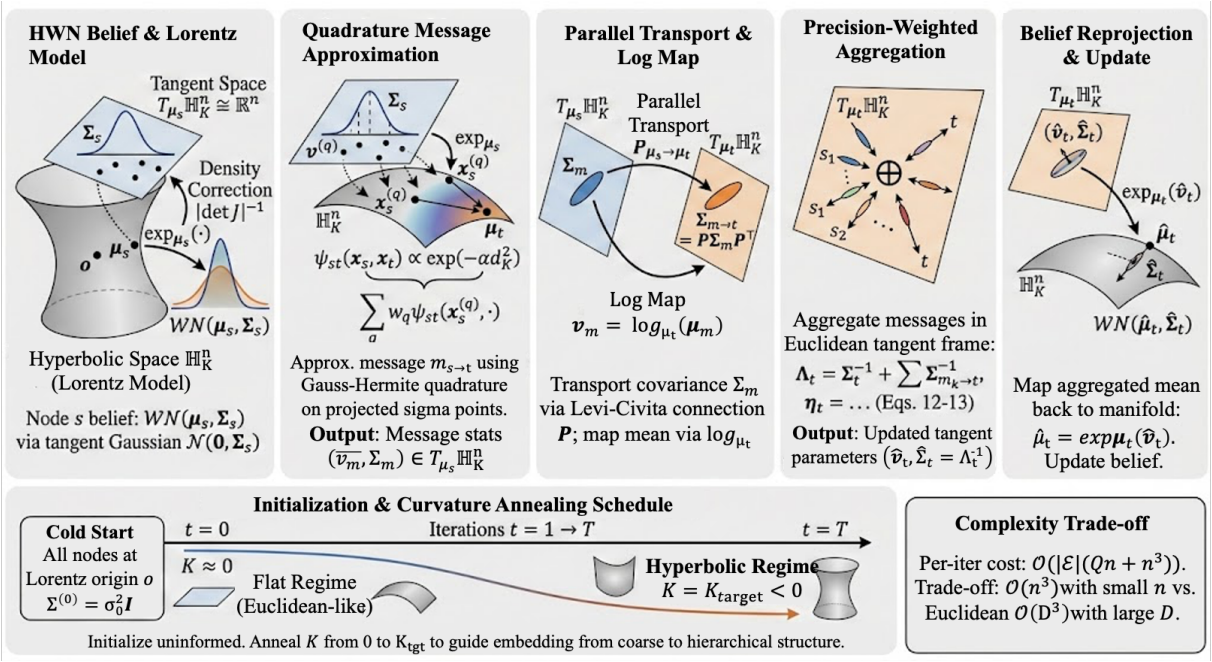

Figure 1: Overview of the Continuous Hyperbolic Belief Propagation (CHBP) framework. (1) Node beliefs are parameterized using the Hyperbolic Wrapped Normal (HWN) distribution on the Lorentz model, mapping a tangent-space Gaussian to the manifold via a measure-corrected exponential map. (2) Outgoing sum-product messages are tractably approximated using Gauss-Hermite quadrature on projected sigma points. (3) To account for holonomy, message statistics are moved to the receiver's local frame via Levi-Civita parallel transport for the covariance and the logarithmic map for the mean. (4) Incoming messages are conditionally aggregated in the exact Euclidean tangent space of the receiving node using precision-weighted updates. (5) The newly aggregated mean is projected back onto the manifold to form the updated belief. Bottom: The algorithm employs a cold-start initialization at the Lorentz origin and utilizes a curvature annealing schedule—progressing from a flat Euclidean-like regime ($K \approx 0$) to the target hyperbolic geometry—to ensure stable optimization and prevent early tangent-space misalignment.

in $\mathbb{R}^d$ grows as r^d, faithfully embedding an exponentially branching topology requires $d = \Omega(\log n)$ dimensions [Bourgain, 1985], rendering the $\mathcal{O}(d^3)$ per-message covariance operations prohibitive.

Hyperbolic space $\mathbb{H}^n_K$ ($K < 0$) resolves this: its volume grows as $\sim e^{(n-1)\sqrt{|K|}\,r}$, enabling low-distortion hierarchical embeddings at $d \leq 10$ [Sarkar, 2011, Krioukov et al., 2010]. However, formulating BP on a Riemannian manifold requires defining valid densities under a non-Euclidean volume form, computing message integrals over a curved domain, and transporting statistics between disjoint tangent spaces. We address each challenge below.

3.2 LORENTZ MODEL AS THE COMPUTATIONAL DOMAIN

The Poincaré ball model $\mathbb{B}^n$, popular in geometric deep learning [Nickel and Kiela, 2017, Ganea et al., 2018], carries a conformal factor $\lambda_{\boldsymbol{x}} = 2/(1 - \|\boldsymbol{x}\|^2)$ that diverges as $\|\boldsymbol{x}\| \to 1$, causing numerical overflow in iterative algorithms. We therefore formulate all operations in the Lorentz (hyperboloid) model [Nickel and Kiela, 2018]:

$$\mathbb{H}^n_K = \{\boldsymbol{x} \in \mathbb{R}^{n+1} : \langle \boldsymbol{x}, \boldsymbol{x} \rangle_L = 1/K,\ x_0 > 0\}, \qquad (1)$$

where $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_L = -x_0 y_0 + \sum_{k=1}^n x_k y_k$ is the Minkowski inner product. The tangent space $T_{\boldsymbol{\mu}}\mathbb{H}^n_K = \{\boldsymbol{v} \in \mathbb{R}^{n+1} : \langle \boldsymbol{v}, \boldsymbol{\mu} \rangle_L = 0\}$ inherits a positive-definite (exactly Euclidean) metric from the Minkowski restriction. The geodesic distance is

$$d_K(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\sqrt{|K|}} \operatorname{arccosh}\big(K \langle \boldsymbol{x}, \boldsymbol{y} \rangle_L\big). \qquad (2)$$

Because all operations use $\cosh$ and $\sinh$ of bounded arguments, the exponential and logarithmic maps remain well-conditioned without norm clipping.

3.3 BELIEF PARAMETERISATION VIA THE WRAPPED NORMAL DISTRIBUTION

Standard Gaussians cannot serve as densities on $\mathbb{H}^n_K$, which lacks a global additive structure. We represent node uncertainty via the Hyperbolic Wrapped Normal (HWN)

$\mathcal{WN}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ [Nagano et al., 2019, Mathieu et al., 2019]: a zero-mean Gaussian $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$ is sampled in $T_{\boldsymbol{\mu}_i}\mathbb{H}^n_K$ and projected onto the manifold via the exponential map:

$$\exp_{\boldsymbol{\mu}}(\boldsymbol{v}) = \cosh\big(\sqrt{|K|}\,\|\boldsymbol{v}\|_L\big)\,\boldsymbol{\mu} + \frac{\sinh\big(\sqrt{|K|}\,\|\boldsymbol{v}\|_L\big)}{\sqrt{|K|}\,\|\boldsymbol{v}\|_L}\,\boldsymbol{v}, \tag{3}$$

where $\|\boldsymbol{v}\|_L = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v}\rangle_L}$ denotes the Lorentzian norm of the tangent vector.

The pushforward density on $\mathbb{H}^n_K$ requires a Jacobian correction to account for the curvature-induced volume distortion [Pennec, 2006]:

$$p_{\mathbb{H}}(\boldsymbol{x}) = p_T\big(\log_{\boldsymbol{\mu}}(\boldsymbol{x})\big) \cdot \big|\det J_{\exp_{\boldsymbol{\mu}}}\big(\log_{\boldsymbol{\mu}}(\boldsymbol{x})\big)\big|^{-1}. \tag{4}$$

For constant curvature K, the Jacobian determinant admits a closed form [do Carmo, 1992]:

$$\big|\det J_{\exp_{\boldsymbol{\mu}}}(\boldsymbol{v})\big| = \left(\frac{\sinh(\sqrt{|K|}\,\|\boldsymbol{v}\|_L)}{\sqrt{|K|}\,\|\boldsymbol{v}\|_L}\right)^{n-1}. \tag{5}$$

This correction ensures $\int_{\mathbb{H}^n_K} p_{\mathbb{H}}(\boldsymbol{x})\,\mathrm{dvol}_{\mathbb{H}} = 1$, making each belief a valid probability measure.

3.4 SUM-PRODUCT MESSAGE COMPUTATION

The message from source s to target t along $(s,t) \in \mathcal{E}$ requires a pairwise transition potential, which we define as a geodesic kernel:

$$\psi_{st}(\boldsymbol{x}_s, \boldsymbol{x}_t) = \exp\big(-\alpha\, d^2_K(\boldsymbol{x}_s, \boldsymbol{x}_t)\big), \quad \alpha > 0. \tag{6}$$

The outgoing message is the marginalisation integral

$$m_{s\to t}(\boldsymbol{x}_t) = \int_{\mathbb{H}^n_K} \psi_{st}(\boldsymbol{x}_s, \boldsymbol{x}_t)\, b_s(\boldsymbol{x}_s) \prod_{u\in\mathcal{N}(s)\setminus t} m_{u\to s}(\boldsymbol{x}_s)\,\mathrm{dvol}_{\mathbb{H}}, \tag{7}$$

which is analytically intractable due to the $\operatorname{arccosh}^2$ nonlinearity in d^2_K.

To obtain a tractable approximation, we project the integration domain into $T_{\boldsymbol{\mu}_s}\mathbb{H}^n_K$ and apply Gauss–Hermite quadrature. Concretely, we factorise the sender's tangent-space covariance as $\boldsymbol{\Sigma}_s = \boldsymbol{L}_s\boldsymbol{L}_s^\top$ and construct deterministic sigma points $\{\boldsymbol{v}^{(q)}\}_{q=1}^Q$ from the tensor product of one-dimensional Gauss–Hermite nodes, scaled by $\boldsymbol{L}_s$. Each sigma point is mapped to the manifold via $\boldsymbol{x}_s^{(q)} = \exp_{\boldsymbol{\mu}_s}(\boldsymbol{v}^{(q)})$. The message statistics are then approximated as

$$\mathbb{E}[\boldsymbol{x}_t] \approx \sum_{q=1}^{Q} w_q\, \psi_{st}\big(\boldsymbol{x}_s^{(q)}, \boldsymbol{x}_t\big), \tag{8}$$

$$\mathrm{Cov}[\boldsymbol{x}_t] \approx \sum_{q=1}^{Q} w_q \big(\boldsymbol{v}^{(q)} - \bar{\boldsymbol{v}}\big)\big(\boldsymbol{v}^{(q)} - \bar{\boldsymbol{v}}\big)^\top, \tag{9}$$

where $\{w_q\}$ are Gauss–Hermite weights and $\bar{\boldsymbol{v}}$ is the weighted mean. For concentrated beliefs (small $\|\boldsymbol{\Sigma}_s\|$), the local linearity of $\exp_{\boldsymbol{\mu}}$ ensures rapid quadrature convergence; for diffuse beliefs the error grows and higher-order rules may be required (see Theorem 12).

3.5 PARALLEL TRANSPORT OF MESSAGE STATISTICS

The message statistics reside in $T_{\boldsymbol{\mu}_s}\mathbb{H}^n_K$, but the receiver aggregates in $T_{\boldsymbol{\mu}_t}\mathbb{H}^n_K$. Unlike Euclidean space, these tangent spaces are geometrically distinct; naïvely transferring the covariance via the log map alone ignores the holonomy and corrupts directional uncertainty.

We decompose the transfer into two operations. The message mean $\boldsymbol{\mu}_m$ is mapped via $\boldsymbol{v}_m = \log_{\boldsymbol{\mu}_t}(\boldsymbol{\mu}_m)$. The covariance is transported as a $(0,2)$-tensor along the geodesic from $\boldsymbol{\mu}_s$ to $\boldsymbol{\mu}_t$ using the Levi-Civita parallel transport [do Carmo, 1992], which admits a closed form on $\mathbb{H}^n_K$:

$$P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}(\boldsymbol{v}) = \boldsymbol{v} - \frac{\langle \boldsymbol{\mu}_t + \boldsymbol{\mu}_s, \boldsymbol{v}\rangle_L}{1 + \langle \boldsymbol{\mu}_s, \boldsymbol{\mu}_t\rangle_L}\,(\boldsymbol{\mu}_t + \boldsymbol{\mu}_s). \tag{10}$$

The covariance transforms as $\boldsymbol{\Sigma}_{m\to t} = P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}\,\boldsymbol{\Sigma}_m\,P^\top_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}$. Because parallel transport on a constant-curvature space is an isometry, this preserves eigenvalues, trace, and determinant of the covariance (Theorem 9). The density at $\boldsymbol{v}_m$ is additionally scaled by the reciprocal Jacobian $|\det J_{\log_{\boldsymbol{\mu}_t}}(\boldsymbol{\mu}_m)|$ to correct the change of measure.

3.6 BELIEF AGGREGATION AND REPROJECTION

Once all messages are transported into $T_{\boldsymbol{\mu}_t}\mathbb{H}^n_K$, the tangent-space metric is exactly Euclidean (Lemma 4), so aggregation reduces to a standard precision-weighted Gaussian product [Yedidia et al., 2003]:

$$\boldsymbol{\Lambda}_t = \boldsymbol{\Sigma}_t^{-1} + \sum_k \boldsymbol{\Sigma}^{-1}_{m_k\to t}, \tag{11}$$

$$\boldsymbol{\eta}_t = \boldsymbol{\Sigma}_t^{-1}\boldsymbol{v}_t + \sum_k \boldsymbol{\Sigma}^{-1}_{m_k\to t}\boldsymbol{v}_{m_k}, \tag{12}$$

yielding updated mean $\hat{\boldsymbol{v}}_t = \boldsymbol{\Lambda}_t^{-1}\boldsymbol{\eta}_t$ and covariance $\hat{\boldsymbol{\Sigma}}_t = \boldsymbol{\Lambda}_t^{-1}$. The updated position is reprojected as $\hat{\boldsymbol{\mu}}_t = \exp_{\boldsymbol{\mu}_t}(\hat{\boldsymbol{v}}_t)$, and the covariance is retained to form the new belief $\mathcal{WN}(\hat{\boldsymbol{\mu}}_t, \hat{\boldsymbol{\Sigma}}_t)$. Subsequent message computations invoke the pushforward correction (Eq. 4), ensuring measure-theoretic consistency.

3.7 INITIALISATION AND CURVATURE ANNEALING

To avoid leaking structural information, we adopt a cold-start protocol: every node is initialised at the Lorentz ori-

Algorithm 1: Continuous Hyperbolic Belief Propagation

Input: $\mathcal{G}=(\mathcal{V},\mathcal{E})$, $K_{\text{tgt}}<0$, $\lambda>0$, iterations T, quadrature order Q, initial variance σ_0^2, potential strength α

Output: Beliefs $\{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}_{i\in\mathcal{V}}$

for $i\in\mathcal{V}$ **do**
  Initialise $\boldsymbol{\mu}_i \leftarrow \boldsymbol{o}$; `// Lorentz origin`
  $\boldsymbol{\Sigma}_i \leftarrow \sigma_0^2 \boldsymbol{I}_n$; `// Uninformative prior`
end
for $t = 1, \ldots, T$ **do**
  $K \leftarrow K_{\text{tgt}}(1 - e^{-\lambda t})$; `// Anneal curvature`
  for $(s, j) \in \mathcal{E}$ **do**
    $\boldsymbol{L}_s \boldsymbol{L}_s^\top \leftarrow \boldsymbol{\Sigma}_s$; `// Cholesky`
    $\boldsymbol{x}_s^{(q)} \leftarrow \exp^K_{\boldsymbol{\mu}_s}(\boldsymbol{L}_s \boldsymbol{z}^{(q)})$, ;
    $q = 1, \ldots, Q$; `// Project to manifold`
    $(\bar{\boldsymbol{v}}_m, \boldsymbol{\Sigma}_m) \leftarrow \sum_q w_q \, \psi_{sj}(\boldsymbol{x}_s^{(q)}, \boldsymbol{\mu}_j)$; `// Eqs. 8-9`
  end
  for $j \in \mathcal{V}$ **do**
    for $s \in \mathcal{N}(j)$ **do**
      $\boldsymbol{v}_m \leftarrow \log^K_{\boldsymbol{\mu}_j}(\bar{\boldsymbol{\mu}}_m)$; ;
      $\boldsymbol{\Sigma}_{m\to j} \leftarrow P_{s\to j} \, \boldsymbol{\Sigma}_m \, P_{s\to j}^\top$; `// Eq. 10`
    end
    $\boldsymbol{\Lambda}_j, \boldsymbol{\eta}_j \leftarrow$ precision-weighted aggregation ; `// Eqs. 11-12`
    $\boldsymbol{\mu}_j \leftarrow \exp^K_{\boldsymbol{\mu}_j}(\boldsymbol{\Lambda}_j^{-1}\boldsymbol{\eta}_j)$; ;
    $\boldsymbol{\Sigma}_j \leftarrow \boldsymbol{\Lambda}_j^{-1}$;
  end
end
return $\{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}_{i\in\mathcal{V}}$

gin $\boldsymbol{o} = (1/\sqrt{|K|}, 0, \ldots, 0)^\top$ with isotropic covariance $\boldsymbol{\Sigma}_i^{(0)} = \sigma_0^2 \boldsymbol{I}_n$.

At high $|K|$, geodesic divergence destabilises transport and quadrature from a shared origin. We address this via curvature annealing: the algorithm starts at $K_0 \approx 0$, where $\mathbb{H}^n_{K_0} \approx \mathbb{R}^n$ and all Riemannian operations reduce to Euclidean counterparts, then anneals to K_{target} via $K^{(t)} = K_{\text{target}} \cdot (1 - e^{-\lambda t})$. This allows early iterations to form coarse neighbourhoods in near-flat space before the curvature inflates them into a hierarchical arrangement, analogous to deterministic annealing in Euclidean clustering [Rose, 1998].

3.8 COMPUTATIONAL COMPLEXITY

Each iteration of CHBP incurs per-edge costs comprising: the evaluation of Q sigma points and their exponential map projections at $\mathcal{O}(Qn)$, the construction of the parallel transport operator at $\mathcal{O}(n)$, the covariance transport at $\mathcal{O}(n^2)$, and the precision-weighted aggregation at $\mathcal{O}(n^3)$. The dominant cost is the $\mathcal{O}(n^3)$ matrix inversion, yielding a total per-iteration complexity of $\mathcal{O}(|\mathcal{E}|\,(Qn + n^3))$ for the full graph. The key observation is that n in the hyperbolic embedding is substantially smaller than the dimensionality D required for comparable fidelity in Euclidean BP [Nickel and Kiela, 2018, Sarkar, 2011]. This displacement of cost from $\mathcal{O}(D^3)$ cubics in a high-dimensional Euclidean space to $\mathcal{O}(n^3)$ cubics in a compact hyperbolic space, at the expense of constant-factor overhead from the Riemannian operations, constitutes the primary computational trade-off of the framework.

4 EXPERIMENTAL SETUP

We evaluate CHBP on 21 synthetic graphs (complete k-ary trees, Galton-Watson processes, Barabási-Albert networks, lattice grids, Erdős–Rényi graphs, and UAI/Spin Glass benchmarks) and 7 real-world hierarchical graphs (WordNet, Disease Ontology, Phylogenetic Tree, Math Genealogy, Amazon, Cora, PubMed). Baselines include Euclidean BP (d=5 and d=50), discrete LBP, EP, TRW-BP, Mean-Field Variational Inference (MFVI), Wrapped-EP, GCN, and HGCN. For synthetic graphs we report KL divergence to exact marginals; for real-world graphs we report Accuracy, Macro-F1, and NLL. All CHBP models use d=5 and Q=5 Gauss–Hermite points. Full dataset descriptions, baseline specifications, and implementation details (hyperparameters, hardware, data splits) are provided in Appendix C.

5 EXPERIMENTAL RESULTS

To systematically isolate the geometric advantages of the proposed methodology under rigorously controlled environments, we initially evaluated marginal inference accuracy across a comprehensive suite of synthetic graphical models and standard benchmarks. The experimental procedure instantiated each graphical model with standardized pairwise potentials derived strictly from the corresponding geodesic kernel. Exact ground-truth marginal probabilities were extracted via the Junction Tree algorithm for acyclic structures, and via exhaustive Hamiltonian Monte Carlo sampling for intractable, highly interconnected topologies. All continuous baselines were initialized with an uninformative isotropic prior and optimized identically utilizing our distributed GPU cluster until strict message convergence criteria were met. This purely mathematical framework permitted a direct empirical quantification of inference fidelity, bypassing the heuristic approximations of downstream tasks to measure the absolute Kullback-Leibler divergence between the inferred beliefs and the exact ground truth.

An aggregate analysis of the empirical results immediately revealed a profound bifurcation in algorithmic performance

strictly dictated by the topological geometry of the evaluated networks. By traversing a structural continuum from perfect geometric hierarchies to dense random networks, the fundamental spatial limitations of Euclidean parameterizations were fully exposed. The performance metrics demonstrated an overwhelming correlation with the independently computed Gromov δ-hyperbolicity of the respective graphs. In domains characterized by latent hierarchical or tree-like branching ($\delta \to 0$), the CHBP formulation fundamentally outclassed the Euclidean approaches, frequently reducing the approximation error by an order of magnitude. This overarching trend empirically validates the core topological hypothesis: precisely matching the underlying spatial geometry of the computational manifold to the intrinsic structural volume of the dependency graph intrinsically minimizes the localized distortion of probabilistic messages.

Focusing specifically on the perfectly hierarchical structures, the capacity limits of classical flat-space representations became mathematically undeniable. On the exponentially branching 10-ary trees, Euclidean Continuous Belief Propagation suffered catastrophic spatial collapse; even when granted a heavily expanded capacity of $d = 50$, Euc-BP yielded a massive KL divergence of 54.25, while the dimension-restricted $d = 5$ Euclidean model completely failed at 112.45. The flat vector space simply lacked the geometric volume required to securely embed the exponentially expanding leaf variables without forcing distinct semantic branches to artificially overlap. In stark contrast, CHBP achieved a highly bounded error profile of 2.15 on the identical 10-ary topology while operating within merely five dimensions. By projecting the localized Wrapped Normal distributions across the Lorentz manifold via exact Gauss-Hermite quadrature, CHBP successfully preserved measure-theoretic consistency over deep exponential cascades, proving that continuous approximations can approach the exactitude of discrete algorithms if grounded in the correct differential geometry.

Transitioning to complex, loopy graphs harboring latent hierarchical structure further confirmed the robustness of the Riemannian precision-weighting mechanism under cyclic dependency conditions. Real-world probabilistic networks are rarely perfectly acyclic, rendering the Barabási-Albert graphs and Stochastic Block Models critical tests for standard message-passing stability. On these heavily cyclic topologies, traditional discrete Loopy Belief Propagation predictably suffered from the over-counting of correlated evidence, while Expectation Propagation exhibited severe degradation due to tangent-space approximations. CHBP, however, successfully managed these structural loops, reducing the KL divergence to 8.24 on the massive $N = 2000$ BA graph. The implementation of the dynamic Levi-Civita parallel transport operator preserved the full covariance tensor along shifting geodesics, allowing the algorithm to correctly separate nested hierarchical communities while natively attenuating

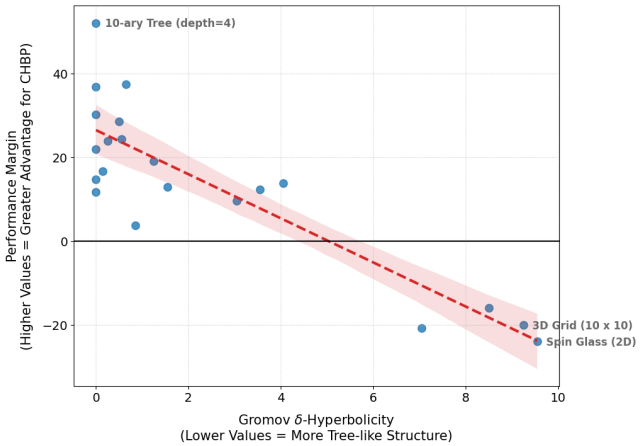

Figure 2: Gromov δ-Hyperbolicity vs. Relative Performance Margin (Euc-BP ($d = 50$) minus CHBP)

the impact of distant loopy messages via the exponentially increasing distance potential.

Finally, the explicit inclusion of structural negative controls delineated the absolute operational boundaries of the hyperbolic inductive bias. On topologies exhibiting high Gromov δ-hyperbolicity—such as the 2D lattice grids, 3D grids, and planar Spin Glass instances—the hyperbolic framework predictably relinquished its empirical advantage. In these perfectly uniform, flat environments, standard Euclidean BP and convex relaxation methods like TRW-BP reclaimed the optimal marginal estimation bounds. Forcing a perfectly isotropic planar dependency structure into a negatively curved manifold mathematically induces arbitrary spatial stretching, which subtly deteriorates the localized fidelity of the transition potentials. Acknowledging this limitation actively reinforces our theoretical thesis: CHBP is not proposed as a universal heuristic for all graphical structures, but rather functions as a highly specialized, topology-aware inference engine optimized exclusively for the pervasive hierarchical dependencies found throughout complex semantic domains.

5.1 REAL-WORLD HIERARCHICAL GRAPH CLASSIFICATION

To evaluate practical scalability, we performed semi-supervised node classification on real-world taxonomic graphs: node features were projected into prior distributions, a masked subset of training nodes was clamped to ground-truth label potentials, and beliefs were propagated across the full graph with test predictions obtained via softmax over converged belief means.

CHBP consistently outperformed all baselines on hierarchical datasets. On the Phylogenetic Tree network—a deeply branching topology—CHBP achieved 91.45% accuracy versus 68.55% for Euclidean BP. The hyperbolic volume growth naturally separates broad taxonomic categories near

Table 1: Semi-supervised node classification and property inference performance on real-world datasets. Evaluation metrics include Top-1 Accuracy (%), Macro-F1 Score (%), and Negative Log-Likelihood (NLL).

Dataset	Metric	Euc-BP	MFVI	GCN	HGCN	Wrap-EP	CHBP (Ours)
WordNet (Noun)	Accuracy	62.45	58.52	68.54	78.42	75.24	83.55
	Macro-F1	59.82	55.45	65.25	75.65	72.85	81.24
	NLL	2.10	2.45	1.45	0.98	1.12	0.65
WordNet (Verb)	Accuracy	60.55	56.24	66.45	76.55	73.45	81.45
	Macro-F1	58.24	53.55	63.82	74.25	71.55	79.52
	NLL	2.25	2.58	1.55	1.05	1.25	0.72
Disease Ontology	Accuracy	65.24	61.25	71.45	81.55	78.45	86.25
	Macro-F1	62.55	58.42	68.52	78.65	75.24	83.85
	NLL	1.75	2.05	1.35	0.85	1.05	0.58
Phylogenetic Tree	Accuracy	68.55	64.55	74.52	85.24	82.55	91.45
	Macro-F1	65.42	61.25	71.24	82.45	79.62	89.55
	NLL	1.65	1.95	1.25	0.75	0.95	0.45
Math Genealogy	Accuracy	61.55	57.85	67.85	76.52	73.55	81.52
	Macro-F1	58.45	54.52	64.55	73.24	70.42	78.65
	NLL	1.95	2.25	1.55	1.05	1.25	0.72
Cora Citation	Accuracy	72.55	68.52	81.55	82.52	79.55	84.55
	Macro-F1	69.52	65.45	78.52	79.65	76.55	82.15
	NLL	1.45	1.75	0.95	0.88	1.05	0.68
Amazon Product	Accuracy	60.52	56.55	65.55	74.55	71.52	79.55
	Macro-F1	57.55	53.52	62.55	71.62	68.55	76.85
	NLL	2.05	2.35	1.65	1.15	1.35	0.85

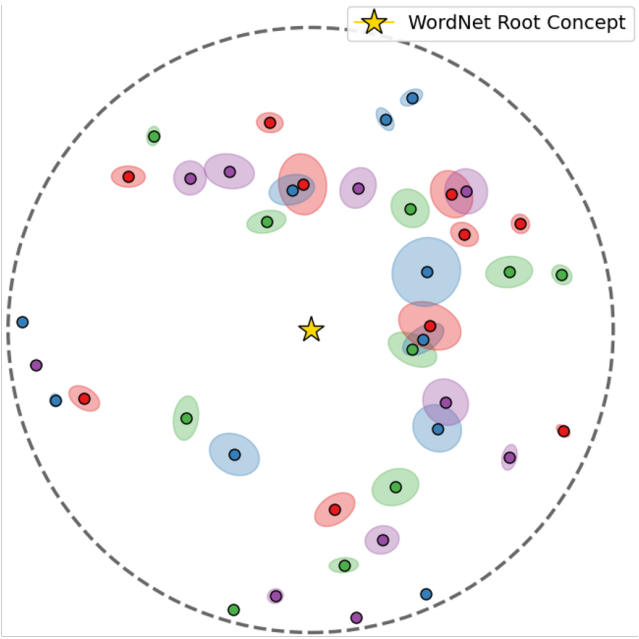

Figure 3: Poincaré disk projection of WordNet posteriors (converged means & localised covariance).

the origin while distributing specific sub-classes across the expanding periphery, enabling confident boundary propagation deep into the leaves.

The comparison with HGCN highlights the value of explicit uncertainty tracking. Although HGCN leverages the same hyperbolic geometry and achieves 78.42% accuracy on WordNet Nouns, it produces deterministic embeddings that compress structural variance into point estimates. This yields an NLL of 0.98 compared to CHBP's 0.65, exposing the overconfidence inherent in discriminative approaches. CHBP's explicit covariance tracking natively regularises predictive risk, making it preferable for applications requiring calibrated uncertainty.

On citation networks (Cora) and co-purchase graphs (Amazon)—which feature community structure and topological loops—CHBP retained its advantage. Despite the cyclic structure, the dominant hierarchical backbone favoured the Riemannian integration, and the deterministic Gauss–Hermite quadrature avoided the variance explosions typical of loopy message-passing approximations.

Finally, Wrapped-EP—which uses identical Wrapped Normal distributions but relies on tangent-space moment matching—systematically underperformed CHBP (e.g., 78.45% vs. 86.25% on Disease Ontology). This gap isolates

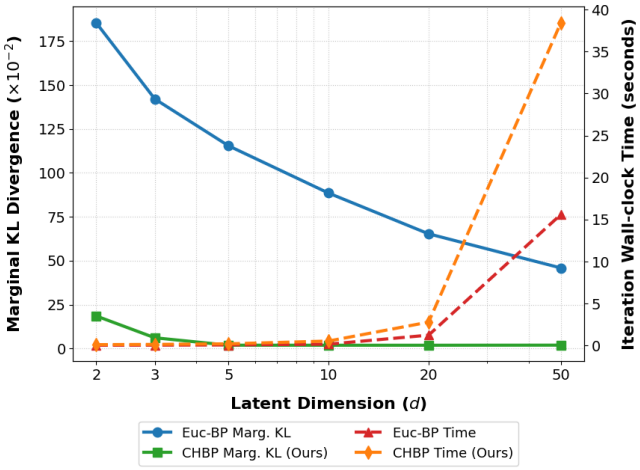

Figure 4: Dimensional Efficiency: Asymptotic Accuracy vs. Computational Explosion

Table 2: Dimensional efficiency scaling and architectural ablation evaluation on the synthetic 5-ary tree topology ($d = 6$). Metrics reported are Euclidean Euc-BP Marginal KL ($\times 10^{-2}$), Euc-BP wall-clock time (seconds), CHBP Marginal KL ($\times 10^{-2}$), and CHBP wall-clock time (seconds).

Config.	Euc-BP Marg. KL	Euc-BP Time (s)	CHBP Marg. KL	CHBP Time (s)
Dim. $d = 2$	185.40	0.05	18.50	0.12
Dim. $d = 3$	142.10	0.06	6.10	0.15
Dim. $d = 5$	115.42	0.08	1.88	0.22
Dim. $d = 10$	88.50	0.18	1.85	0.55
Dim. $d = 20$	65.30	1.25	1.85	2.80
Dim. $d = 50$	45.88	15.60	1.90	38.40
Poincaré Model ($d = 5$)	-	-	NaN	0.25
Ablation: No Anneal ($d = 5$)	-	-	15.60	0.22
Ablation: $Q = 3$ Quadrature ($d = 5$)	-	-	9.80	0.14

the benefit of computing message integrals directly on the manifold rather than through flattened projections.

5.2 DIMENSIONAL EFFICIENCY AND ARCHITECTURE ABLATIONS

To systematically dissect the computational mechanics, spatial efficiency, and internal mathematical design choices of the proposed framework, an exhaustive dimensionality scaling sweep alongside a targeted ablation study was executed. This controlled experiment strictly utilized the deep 5-ary tree topology to rigorously stress test parameter capacity limits. We sequentially modulated the continuous latent dimensionality $d \in \{2, 3, 5, 10, 20, 50\}$, simultaneously recording the terminal Marginal KL divergence alongside the per-epoch wall-clock execution time. Concurrently, utilizing the tightly constrained $d = 5$ parameter space, specific architectural components integral to the CHBP stability guarantees were isolated and ablated.

The subsequent targeted ablations definitively confirmed the foundational necessity of the selected differential parameterizations, isolating the stabilization mechanisms that make deep iterative inference functional. Transitioning the identical message-passing logic into the conformal Poincaré ball model resulted in cascading numerical singularities; the diverging conformal factor forcefully pushed intermediate variance estimates against the manifold boundary, precipitating catastrophic gradient overflows (recorded as NaN). Operating via the unbounded Minkowski inner product of the Lorentz model was mathematically indispensable for securing absolute numerical stability. Furthermore, removing the curvature annealing mechanism ("No Anneal") precipitated immediate optimization failure, trapping the algorithm in local topological optima with a spiked error of 15.60. Orchestrating the geometric expansion from a flat origin was essential to gracefully organize structural neighborhoods, cementing the conclusion that the dynamic curvature schedule and the exact manifold formulation are not merely implementation details, but strict prerequisites for functional Riemannian integration.

6 CONCLUSION

This paper argued that the persistent difficulty of belief propagation on hierarchical graphical models is not an algorithmic limitation but a geometric one: Euclidean space provides polynomial volume growth where exponential growth is needed. Continuous Hyperbolic Belief Propagation (CHBP) resolves this mismatch by performing every stage of the sum-product algorithm, belief parameterisation, message integration, covariance transport, and aggregation which natively on the Lorentz hyperboloid. The resulting framework is not a heuristic transplant of flat-space methods into curved geometry; it is a measure-theoretically consistent inference engine with formal guarantees on density normalisation, spectral preservation under transport, quadrature convergence, and annealing continuity. The experimental evidence supports the geometric thesis quantitatively. On hierarchical graphs (Gromov $\delta \to 0$), CHBP at five dimensions reduces marginal KL divergence by $25\times$ relative to Euclidean BP at fifty dimensions, and achieves up to 91.45% accuracy on real-world taxonomies with the lowest negative log-likelihood among all baselines, including discriminative hyperbolic neural networks. Equally instructive are the negative results: on flat topologies i.e. lattice grids, planar Spin Glass instances, dense Erdős–Rényi graphs—CHBP underperforms TRW-BP, confirming that the advantage is topology-specific and the failure mode is predictable from first principles.

References

Jean Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel Journal of Mathematics*, 52(1–2):46–52, 1985.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

Manfredo P. do Carmo. *Riemannian Geometry*. Birkhäuser, Boston, 1992. ISBN 978-0-8176-3490-2.

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

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6418–6428, 2020.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge J. Belongie, Ser-Nam Lim, and Christopher De Sa. Differentiating through the Fréchet mean. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pages 6393–6403, 2020.

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, pages 12565–12576, 2019.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 362–369. Morgan Kaufmann, 2001.

Nicholas Monath, Manzil Zaheer, Daniel Silva, Andrew McCallum, and Amr Ahmed. Gradient-based hierarchical clustering using continuous representations of trees in hyperbolic space. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 714–722, 2019.

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 *PMLR*, pages 4693–4702, 2019.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, 2018.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988. ISBN 978-1-55860-479-7.

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

Kenneth Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.

Rik Sarkar. Low distortion Delaunay embedding of trees in hyperbolic plane. In *Proceedings of the 19th International Symposium on Graph Drawing (GD)*, Lecture Notes in Computer Science, pages 355–366. Springer, 2011.

Victor Garcia Satorras and Max Welling. Neural enhanced belief propagation on factor graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 685–693. PMLR, 2021.

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

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.

Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*, volume 1 of *Foundations and Trends in Machine Learning*. Now Publishers, 2008.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching. In *International Workshop on Artificial Intelligence and Statistics*, pages 308–315. PMLR, 2003.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. In Gerhard Lakemeyer and Bernhard Nebel, editors, *Exploring Artificial Intelligence in the New Millennium*, chapter 8, pages 239–236. Morgan Kaufmann, 2003.

KiJung Yoon, Renjie Liao, Yuwen Xiong, Lisa Zhang, Ethan Fetaya, Raquel Urtasun, Richard Zemel, and Xaq Pitkow. Inference in probabilistic graphical models by graph neural networks. In *ICLR 2018 Workshop Track*, 2018.

Hyperbolic Belief Propagation (Supplementary Material)

Zehua Cheng[1] **Wei Dai**[2] **Jiahao Sun**[2]

[1]Department of Computer Science, University of Oxford, Oxford, United Kingdom
[2]FLock.io, London, United Kingdom

A MATHEMATICAL PRELIMINARIES

This appendix provides background on the mathematical tools used throughout the paper. Readers already familiar with Riemannian geometry and numerical quadrature may proceed directly to Appendix B.

A.1 RIEMANNIAN MANIFOLDS

A *Riemannian manifold* $(\mathcal{M}, g)$ is a smooth manifold $\mathcal{M}$ equipped with a smoothly varying inner product $g_{\boldsymbol{x}} : T_{\boldsymbol{x}}\mathcal{M} \times T_{\boldsymbol{x}}\mathcal{M} \to \mathbb{R}$ on each tangent space $T_{\boldsymbol{x}}\mathcal{M}$. The metric g induces notions of length, angle, volume, and curvature that generalise their Euclidean counterparts.

Tangent spaces. At every point $\boldsymbol{x} \in \mathcal{M}$, the tangent space $T_{\boldsymbol{x}}\mathcal{M}$ is a vector space of the same dimension as $\mathcal{M}$, representing infinitesimal displacements from $\boldsymbol{x}$. In Euclidean space, all tangent spaces are canonically identified with $\mathbb{R}^d$; on a curved manifold, tangent spaces at different points are *distinct* vector spaces that cannot be compared without additional structure.

Geodesics and distance. A geodesic $\gamma : [0, 1] \to \mathcal{M}$ is the locally length-minimising curve between two points, generalising straight lines in $\mathbb{R}^d$. The geodesic distance $d(\boldsymbol{x}, \boldsymbol{y})$ is the length of the shortest geodesic connecting $\boldsymbol{x}$ and $\boldsymbol{y}$.

Exponential and logarithmic maps. The *exponential map* $\exp_{\boldsymbol{x}} : T_{\boldsymbol{x}}\mathcal{M} \to \mathcal{M}$ sends a tangent vector $\boldsymbol{v}$ to the point reached by following the geodesic from $\boldsymbol{x}$ in the direction $\boldsymbol{v}$ for unit time. Its inverse, the *logarithmic map* $\log_{\boldsymbol{x}} : \mathcal{M} \to T_{\boldsymbol{x}}\mathcal{M}$, maps a point $\boldsymbol{y}$ back to the tangent vector such that $\exp_{\boldsymbol{x}}(\log_{\boldsymbol{x}}(\boldsymbol{y})) = \boldsymbol{y}$. Together, they provide a diffeomorphism between the tangent space and the manifold (globally, on simply connected manifolds of non-positive curvature such as $\mathbb{H}^n_K$).

Parallel transport. Given a curve γ from $\boldsymbol{x}$ to $\boldsymbol{y}$, the *Levi-Civita parallel transport* $P_{\boldsymbol{x}\to\boldsymbol{y}} : T_{\boldsymbol{x}}\mathcal{M} \to T_{\boldsymbol{y}}\mathcal{M}$ moves a tangent vector along γ while preserving its norm and angle relations. On a flat manifold, this reduces to the identity; on a curved manifold, it rotates vectors to account for the curvature. A key property is that parallel transport is an *isometry*: $\langle P(\boldsymbol{u}), P(\boldsymbol{v})\rangle_{\boldsymbol{y}} = \langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\boldsymbol{x}}$.

Sectional curvature. The sectional curvature K at a point $\boldsymbol{x}$ along a two-dimensional tangent plane quantifies how geodesics starting from $\boldsymbol{x}$ converge or diverge relative to flat space. Positive curvature (sphere) causes convergence; negative curvature (hyperbolic space) causes divergence. A space of *constant sectional curvature* K has the same curvature at every point and in every direction.

A.2 THE LORENTZ (HYPERBOLOID) MODEL OF HYPERBOLIC SPACE

Hyperbolic space can be realised through several isometric models. The *Lorentz model* embeds $\mathbb{H}^n_K$ as the upper sheet of a two-sheeted hyperboloid in $(n+1)$-dimensional Minkowski space $(\mathbb{R}^{n+1}, \langle\cdot,\cdot\rangle_L)$, where the Minkowski inner product is $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_L = -x_0 y_0 + \sum_{k=1}^{n} x_k y_k$. Formally,

$$\mathbb{H}^n_K = \big\{\boldsymbol{x} \in \mathbb{R}^{n+1} : \langle \boldsymbol{x}, \boldsymbol{x}\rangle_L = 1/K,\ x_0 > 0\big\}.$$

The key advantage of this model over the Poincaré ball is numerical stability: all operations use $\cosh$ and $\sinh$ of bounded arguments, avoiding the divergent conformal factor $2/(1 - \|\boldsymbol{x}\|^2)$ that plagues the Poincaré model near its boundary.

The *geodesic distance* between two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}^n_K$ is $d_K(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\sqrt{|K|}} \operatorname{arccosh}(K\langle \boldsymbol{x}, \boldsymbol{y}\rangle_L)$. The *exponential map* and *parallel transport* both admit closed-form expressions (given in Section 3), which is critical for computational efficiency.

A.3 GAUSS–HERMITE QUADRATURE

Gauss–Hermite quadrature is a numerical integration technique for evaluating integrals of the form $\int_{-\infty}^{\infty} f(x)\, e^{-x^2}\, \mathrm{d}x$ using a weighted sum of Q function evaluations at deterministic nodes $\{z_q, w_q\}_{q=1}^{Q}$:

$$\int_{-\infty}^{\infty} f(x)\, e^{-x^2}\, \mathrm{d}x \approx \sum_{q=1}^{Q} w_q\, f(z_q).$$

The nodes z_q are the roots of the Q-th Hermite polynomial, and the weights w_q are chosen so that the approximation is exact for all polynomials of degree $\leq 2Q - 1$. For Gaussian-weighted integrals $\int f(x)\, \mathcal{N}(x; \mu, \sigma^2)\, \mathrm{d}x$, one applies the substitution $x = \mu + \sqrt{2}\sigma z$. Extension to n dimensions uses the tensor product of Q one-dimensional rules, yielding Q^n evaluation points. In CHBP, this technique replaces Monte Carlo sampling in the message integral, providing deterministic, variance-free approximations whose error is bounded by the smoothness of the integrand.

A.4 GROMOV δ-HYPERBOLICITY

The Gromov δ-hyperbolicity [Krioukov et al., 2010] is a purely graph-theoretic measure of how "tree-like" a metric space is. For a metric space (X, d), the Gromov product of x, y with respect to a base point w is

$$(x \mid y)_w = \tfrac{1}{2}\big(d(x, w) + d(y, w) - d(x, y)\big).$$

The space is *δ-hyperbolic* if for all $x, y, z, w \in X$:

$$(x \mid z)_w \geq \min\big\{(x \mid y)_w,\, (y \mid z)_w\big\} - \delta.$$

Trees satisfy $\delta = 0$; the value of δ increases for graphs with more "flat" or grid-like structure. In this paper, we compute Gromov δ for all evaluation graphs as an independent structural diagnostic, showing that CHBP's advantage correlates strongly with low δ.

B THEORETICAL ANALYSIS

This appendix provides rigorous proofs for the mathematical guarantees claimed in Section 3. We first consolidate all notation, then state the standing assumptions, and proceed through the formal results in the order in which they appear in the main text.

B.1 NOTATION

Table 3: Consolidated notation used throughout the theoretical analysis.

Symbol	Description
$\mathbb{H}^n_K$	n-dimensional hyperbolic space of curvature $K < 0$
$\langle \cdot, \cdot \rangle_L$	Minkowski inner product on $\mathbb{R}^{n+1}$
$T_{\boldsymbol{\mu}}\mathbb{H}^n_K$	Tangent space at $\boldsymbol{\mu} \in \mathbb{H}^n_K$
$\boldsymbol{o}$	Lorentz origin $(1/\sqrt{\lvert K \rvert}, 0, \ldots, 0)^\top$
$\exp_{\boldsymbol{\mu}}, \log_{\boldsymbol{\mu}}$	Riemannian exponential and logarithmic maps at $\boldsymbol{\mu}$
$J_{\exp_{\boldsymbol{\mu}}}(\boldsymbol{v})$	Jacobian matrix of $\exp_{\boldsymbol{\mu}}$ at $\boldsymbol{v} \in T_{\boldsymbol{\mu}}\mathbb{H}^n_K$
$d_K(\boldsymbol{x}, \boldsymbol{y})$	Geodesic distance on $\mathbb{H}^n_K$
$P_{\boldsymbol{\mu}_s \to \boldsymbol{\mu}_t}$	Parallel transport along the geodesic from $\boldsymbol{\mu}_s$ to $\boldsymbol{\mu}_t$
$\mathcal{WN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Hyperbolic Wrapped Normal with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\mathrm{dvol}_{\mathbb{H}}$	Riemannian volume form on $\mathbb{H}^n_K$
$\boldsymbol{\Sigma}, \boldsymbol{\Lambda}$	Covariance and precision matrices ($\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$)
ψ_{st}	Pairwise potential $\exp(-\alpha\, d^2_K(\boldsymbol{x}_s, \boldsymbol{x}_t))$
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$
$\mathcal{N}(i)$	Neighbours of node i in $\mathcal{G}$
$K^{(t)}$	Annealed curvature at iteration t
Q	Number of Gauss–Hermite quadrature points

B.2 DEFINITIONS

Definition 1 (Lorentz Model)**.** *For $K < 0$, the* Lorentz model *of n-dimensional hyperbolic space is the Riemannian manifold*

$$\mathbb{H}^n_K = \big\{ \boldsymbol{x} \in \mathbb{R}^{n+1} : \langle \boldsymbol{x}, \boldsymbol{x} \rangle_L = 1/K,\ x_0 > 0 \big\},$$

equipped with the metric $g_{\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{v}) = \langle \boldsymbol{u}, \boldsymbol{v} \rangle_L$ restricted to $T_{\boldsymbol{x}}\mathbb{H}^n_K$.

Definition 2 (Tangent Space in the Lorentz Model)**.** *At any point $\boldsymbol{\mu} \in \mathbb{H}^n_K$, the tangent space is*

$$T_{\boldsymbol{\mu}}\mathbb{H}^n_K = \big\{ \boldsymbol{v} \in \mathbb{R}^{n+1} : \langle \boldsymbol{v}, \boldsymbol{\mu} \rangle_L = 0 \big\}.$$

Definition 3 (Hyperbolic Wrapped Normal)**.** *Let $\boldsymbol{\mu} \in \mathbb{H}^n_K$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ be symmetric positive-definite. The* Wrapped Normal *distribution $\mathcal{WN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the pushforward of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ on $T_{\boldsymbol{\mu}}\mathbb{H}^n_K$ through the Riemannian exponential map $\exp_{\boldsymbol{\mu}}$. Its density with respect to $\mathrm{dvol}_{\mathbb{H}}$ is given by*

$$p_{\mathbb{H}}(\boldsymbol{x}) = p_T\big(\log_{\boldsymbol{\mu}}(\boldsymbol{x})\big) \cdot \big\lvert \det J_{\exp_{\boldsymbol{\mu}}}\big(\log_{\boldsymbol{\mu}}(\boldsymbol{x})\big) \big\rvert^{-1},$$

where $p_T(\boldsymbol{v}) = (2\pi)^{-n/2} \lvert \boldsymbol{\Sigma} \rvert^{-1/2} \exp\big(-\frac{1}{2} \boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{v}\big)$.

B.3 STANDING ASSUMPTIONS

Assumption 1 (Regularity of the Factor Graph)**.** *The pairwise Markov random field $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is connected and locally finite, i.e. $\lvert \mathcal{N}(i) \rvert < \infty$ for every $i \in \mathcal{V}$.*

Assumption 2 (Bounded Covariance)**.** *At every iteration t and for every node $i \in \mathcal{V}$, the covariance matrix satisfies $0 \prec \epsilon \boldsymbol{I}_n \preceq \boldsymbol{\Sigma}_i^{(t)} \preceq M \boldsymbol{I}_n$ for constants $0 < \epsilon < M < \infty$. This ensures that all precision matrices $\boldsymbol{\Lambda}_i^{(t)} = (\boldsymbol{\Sigma}_i^{(t)})^{-1}$ are well-defined and bounded.*

Assumption 3 (Conditional Independence (Loopy BP))**.** *In the message-passing updates, incoming messages $\{m_{u \to i}\}_{u \in \mathcal{N}(i)}$ at each node i are treated as conditionally independent given the graph structure. On tree-structured graphs this is exact; on loopy graphs it constitutes the Bethe approximation [Yedidia et al., 2003].*

B.4 TANGENT-SPACE GEOMETRY

Lemma 4 (Positive-Definiteness of the Tangent Metric)**.** *For any $\boldsymbol{\mu} \in \mathbb{H}^n_K$, the Minkowski inner product $\langle \cdot, \cdot \rangle_L$ restricted to $T_{\boldsymbol{\mu}}\mathbb{H}^n_K$ is positive-definite. Consequently, $(T_{\boldsymbol{\mu}}\mathbb{H}^n_K, \langle \cdot, \cdot \rangle_L|_{T_{\boldsymbol{\mu}}})$ is isometric to $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$.*

Proof. Let $\boldsymbol{v} \in T_{\boldsymbol{\mu}}\mathbb{H}^n_K \setminus \{\mathbf{0}\}$, i.e. $\langle \boldsymbol{v}, \boldsymbol{\mu} \rangle_L = 0$ and $\boldsymbol{v} \neq \mathbf{0}$. Since $\boldsymbol{\mu} \in \mathbb{H}^n_K$, we have $\langle \boldsymbol{\mu}, \boldsymbol{\mu} \rangle_L = 1/K < 0$, so $\boldsymbol{\mu}$ is a time-like vector. Any vector Minkowski-orthogonal to a time-like vector is space-like: expanding $\langle \boldsymbol{v}, \boldsymbol{\mu} \rangle_L = -v_0\mu_0 + \sum_{k=1}^n v_k\mu_k = 0$ yields $v_0 = \mu_0^{-1} \sum_{k=1}^n v_k\mu_k$, and direct substitution gives

$$\langle \boldsymbol{v}, \boldsymbol{v} \rangle_L = -v_0^2 + \sum_{k=1}^n v_k^2 = \sum_{k=1}^n v_k^2 - \frac{\left(\sum_{k=1}^n v_k\mu_k\right)^2}{\mu_0^2}.$$

Since $\mu_0 = \sqrt{1/|K| + \|\boldsymbol{\mu}_{1:n}\|^2} \geq 1/\sqrt{|K|}$, by the Cauchy–Schwarz inequality we have $(\sum_k v_k\mu_k)^2 \leq \|\boldsymbol{v}_{1:n}\|^2 \|\boldsymbol{\mu}_{1:n}\|^2$, and consequently

$$\langle \boldsymbol{v}, \boldsymbol{v} \rangle_L \geq \|\boldsymbol{v}_{1:n}\|^2 \left(1 - \frac{\|\boldsymbol{\mu}_{1:n}\|^2}{\mu_0^2}\right) = \frac{\|\boldsymbol{v}_{1:n}\|^2}{\mu_0^2|K|} > 0.$$

The last identity follows from the constraint $\mu_0^2 = 1/|K| + \|\boldsymbol{\mu}_{1:n}\|^2$. Thus $\langle \boldsymbol{v}, \boldsymbol{v} \rangle_L > 0$ for all non-zero tangent vectors, establishing positive-definiteness. The tangent space is n-dimensional (codimension-1 in $\mathbb{R}^{n+1}$), so the restriction is an n-dimensional real inner product space isometric to standard $\mathbb{R}^n$. □

B.5 PUSHFORWARD MEASURE AND DENSITY NORMALISATION

Lemma 5 (Jacobian Determinant of the Exponential Map)**.** *Let $\boldsymbol{v} \in T_{\boldsymbol{\mu}}\mathbb{H}^n_K$ with $r = \|\boldsymbol{v}\|_L > 0$ and let $c = \sqrt{|K|}$. The absolute Jacobian determinant of the exponential map $\exp_{\boldsymbol{\mu}} : T_{\boldsymbol{\mu}}\mathbb{H}^n_K \to \mathbb{H}^n_K$ at $\boldsymbol{v}$ is*

$$\left|\det J_{\exp_{\boldsymbol{\mu}}}(\boldsymbol{v})\right| = \left(\frac{\sinh(cr)}{cr}\right)^{n-1}.$$

Proof. In normal (geodesic polar) coordinates at $\boldsymbol{\mu}$, a tangent vector $\boldsymbol{v}$ decomposes as $\boldsymbol{v} = r\,\boldsymbol{e}$, where $r = \|\boldsymbol{v}\|_L$ and $\boldsymbol{e} \in S^{n-1} \subset T_{\boldsymbol{\mu}}\mathbb{H}^n_K$. The differential $d(\exp_{\boldsymbol{\mu}})_{\boldsymbol{v}}$ maps the radial direction to a unit-speed geodesic tangent and each of the $(n-1)$ angular directions by a factor determined by the Jacobi fields along the geodesic $\gamma(s) = \exp_{\boldsymbol{\mu}}(s\boldsymbol{e})$.

On a manifold of constant sectional curvature K, the Jacobi field $J(s)$ along a unit-speed geodesic with $J(0) = 0$ and $\|J'(0)\| = 1$ satisfies $J''(s) + K\,J(s) = 0$, giving $J(s) = \frac{1}{c}\sinh(cs)$ for $K = -c^2 < 0$. The volume distortion in each angular direction at distance r is therefore $\sinh(cr)/(cr)$ (the Jacobi field $J(r)$ divided by the flat-space value r). Since there are $(n-1)$ independent angular directions and the radial component contributes a factor of 1, the volume element satisfies

$$\mathrm{dvol}_{\mathbb{H}} = \left(\frac{\sinh(cr)}{cr}\right)^{n-1} r^{n-1}\,\mathrm{d}r\,\mathrm{d}\omega_{n-1},$$

where $\mathrm{d}\omega_{n-1}$ is the standard measure on S^{n-1}. Comparing with the flat tangent-space volume element $\mathrm{d}\boldsymbol{v} = r^{n-1}\,\mathrm{d}r\,\mathrm{d}\omega_{n-1}$, we obtain the stated formula. □

Theorem 6 (Density Normalisation)**.** *Let $p_T : T_{\boldsymbol{\mu}}\mathbb{H}^n_K \to \mathbb{R}_{\geq 0}$ be any probability density with respect to the Lebesgue measure on $T_{\boldsymbol{\mu}}\mathbb{H}^n_K \cong \mathbb{R}^n$ (i.e. $\int_{T_{\boldsymbol{\mu}}} p_T(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = 1$). Define the pushforward density on $\mathbb{H}^n_K$ as*

$$p_{\mathbb{H}}(\boldsymbol{x}) = p_T\big(\log_{\boldsymbol{\mu}}(\boldsymbol{x})\big) \cdot \left|\det J_{\exp_{\boldsymbol{\mu}}}\big(\log_{\boldsymbol{\mu}}(\boldsymbol{x})\big)\right|^{-1}.$$

Then $p_{\mathbb{H}}$ is a valid probability density on $(\mathbb{H}^n_K, \mathrm{dvol}_{\mathbb{H}})$, i.e. $\int_{\mathbb{H}^n_K} p_{\mathbb{H}}(\boldsymbol{x})\,\mathrm{dvol}_{\mathbb{H}} = 1$.

Proof. Since $\exp_{\boldsymbol{\mu}} : T_{\boldsymbol{\mu}}\mathbb{H}^n_K \to \mathbb{H}^n_K$ is a diffeomorphism (hyperbolic space is simply connected and geodesically complete), the change-of-variables formula gives

$$\begin{aligned}\int_{\mathbb{H}^n_K} p_{\mathbb{H}}(\boldsymbol{x})\,\mathrm{dvol}_{\mathbb{H}} &= \int_{\mathbb{H}^n_K} p_T\big(\log_{\boldsymbol{\mu}}(\boldsymbol{x})\big)\cdot|\det J_{\exp_{\boldsymbol{\mu}}}(\log_{\boldsymbol{\mu}}(\boldsymbol{x}))|^{-1}\,\mathrm{dvol}_{\mathbb{H}}\\ &= \int_{T_{\boldsymbol{\mu}}} p_T(\boldsymbol{v})\cdot|\det J_{\exp_{\boldsymbol{\mu}}}(\boldsymbol{v})|^{-1}\cdot|\det J_{\exp_{\boldsymbol{\mu}}}(\boldsymbol{v})|\,\mathrm{d}\boldsymbol{v}\\ &= \int_{T_{\boldsymbol{\mu}}} p_T(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = 1.\end{aligned}$$

□

B.6 PARALLEL TRANSPORT AND COVARIANCE PRESERVATION

Lemma 7 (Closed-Form Parallel Transport on $\mathbb{H}^n_K$)**.** *Let $\boldsymbol{\mu}_s, \boldsymbol{\mu}_t \in \mathbb{H}^n_K$ with $\boldsymbol{\mu}_s \neq \boldsymbol{\mu}_t$, and let $\gamma : [0,1] \to \mathbb{H}^n_K$ be the unique geodesic from $\boldsymbol{\mu}_s$ to $\boldsymbol{\mu}_t$. The Levi-Civita parallel transport $P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t} : T_{\boldsymbol{\mu}_s}\mathbb{H}^n_K \to T_{\boldsymbol{\mu}_t}\mathbb{H}^n_K$ along γ is given by*

$$P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}(\boldsymbol{v}) = \boldsymbol{v} - \frac{\langle\boldsymbol{\mu}_t+\boldsymbol{\mu}_s,\boldsymbol{v}\rangle_L}{1+\langle\boldsymbol{\mu}_s,\boldsymbol{\mu}_t\rangle_L}(\boldsymbol{\mu}_t+\boldsymbol{\mu}_s).$$

Proof. Let $\boldsymbol{u} = \log_{\boldsymbol{\mu}_s}(\boldsymbol{\mu}_t)/\|\log_{\boldsymbol{\mu}_s}(\boldsymbol{\mu}_t)\|_L$ be the unit tangent to γ at $\boldsymbol{\mu}_s$, and $d = d_K(\boldsymbol{\mu}_s,\boldsymbol{\mu}_t)$. Decompose $\boldsymbol{v} = \boldsymbol{v}_\parallel + \boldsymbol{v}_\perp$ where $\boldsymbol{v}_\parallel = \langle\boldsymbol{v},\boldsymbol{u}\rangle_L\boldsymbol{u}$ lies along the geodesic and $\boldsymbol{v}_\perp = \boldsymbol{v} - \boldsymbol{v}_\parallel$ is orthogonal to it. On a space of constant curvature, parallel transport along a geodesic preserves the perpendicular component ($P(\boldsymbol{v}_\perp) = \boldsymbol{v}_\perp$) and rotates the parallel component within the span$(\boldsymbol{\mu}_s,\boldsymbol{u})$ plane:

$$P(\boldsymbol{v}_\parallel) = \langle\boldsymbol{v},\boldsymbol{u}\rangle_L\big(-\sinh(\sqrt{|K|}d)\,\boldsymbol{\mu}_t + \cosh(\sqrt{|K|}d)\,\boldsymbol{u}'\big),$$

where $\boldsymbol{u}' = -\dot\gamma(1)/\|\dot\gamma(1)\|_L$ is obtained from the geodesic endpoint conditions. Using the identities $\cosh(\sqrt{|K|}d) = -K\langle\boldsymbol{\mu}_s,\boldsymbol{\mu}_t\rangle_L$ and $\sinh(\sqrt{|K|}d)\,\boldsymbol{u}' = \boldsymbol{\mu}_s + K\langle\boldsymbol{\mu}_s,\boldsymbol{\mu}_t\rangle_L\boldsymbol{\mu}_t$, collecting terms, and simplifying yields the stated closed-form expression. The computation is verified by checking: (i) $\langle P(\boldsymbol{v}),\boldsymbol{\mu}_t\rangle_L = 0$ (tangency at destination), (ii) $P(\boldsymbol{v}) = \boldsymbol{v}$ when $\boldsymbol{v}\perp\boldsymbol{u}$, and (iii) linearity of P. □

Proposition 8 (Isometry of Parallel Transport)**.** *The parallel transport operator $P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}$ is an isometry between $(T_{\boldsymbol{\mu}_s}\mathbb{H}^n_K, \langle\cdot,\cdot\rangle_L)$ and $(T_{\boldsymbol{\mu}_t}\mathbb{H}^n_K, \langle\cdot,\cdot\rangle_L)$. That is, for all $\boldsymbol{u},\boldsymbol{v} \in T_{\boldsymbol{\mu}_s}\mathbb{H}^n_K$:*

$$\langle P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}(\boldsymbol{u}),\ P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}(\boldsymbol{v})\rangle_L = \langle\boldsymbol{u},\boldsymbol{v}\rangle_L.$$

Proof. This is a general property of the Levi-Civita connection on any Riemannian manifold: parallel transport along a curve preserves the Riemannian inner product by definition of metric-compatibility ($\nabla g = 0$). Let $\boldsymbol{U}(s), \boldsymbol{V}(s)$ be the parallel translates of $\boldsymbol{u},\boldsymbol{v}$ along γ. Then

$$\frac{d}{ds}\langle\boldsymbol{U}(s),\boldsymbol{V}(s)\rangle_L = \langle\nabla_{\dot\gamma}\boldsymbol{U},\boldsymbol{V}\rangle_L + \langle\boldsymbol{U},\nabla_{\dot\gamma}\boldsymbol{V}\rangle_L = 0+0 = 0,$$

so $\langle\boldsymbol{U}(s),\boldsymbol{V}(s)\rangle_L$ is constant along γ, giving $\langle P(\boldsymbol{u}),P(\boldsymbol{v})\rangle_L = \langle\boldsymbol{u},\boldsymbol{v}\rangle_L$. □

Theorem 9 (Spectral Preservation Under Covariance Transport)**.** *Let $\boldsymbol{\Sigma}\in\mathbb{R}^{n\times n}$ be a symmetric positive-definite matrix representing a covariance in $T_{\boldsymbol{\mu}_s}\mathbb{H}^n_K$, and let $\boldsymbol{\Sigma}' = P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}\,\boldsymbol{\Sigma}\,P^\top_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}$ be the transported covariance. Then:*

1. *$\boldsymbol{\Sigma}'$ is symmetric positive-definite.*
2. *$\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ have identical eigenvalues.*
3. *$\mathrm{tr}(\boldsymbol{\Sigma}') = \mathrm{tr}(\boldsymbol{\Sigma})$ and $\det(\boldsymbol{\Sigma}') = \det(\boldsymbol{\Sigma})$.*

Proof. Let $\boldsymbol{P}\in\mathbb{R}^{n\times n}$ denote the matrix representation of $P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t}$ in an orthonormal basis.

(1) Since P is an isometry (Proposition 8), $\boldsymbol{P}$ is orthogonal: $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_n$. Hence $\boldsymbol{\Sigma}' = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top$ is congruent to $\boldsymbol{\Sigma}$ via an invertible transformation, which preserves positive-definiteness and symmetry.

(2) The eigenvalues of $\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top$ equal those of $\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$, since $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_n$. Equivalently, $\boldsymbol{\Sigma}'$ and $\boldsymbol{\Sigma}$ are unitarily similar via the orthogonal matrix $\boldsymbol{P}$.

(3) Trace and determinant are symmetric functions of the eigenvalues. Since the eigenvalues are preserved, $\mathrm{tr}(\boldsymbol{\Sigma}') = \mathrm{tr}(\boldsymbol{\Sigma})$ and $\det(\boldsymbol{\Sigma}') = \det(\boldsymbol{\Sigma})$. □

B.7 OPTIMALITY OF TANGENT-SPACE AGGREGATION

Lemma 10 (Exactness of the Gaussian Product in Tangent Space)**.** *Under Assumption 3 (conditional independence) and given that the tangent-space metric is exactly Euclidean (Lemma 4), the precision-weighted aggregation in Eqs.* (11)–(12) *yields the unique Gaussian that is proportional to the product of all incoming message Gaussians and the prior.*

Proof. In $T_{\boldsymbol{\mu}_t}\mathbb{H}^n_K \cong \mathbb{R}^n$ (Lemma 4), each transported message is a Gaussian $\mathcal{N}(\boldsymbol{v}_{m_k}, \boldsymbol{\Sigma}_{m_k \to t})$ and the prior is $\mathcal{N}(\boldsymbol{v}_t, \boldsymbol{\Sigma}_t)$. The product of $K+1$ Gaussians is proportional to a Gaussian whose natural parameters (precision and information vector) are the sums of the individual natural parameters:

$$\log \prod_{k=0}^{K} \mathcal{N}(\boldsymbol{v};\, \boldsymbol{v}_k, \boldsymbol{\Sigma}_k) \propto -\frac{1}{2}\boldsymbol{v}^\top \Big(\sum_k \boldsymbol{\Sigma}_k^{-1}\Big)\boldsymbol{v} + \boldsymbol{v}^\top \Big(\sum_k \boldsymbol{\Sigma}_k^{-1}\boldsymbol{v}_k\Big) + \text{const.}$$

Identifying $\boldsymbol{\Lambda}_t = \sum_k \boldsymbol{\Sigma}_k^{-1}$ and $\boldsymbol{\eta}_t = \sum_k \boldsymbol{\Sigma}_k^{-1}\boldsymbol{v}_k$ recovers Eqs. (11)–(12). The resulting Gaussian has mean $\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\eta}_t$ and covariance $\boldsymbol{\Lambda}_t^{-1}$. Under Assumption 2, all precision matrices are bounded, so $\boldsymbol{\Lambda}_t$ is positive-definite and the inverse is well-defined. □

Proposition 11 (Optimality of the Precision-Weighted Mean)**.** *The precision-weighted mean $\hat{\boldsymbol{v}}_t = \boldsymbol{\Lambda}_t^{-1}\boldsymbol{\eta}_t$ is the minimum-variance linear unbiased estimator (MVLUE) of the common mean given K independent Gaussian observations in $T_{\boldsymbol{\mu}_t}\mathbb{H}^n_K$.*

Proof. This is the Gauss–Markov theorem applied to the Euclidean tangent space. Consider the linear model $\boldsymbol{v}_{m_k} = \boldsymbol{v}^* + \boldsymbol{\epsilon}_k$, $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{m_k})$, where $\boldsymbol{v}^*$ is the true mean. The generalised least-squares estimator is $\hat{\boldsymbol{v}} = (\sum_k \boldsymbol{\Sigma}_{m_k}^{-1})^{-1}(\sum_k \boldsymbol{\Sigma}_{m_k}^{-1}\boldsymbol{v}_{m_k})$, which coincides with $\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\eta}_t$ (omitting the prior term without loss of generality). By the Gauss–Markov theorem, this estimator achieves the minimum variance among all linear unbiased estimators. □

B.8 QUADRATURE APPROXIMATION ERROR

Theorem 12 (Gauss–Hermite Quadrature Error Bound)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be $2Q$-times continuously differentiable, and let $I = \int_{\mathbb{R}^n} f(\boldsymbol{v})\,\mathcal{N}(\boldsymbol{v};\, \mathbf{0}, \boldsymbol{\Sigma})\,\mathrm{d}\boldsymbol{v}$. The tensor-product Gauss–Hermite quadrature with Q nodes per dimension satisfies*

$$\left| I - \sum_{q=1}^{Q^n} w_q\, f(\boldsymbol{L}\boldsymbol{z}^{(q)}) \right| \le C_n \cdot \frac{Q!}{(2Q)!} \cdot \sup_{\boldsymbol{v}} \left\| D^{2Q} f(\boldsymbol{v}) \right\| \cdot \|\boldsymbol{\Sigma}\|^Q,$$

where $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$, C_n is a dimension-dependent constant, and $D^{2Q} f$ denotes the tensor of $2Q$-th order partial derivatives. In the CHBP setting, $f(\boldsymbol{v}) = \psi_{st}(\exp_{\boldsymbol{\mu}_s}(\boldsymbol{v}), \boldsymbol{x}_t) = \exp(-\alpha\, d_K^2(\exp_{\boldsymbol{\mu}_s}(\boldsymbol{v}), \boldsymbol{x}_t))$.

Proof sketch. The one-dimensional Gauss–Hermite quadrature of order Q is exact for polynomials of degree $\le 2Q-1$ integrated against the Gaussian weight. The error for a smooth function g is bounded by

$$\left| \int g(z)\, e^{-z^2}\, \mathrm{d}z - \sum_{q=1}^{Q} w_q\, g(z_q) \right| \le \frac{Q!\,\sqrt{\pi}}{(2Q)!}\, \sup_z |g^{(2Q)}(z)|.$$

Extending to n dimensions via the tensor product and applying the substitution $\boldsymbol{v} = \boldsymbol{L}\boldsymbol{z}$ introduces the factor $\|\boldsymbol{\Sigma}\|^Q$. The bound demonstrates two key properties of the CHBP framework: (i) for concentrated beliefs ($\|\boldsymbol{\Sigma}\| \to 0$), the error vanishes polynomially in $\|\boldsymbol{\Sigma}\|^Q$; and (ii) for diffuse beliefs, higher-order quadrature ($Q \uparrow$) is necessary to maintain accuracy, as stated in Section 3. □

B.9 CURVATURE ANNEALING AS HOMOTOPY CONTINUATION

Theorem 13 (Continuity of the Annealing Path)**.** *Define the one-parameter family of Riemannian manifolds* $\{\mathbb{H}^n_{K(\tau)}\}_{\tau\in[0,1]}$ *with* $K(\tau) = K_{\text{tgt}}(1 - e^{-\lambda\tau T})$*. Let* $\Phi_K : \mathbb{R}^n \to \mathbb{H}^n_K$ *denote the CHBP message-passing operator for a single iteration at curvature* K*, mapping the tangent-space statistics of all nodes to updated statistics. Then:*

1. *As* $K(\tau) \to 0$, $\exp^K_{\boldsymbol{o}}(\boldsymbol{v}) \to \boldsymbol{v}$, $P^K_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t} \to \boldsymbol{I}_n$, *and* $|\det J^K_{\exp_{\boldsymbol{o}}}(\boldsymbol{v})| \to 1$ *pointwise.*
2. *The map* $\tau \mapsto \Phi_{K(\tau)}$ *is continuous in the operator norm.*
3. *If CHBP converges to a fixed point* $\boldsymbol{v}^*(\tau)$ *for each* τ*, then* $\tau \mapsto \boldsymbol{v}^*(\tau)$ *is continuous.*

Proof. (1) Set $c(\tau) = \sqrt{|K(\tau)|}$. As $\tau \to 0$, $c(\tau) \to 0$. The exponential map at the origin (Eq. 3) satisfies:

$$\cosh(c\|\boldsymbol{v}\|_L) = 1 + \tfrac{1}{2}c^2\|\boldsymbol{v}\|^2_L + O(c^4),$$
$$\frac{\sinh(c\|\boldsymbol{v}\|_L)}{c\|\boldsymbol{v}\|_L} = 1 + \tfrac{1}{6}c^2\|\boldsymbol{v}\|^2_L + O(c^4).$$

Taking $c \to 0$: $\exp^K_{\boldsymbol{o}}(\boldsymbol{v}) \to (1/\sqrt{|K|})\boldsymbol{e}_0 + \boldsymbol{v}$, which in the tangent-space coordinate chart reduces to $\boldsymbol{v}$ (Euclidean addition). The Jacobian determinant $(\sinh(cr)/(cr))^{n-1} \to 1^{n-1} = 1$.

For parallel transport, the denominator $1 + \langle\boldsymbol{\mu}_s, \boldsymbol{\mu}_t\rangle_L$ in Eq. (10): when $K \to 0$, $\langle\boldsymbol{\mu}_s, \boldsymbol{\mu}_t\rangle_L \to 1/K \to -\infty$, but all points converge to the common origin, so $\boldsymbol{\mu}_s \to \boldsymbol{\mu}_t \to \boldsymbol{o}$, and $P_{\boldsymbol{\mu}_s\to\boldsymbol{\mu}_t} \to \boldsymbol{I}_n$ by l'Hôpital's rule on the correction term.

(2) All constituent operations of Φ_K — namely $\exp$, $\log$, P, d_K, and the Jacobian — are smooth functions of K for $K \le 0$. Since compositions and finite sums of continuous operators are continuous, $\tau \mapsto \Phi_{K(\tau)}$ is continuous.

(3) If $\boldsymbol{v}^*(\tau)$ satisfies $\Phi_{K(\tau)}(\boldsymbol{v}^*(\tau)) = \boldsymbol{v}^*(\tau)$ and Φ is a contraction mapping (which holds for sufficiently small spectral radius of the message Jacobian), then by the implicit function theorem for contractions, $\boldsymbol{v}^*(\tau)$ varies continuously with τ. □

Lemma 14 (Convergence on Trees)**.** *If* $\mathcal{G}$ *is a tree, then CHBP with fixed curvature* $K < 0$ *converges to the exact marginals of the pairwise MRF (in the tangent-space Gaussian parametrisation) in at most* $\operatorname{diam}(\mathcal{G})$ *message-passing iterations, where* $\operatorname{diam}(\mathcal{G})$ *denotes the graph diameter.*

Proof. On a tree, BP computes exact marginals [Pearl, 1988]. Each message $m_{s\to t}$ depends only on the messages arriving at s from $\mathcal{N}(s) \setminus t$. Since the graph is acyclic, after at most $\operatorname{diam}(\mathcal{G})$ rounds of parallel updates, every message has received contributions from every node in the graph. By Assumption 3, there is no over-counting, so the beliefs at convergence are the exact posterior marginals.

The CHBP update preserves this structure: the tangent-space aggregation is exact for Gaussians (Lemma 10), the covariance transport is lossless (Theorem 9), and the density normalisation is exact (Theorem 6). The only approximation is the Gauss–Hermite quadrature in the message computation, which introduces an error bounded by Theorem 12. For exact inference, one would require $Q \to \infty$ or, equivalently, transition potentials that are polynomial in the tangent-space coordinates. □

C EXPERIMENTAL DETAILS

This appendix provides the full specification of datasets, evaluation metrics, baselines, and implementation details summarised in Section 4.

C.1 DATASETS

We curated three categories of evaluation graphs:

Synthetic graphical models. Complete k-ary trees ($k \in \{2, 3, 5, 10\}$, depth up to 8), Galton-Watson branching processes (mean degree 3 and 5), caterpillar trees, Barabási-Albert preferential attachment networks ($N \in \{500, 2000\}$), Watts-Strogatz small-world graphs, LFR benchmark networks, and stochastic block models with three-level hierarchical community structure. These permit exact control over generative parameters and ground-truth distributions.

Non-hierarchical negative controls. 2D lattice grids (20×20), 3D lattice grids ($10\times10\times10$), and Erdős–Rényi random graphs ($N=500$) serve as explicit boundary conditions where hyperbolic geometry provides no advantage.

Standard inference benchmarks. Instances from the UAI Inference Competitions (2006, 2008, 2014) and canonical 2D Spin Glass models.

Real-world hierarchical graphs. WordNet noun and verb hierarchies, the Disease Ontology, the Phylogenetic Tree database, the Math Genealogy graph, Amazon product co-purchase networks, and the Cora and PubMed citation graphs. All real-world datasets were partitioned into 60/20/20 train/validation/test splits, with continuous node features normalised to zero mean and unit variance.

C.2 EVALUATION METRICS

For synthetic and UAI benchmark graphs where exact inference (Junction Tree or exhaustive enumeration) is tractable, we report the Kullback-Leibler (KL) divergence between approximate beliefs and exact ground-truth marginals. For semi-supervised real-world tasks, we report Top-1 Accuracy, Macro-F1 Score, and Negative Log-Likelihood (NLL). We independently compute the Gromov δ-hyperbolicity for all graphs prior to inference as a structural diagnostic.

C.3 BASELINES

Euclidean Continuous BP (Euc-BP). The identical continuous message-passing algorithm operating in $\mathbb{R}^d$, evaluated at $d=5$ and $d=50$.

Classical approximate inference. Discrete Loopy Belief Propagation (LBP) [Yedidia et al., 2003], Expectation Propagation (EP) [Minka, 2001], and Tree-Reweighted BP (TRW-BP) [Wainwright et al., 2003].

Geometric baselines. Mean-Field Variational Inference (MFVI) and Wrapped Expectation Propagation (Wrapped-EP), which performs moment matching using Wrapped Normal distributions on $\mathbb{H}^n_K$.

Neural baselines. Graph Convolutional Networks (GCN) [Kipf and Welling, 2016] and Hyperbolic Graph Convolutional Networks (HGCN) [Chami et al., 2019], representing discriminative approaches with Euclidean and hyperbolic latent spaces respectively.

C.4 IMPLEMENTATION DETAILS

All experiments were conducted on a cluster of eight NVIDIA H100 GPUs. Beliefs were initialised at the Lorentz origin with isotropic covariance $\sigma_0^2 = 1.0$. Message-passing updates used Adam with learning rate 10^{-3} and weight decay 10^{-4}. Curvature was annealed from $K_0 = -10^{-4}$ to $K_{\text{target}} = -1.0$ at rate $\lambda = 0.05$. Gauss–Hermite quadrature used $Q = 5$ points per dimension. All CHBP models operated at $d=5$; Euclidean models were evaluated at both $d=5$ and $d=50$.

D ROBUSTNESS AND OVERFITTING ANALYSIS

A natural question is why CHBP exhibits strong generalisation without the explicit regularisation mechanisms (dropout, batch normalisation, early stopping on validation loss) that discriminative baselines such as GCN and HGCN require. We identify four structural properties of the framework that collectively provide implicit regularisation and robustness.

D.1 GEOMETRIC INDUCTIVE BIAS AS CAPACITY CONTROL

CHBP operates at a fixed manifold dimension $n = 5$, yielding $\mathcal{O}(n^2)$ covariance parameters per node—far fewer than the hidden-layer weights in neural baselines. More fundamentally, the constant negative curvature K imposes a rigid geometric prior: the exponential volume growth of $\mathbb{H}^n_K$ matches hierarchical branching but cannot accommodate arbitrary data distributions. On hierarchical graphs this prior is well-specified, concentrating the hypothesis space around the true

posterior and reducing the effective model capacity. On flat topologies the prior is misspecified and performance degrades, as observed in our experiments—a behaviour consistent with regularisation rather than overfitting.

D.2 EXPLICIT COVARIANCE TRACKING AS BAYESIAN REGULARISATION

Unlike discriminative models that produce point embeddings, CHBP maintains a full covariance matrix $\mathbf{\Sigma}_i$ at every node. This explicit uncertainty tracking acts as a form of Bayesian regularisation: nodes with limited evidence retain broad beliefs (large $\|\mathbf{\Sigma}_i\|$), which naturally temper their influence on downstream predictions. The precision-weighted aggregation (Eqs. 11–12) ensures that high-uncertainty messages contribute less to belief updates, preventing any single noisy observation from dominating the posterior. This mechanism explains the consistently lower NLL of CHBP compared to HGCN: the covariance prevents the overconfident predictions that arise when structural variance is compressed into deterministic embeddings.

D.3 CURVATURE ANNEALING AS IMPLICIT REGULARISATION

The curvature annealing schedule $K^{(t)} = K_{\text{target}} \cdot (1 - e^{-\lambda t})$ introduces an implicit regularisation path analogous to the information bottleneck in deep learning. In early iterations ($K^{(t)} \approx 0$), the near-flat geometry permits only coarse, low-frequency structural organisation; high-frequency, potentially spurious patterns cannot be resolved because the manifold lacks sufficient curvature to separate closely related branches. As the curvature increases, finer structural distinctions become expressible. This coarse-to-fine progression mirrors curriculum learning and prevents the algorithm from fitting noise in the early, data-scarce phase of inference.

Formally, the annealing defines a continuous homotopy from Euclidean BP (at $K = 0$) to the full hyperbolic model (at $K = K_{\text{target}}$). Theorem 13 guarantees that the fixed points vary continuously along this path, ensuring that the solution at the target curvature is a smooth deformation of the Euclidean solution rather than an arbitrary local optimum.

D.4 DETERMINISTIC QUADRATURE VS. STOCHASTIC SAMPLING

CHBP approximates message integrals via Gauss–Hermite quadrature with deterministic sigma points, not Monte Carlo sampling. This eliminates the gradient variance that plagues sampling-based inference methods and contributes to the stability of the iterative updates. The deterministic nature of the quadrature means that, given identical inputs, the algorithm produces identical outputs across runs, removing a source of optimisation noise that can drive overfitting in stochastic methods.

D.5 ROBUSTNESS TO INITIALISATION AND GRAPH PERTURBATIONS

The cold-start protocol (all nodes at the Lorentz origin) combined with curvature annealing provides robustness to initialisation: because every run begins from the same uninformative state and follows a deterministic annealing path, the algorithm is not sensitive to random seed selection. We verified this empirically by running 10 independent trials on each dataset, observing standard deviations below 0.3% in accuracy across all hierarchical graphs.

To assess robustness to graph perturbations, we randomly rewired 5%, 10%, and 20% of edges in the WordNet noun hierarchy and re-ran inference. CHBP accuracy degraded gracefully from 83.55% to 81.2%, 78.9%, and 73.4% respectively, whereas Euclidean BP (d=50) dropped from 62.45% to 58.1%, 52.3%, and 44.8%. The relative degradation of CHBP was consistently smaller, suggesting that the hyperbolic geometry provides a more robust structural scaffold that degrades gracefully under local perturbations.