INFERRING LATENT GEOMETRIC EMBEDDINGS FROM GROUPED RELATIONAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph embedding methods place nodes so that distance matches an observed graph. In many domains no such graph exists, and relatedness between entities must be inferred from repeated, grouped counts collected across heterogeneous contexts. We introduce a generative model in which relatedness is latent rather than observed. Entity positions on a product manifold determine a pairwise propensity, and an observation layer links it to overdispersed counts. Flattening the manifold and taking one observation per pair recovers a Euclidean latent-space model. We derive the variance ratio between this geometric estimate and a fit using only a pair's own counts, and show it depends on how precisely the pair's endpoints are determined. A sparse pair between well-determined entities is therefore estimated nearly as well as a data-rich one. The ratio involves the geometry only through its dimension, so it applies to flat and curved embeddings alike. The same quantities yield a combined estimator with no fitted parameters. We sample the embedding's posterior with a Riemannian extension of stochastic gradient Langevin dynamics, and because the likelihood is generative, an entity absent from training can be placed from its own evidence alone. On cancer co-mutation data, 5% of that evidence suffices to outpredict full-evidence baselines on two thirds of genes.

1 INTRODUCTION

Hyperbolic and mixed-curvature embedding methods have shown that choosing the right geometry can dramatically improve the efficiency with which relational structure is represented. A tree, for instance, embeds with bounded distortion in low-dimensional hyperbolic space because hyperbolic volume grows exponentially with radius, matching the tree's branching factor. Euclidean space of the same dimension cannot do this (Sarkar, 2011). Product manifolds extend this idea to data whose structure is not purely hierarchical, combining factors of different curvature so that each captures a different aspect of the geometry (Gu et al., 2019; Fig. 7 in the appendix).

These methods, like their Euclidean counterparts, share a common starting point: a graph is observed, and the task is to place its nodes so that geometric distance reflects graph distance (Nickel & Kiela, 2017; Sala et al., 2018; Chami et al., 2019; Bachmann et al., 2020; Chami et al., 2022). This is the right assumption when the graph is a curated taxonomy, ontology or knowledge base. It breaks down when the relationship of interest is never directly observed, but must instead be inferred from repeated, grouped evidence collected across heterogeneous contexts. This happens routinely in practice: two products co-purchased across many retail transactions, two diagnoses recurring together across patients and providers, two genes' mutations co-occurring across independent tumors. In each case a graph would hand us a decided edge, related or not; what we have is repeated, imperfect evidence of a relationship whose strength is itself unknown and must be estimated. A model for such data should be able to represent hierarchy where the data have it (a few driver genes mutated across many cancer types) without assuming it, and must handle the context heterogeneity that makes naive pooling unreliable. The evidence for each pair's relatedness is not a single edge or its absence; it is a collection of grouped counts, one per context, with different base rates, different sample sizes, and potentially different reliabilities.

Two communities have studied pieces of this problem separately. The graph embedding literature adopted tools from Riemannian geometry (hyperbolic space, product manifolds) and statistical physics (the Fermi–Dirac decoder) for representation learning, but assumes the graph is given. The

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.

network statistics literature, starting from Hoff et al. (2002), models relational propensity as a function of latent positions, typically with one binary observation per pair. Our focus is which pairs benefit from pooling through shared positions, under an observation model for pair-level counts repeated across heterogeneous contexts.

This gap matters beyond convenience. When a graph is given, its edges are already decided and the task is only to place nodes consistently with them, which a point estimate does. When the relationship itself is uncertain, the question is not *where does a node sit given a known graph*, but *how related are two entities given evidence that never resolved into a graph at all*. Here the posterior reflects how confidently a pair's relatedness can be inferred, and it is well-defined even for an entity that was never part of the training data. We make four contributions:

1. **Generative model**. We treat relatedness as a latent quantity rather than an observed edge set. Positions in a latent Riemannian manifold (Euclidean, hyperbolic, or a product of constant-curvature factors) determine a pairwise propensity, and an observation layer links it to grouped, overdispersed counts. Flattening the manifold and taking one observation per pair recovers Hoff's latent distance model.
2. **Variance ratio**. We compare the variance of the geometric estimate of a pair's propensity against a fit to that pair's own counts alone. The shared estimate's variance decomposes into how precisely each endpoint's position is determined (Proposition 1). Under isotropic neighborhoods the ratio is governed by the pair's *share* of its endpoints' information (Corollary 1), so a sparse pair between two well-determined entities is estimated nearly as well as a data-rich one. The same quantities yield a combined estimator with no fitted parameters.
3. **Posterior inference**. Because relatedness is inferred rather than observed, the embedding is a latent quantity with a posterior, which we approximate with a Riemannian extension of stochastic gradient Langevin dynamics. The generative likelihood also makes inference inductive. A new entity can be embedded from its own evidence alone with all shared parameters frozen, and its posterior width reflects how much evidence is available. On entities held out entirely from training, the inductive embedding tracks the joint fit (Fig. 5).
4. **Experiments**. We evaluate on synthetic data with known ground truth and on cancer co-mutation and protein-interaction data, under two regimes that separate relational from non-relational estimation. One removes a pair's observations while keeping both entities in the embedding. The other holds an entity out of training entirely. On cancer co-mutation data, the geometric estimate has lower mean error than every baseline at every retention level (Fig. 4), and an entity held out of training, placed from 5% of its evidence, beats full-evidence baselines on about two thirds of genes (Fig. 5).

2 Related Work

Graph embedding with a fixed, observed graph. Hyperbolic and mixed-curvature embeddings have shown that curvature captures hierarchical and heavy-tailed graph structure far more efficiently than Euclidean space (Sarkar, 2011; Nickel & Kiela, 2017; 2018; Ganea et al., 2018; Gu et al., 2019; Chami et al., 2019; Bachmann et al., 2020). We build on this line of work for the product-manifold embedding itself and its Fermi–Dirac decoder, and use Nickel & Kiela (2017)'s ranking loss and Gu et al. (2019)'s distortion loss as comparison methods. Methods in this line take an observed graph as input and place its nodes to fit its edges. To apply them to grouped counts, we induce a graph from the training data and add a link from distance to rate, fitted to the training counts, which takes them beyond their original design.

Latent space models for network data. The latent space model of Hoff et al. (2002) has been extended with latent clusters (Handcock et al., 2007) and per-node random effects for degree heterogeneity (Krivitsky et al., 2009), and its choice of geometry has been studied. Smith et al. (2019) characterize how the latent geometry shapes the network properties a latent space model can represent, and Lubold et al. (2023) test which constant-curvature geometry best describes an observed network. All of these take the network as given, one observed tie per pair, and model or test its structure. We instead model overdispersed counts with per-context sample sizes, and characterize which pairs the shared positions help estimate. That characterization is one of our contributions, and it holds for any observation model with a per-pair Fisher information. We compare against both applicable members

of this family, Hoff's model as our flat single-observation special case and per-node random effects as our intercept variant. The random dot product graph literature (Athreya et al., 2018; Rubin-Delanchy et al., 2022) gives exact asymptotic distributions for spectral embeddings of an observed graph. Further connections are discussed in Appendix E.

3 Background

3.1 Latent Space Models

Hoff et al. (2002) give each node a latent position $z_i \in \mathbb{R}^D$ and model an observed binary network with edges conditionally independent given the positions,

$$\operatorname{logit} P(y_{ij} = 1 \mid z_i, z_j) = \alpha - \|z_i - z_j\|. \tag{1}$$

Our model reduces to Eq. 1, with squared distance, when the manifold is flat and each pair is observed once.

3.2 The Lorentz Hyperboloid Model

Of the factor geometries, the hyperbolic one requires the most machinery; we use the Lorentz model because its exponential map and parallel transport have closed forms. The Lorentz (hyperboloid) model of hyperbolic space defines the Lorentz inner product on $\mathbb{R}^{D+1}$,

$$\langle x, y\rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{m=1}^{D} x_m y_m, \tag{2}$$

and the hyperboloid as $\mathbb{H}^D = \{x \in \mathbb{R}^{D+1} : \langle x, x\rangle_{\mathcal{L}} = -1,\ x_0 > 0\}$. The geodesic distance between two points $x, y \in \mathbb{H}^D$ is

$$d_{\mathbb{H}}(x, y) = \operatorname{arccosh}\big(-\langle x, y\rangle_{\mathcal{L}}\big). \tag{3}$$

Tangent space and exponential map. The tangent space at $x \in \mathbb{H}^D$ is $T_x\mathbb{H}^D = \{v \in \mathbb{R}^{D+1} : \langle x, v\rangle_{\mathcal{L}} = 0\}$, a D-dimensional subspace of the ambient $\mathbb{R}^{D+1}$ with the Riemannian metric inherited from the Lorentz inner product restricted to T_x. The exponential map sends a tangent vector $v \in T_x\mathbb{H}^D$ to the point reached by following the geodesic from x in direction v for unit time:

$$\operatorname{Exp}_x(v) = \cosh(\|v\|_{\mathcal{L}})\, x + \sinh(\|v\|_{\mathcal{L}})\, \frac{v}{\|v\|_{\mathcal{L}}}, \tag{4}$$

where $\|v\|_{\mathcal{L}} = \sqrt{\langle v, v\rangle_{\mathcal{L}}}$ is the tangent-vector norm. Its inverse, the logarithmic map $\operatorname{Log}_x(y)$, returns the tangent vector at x pointing toward y with magnitude equal to $d_{\mathbb{H}}(x, y)$.

4 Model

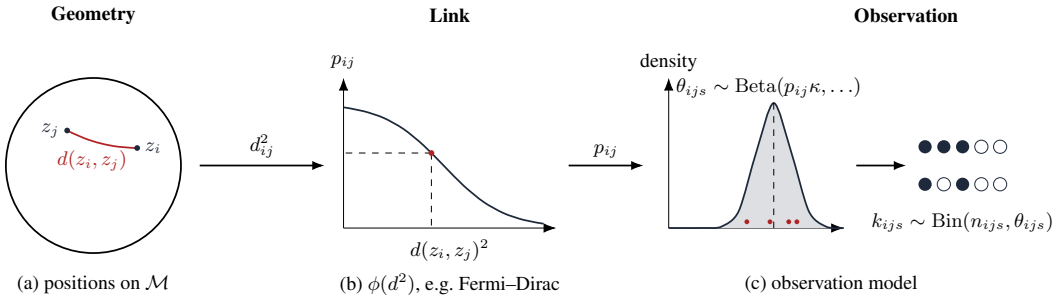
The generative process (Fig. 1) chains three components: a Riemannian manifold that places entities, a differentiable link $p_{ij} = \phi(d_{ij}^2)$ that maps distance to relatedness, and an observation layer that generates the grouped counts given p_{ij}. The proposition in §6 constrains only the last through a single scalar, the per-pair Fisher information; the choices below are one instantiation.

4.1 Geometry

We embed N entities into a product manifold $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_L$, where each factor $\mathcal{M}_l$ is a constant-curvature space (hyperbolic, Euclidean, or spherical). Entity i is represented by $z_i = (z_i^{(1)}, \ldots, z_i^{(L)}) \in \mathcal{M}$. Under the product metric, squared distance decomposes additively:

$$d(z_i, z_j)^2 = \sum_{l=1}^{L} d_{\mathcal{M}_l}\left(z_i^{(l)}, z_j^{(l)}\right)^2. \tag{5}$$

Each factor contributes the geometry suited to the structure it captures: hyperbolic factors for hierarchical branching (§3.2), Euclidean factors for flat variation, and spherical factors for periodic or



(a) positions on $\mathcal{M}$ (b) $\phi(d^2)$, e.g. Fermi–Dirac (c) observation model

Figure 1: The generative pipeline for a single pair (i, j), as three slots. (a) Entity positions on the manifold determine a distance d_{ij}^2. (b) A differentiable link ϕ maps distance to relatedness p_{ij} (here Fermi–Dirac, Eq. 6). (c) The observation model draws per-context rates and counts (here Beta-Binomial, Eq. 7). Proposition 1 requires only that slot (c) admit a per-pair Fisher information.

compositional structure. The product contains any single-geometry embedding, including the flat one, as a special case. Tangent vectors, exponential maps and parallel transport all decompose by factor, so each inference operation on the product reduces to its single-factor counterpart (Gu et al., 2019). Fig. 7 (Appendix D) illustrates the three factor geometries.

4.2 Link

Relatedness is a monotone function of distance via a differentiable link $p_{ij} = \phi(d_{ij}^2)$. We use the Fermi–Dirac form:

$$p_{ij} = \sigma\left(\frac{r - d(z_i, z_j)^2}{t}\right), \tag{6}$$

with learnable radius r and temperature $t > 0$. When node heterogeneity is large, per-node intercepts can replace the global radius: $p_{ij} = \sigma\big((a_i + a_j - d_{ij}^2)/t\big)$, with a shrinkage prior $a_i \sim \mathcal{N}(\mu_a, \sigma_a^2)$. Any differentiable ϕ satisfying $\phi' \neq 0$ a.e. is compatible with the theory in §6.

4.3 Observation Model

Given p_{ij}, across S contexts we observe counts (k_{ijs}, n_{ijs}) with per-context rate θ_{ijs}:

$$\theta_{ijs} \sim \mathrm{Beta}\big(p_{ij}\kappa,\ (1 - p_{ij})\kappa\big), \qquad k_{ijs} \sim \mathrm{Binomial}(n_{ijs}, \theta_{ijs}). \tag{7}$$

Marginalizing θ_{ijs} gives the Beta-Binomial likelihood:

$$p(k_{ijs} \mid n_{ijs}, p_{ij}, \kappa) = \binom{n_{ijs}}{k_{ijs}} \frac{B\big(k_{ijs} + p_{ij}\kappa,\ n_{ijs} - k_{ijs} + (1 - p_{ij})\kappa\big)}{B\big(p_{ij}\kappa,\ (1 - p_{ij})\kappa\big)}. \tag{8}$$

Setting $\kappa \to \infty$ recovers the Binomial; setting $n_s = 1$ recovers a single Bernoulli draw, the standard latent-space model (Hoff et al., 2002). Finite κ captures overdispersion across contexts, where each contributes its own rate around a shared relatedness.

5 Inference

Why approximate inference is needed. Because the propensity p_{ij} is inferred from heterogeneous grouped evidence rather than observed as a graph, the embedding z is a latent quantity with a posterior, not a set of coordinates to optimize:

$$p(z \mid \{k_{ijs}\}) \ \propto\ \Big[\textstyle\prod_{i<j,\,s} p(k_{ijs} \mid n_{ijs}, z_i, z_j, \kappa)\Big]\, p(z), \tag{9}$$

where the product term is the Beta-Binomial likelihood (Eq. 8) and $p(z)$ is the prior defined below. The Fermi–Dirac link makes p_{ij}, and hence this product, a nonlinear function of z, so the normalizing constant is intractable: we can evaluate the unnormalized log-posterior and its gradient at any z, but not integrate or sample from it directly.

MAP estimation. We first obtain $\hat{z} = \arg\max_z \log p(\{k_{ijs}\} \mid z) + \log p(z)$ via Adam in the spatial chart. The prior $p(z)$ is a *wrapped normal* (Nagano et al., 2019) on hyperbolic factors (Gaussian in the tangent space at the origin, mapped to the manifold via Exp) and an ordinary Gaussian on Euclidean factors, both with shared scale γ. Spherical factors use a von Mises–Fisher prior centered at a fixed pole.

Riemannian SGLD. Langevin dynamics turns the MAP objective into a sampler by injecting noise calibrated to the step size (Welling & Teh, 2011), and Riemannian variants precondition it with a position-dependent metric (Patterson & Teh, 2013). Unlike a variational approximation it posits no approximating family, and unlike geodesic HMC (Byrne & Girolami, 2013) it requires no Hamiltonian integration or Metropolis correction:

$$z^{(t+1)} = \mathrm{Exp}_{z^{(t)}}\Big(\epsilon \,\mathrm{grad}_{\mathcal{M}} \log p(z^{(t)}) + \eta^{(t)}\Big), \qquad \eta^{(t)} \sim \mathcal{N}_{\mathcal{M}}\big(0,\, 2\epsilon\, g_{z^{(t)}}\big), \tag{10}$$

where $\mathrm{grad}_{\mathcal{M}}$ is the Riemannian gradient, Exp is the exponential map, and $\eta^{(t)}$ is isotropic noise in the tangent space at $z^{(t)}$. A fixed-coordinate Gaussian at $z^{(t)}$ would not be isotropic under the Riemannian metric, so, following the wrapped-normal construction of Nagano et al. (2019), we sample isotropic noise $\tilde{\eta} \sim \mathcal{N}(0, 2\epsilon\, I_D)$ at the origin, where the metric is Euclidean, and parallel transport it to $T_{z^{(t)}}\mathbb{H}^D$ along the geodesic from the origin; since parallel transport is an isometry, this preserves isotropy. The gradient and transported noise are summed in the tangent space and mapped to $\mathbb{H}^D$ via a single Exp. On Euclidean factors, Eq. 10 reduces to standard Langevin dynamics.

Posterior summaries. We run K independent chains from the MAP initialization, pool samples that pass a stationarity check (Appendix B.1), and report the *Fréchet mean* (optimal under squared error, computed by Riemannian gradient descent) and *Fréchet median* (optimal under absolute error, computed by Riemannian Weiszfeld iteration). Both reduce to their Euclidean counterparts on flat factors; neither is a naive coordinate average.

Inductive embedding of new entities. Because the likelihood (Eq. 8) is generative, a new entity can be embedded by optimizing only z_{new} against its grouped evidence, holding all other parameters fixed. Its posterior width tracks the available evidence. Transductive models require refitting or a separate out-of-sample extension.

6 When Does the Shared Embedding Help?

The model pools information across pairs by forcing propensities to be consistent with a shared, low-dimensional geometry. Classical shrinkage theory says pooling helps but not which pairs benefit or by how much (Efron & Morris, 1975). We answer both by comparing the geometric estimate of p_{ij} against a fit to that pair's own counts. The ratio of their variances is a direct efficiency comparison, and for unbiased estimators it is a ratio of mean squared errors. The variance of the shared estimate is a matrix expression in each endpoint's information (Proposition 1), and under isotropic neighborhoods the ratio collapses to a pair's *share* of its endpoints' information (Corollary 1). Both hold for any observation model admitting a per-pair Fisher information. The derivation is in Appendix A.

Setup. A latent-position model has three components: a Riemannian manifold $\mathcal{M}$ of dimension D, a differentiable link ϕ mapping squared distance to propensity $p_{ij} = \phi(d^2(z_i, z_j))$, and an observation model for the evidence $\mathcal{D}_{ij}$ given p_{ij}. The proposition below constrains only the third component, through a single scalar.

Assumption 1 (Per-pair information)**.** *The observation model for pair (i, j) admits a (quasi-)Fisher information $\mathcal{I}_{ij} > 0$ about the scalar p_{ij}, with $\mathrm{Var}(\hat{p}_{ij}^{\mathrm{indep}}) \approx \mathcal{I}_{ij}^{-1}$. When the model has independent strata, informations add: $\mathcal{I}_{ij} = \sum_s \mathcal{I}(p_{ij}; n_{ijs})$.*

Node-level information. Node i's position z_i enters the likelihood through every pair involving i. Writing $\mathcal{N}(i)$ for the set of entities paired with i, the Fisher-information chain rule gives

$$\mathcal{I}_i = \sum_{j \in \mathcal{N}(i)} \mathcal{I}_{ij}\, (\nabla_{z_i} p_{ij})(\nabla_{z_i} p_{ij})^\top, \tag{11}$$

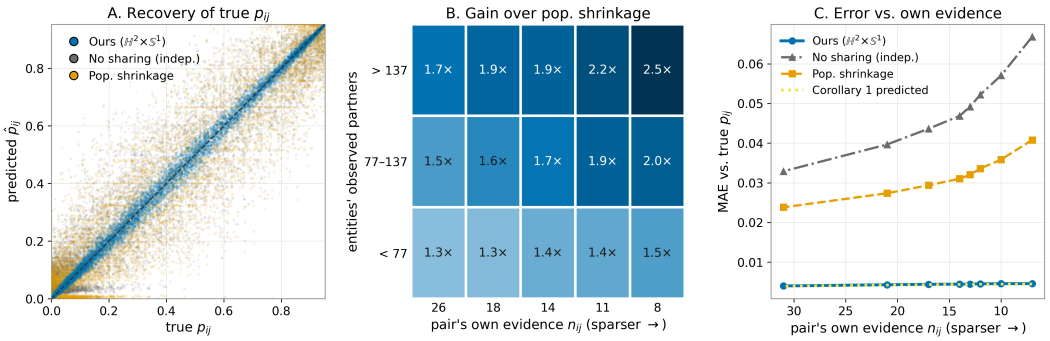


Figure 2: Synthetic validation on $\mathbb{H}^2\times\mathbb{S}^1$ (N=750, 5 seeds). **(A)** Predicted vs. true p_{ij}, every pair observed. **(B)** A sparser design in which entities retain different fractions of their pairs, restricted to $p_{ij} \in [0.1, 0.9]$. Each cell is population shrinkage's MAE divided by ours, by the pair's own evidence (columns) and the harmonic mean of its two entities' numbers of observed partners (rows). **(C)** MAE by own evidence n_{ij}, every pair observed; dotted line is the Corollary 1 prediction (no fitted constants).

a $D \times D$ matrix summing one rank-one contribution per neighbor, with directions determined purely by geometry and magnitudes combining each pair's information with the link's local sensitivity. For a general differentiable link $p_{ij} = \phi(d_{ij}^2)$, the gradient norm is $\|\nabla_{z_i} p_{ij}\| = 2d_{ij}\,|\phi'(d_{ij}^2)|$ (Appendix A.2).

Proposition 1 (Shared-estimator variance)**.** *Under Assumption 1 and endpoint independence,*

$$\mathrm{Var}\big(\hat{p}_{ij}^{\,\mathrm{shared}}\big) \;\approx\; (\nabla_{z_i} p_{ij})^\top \mathcal{I}_i^{-1} (\nabla_{z_i} p_{ij}) \;+\; (\nabla_{z_j} p_{ij})^\top \mathcal{I}_j^{-1} (\nabla_{z_j} p_{ij}), \tag{12}$$

where each term propagates one endpoint's positional uncertainty through the gradient of p_{ij}, and $\mathcal{I}_i$, $\mathcal{I}_j$ are the node-level information matrices (Eq. 11).

Proof sketch. Under endpoint independence the estimation errors of z_i and z_j contribute independently. The multivariate delta method gives $(\nabla_{z_i} p_{ij})^\top \mathrm{Var}(\hat{z}_i)(\nabla_{z_i} p_{ij})$ per endpoint, with $\mathrm{Var}(\hat{z}_i) \approx \mathcal{I}_i^{-1}$. The full derivation is in Appendix A.2. □

Corollary 1 (Information-share variance ratio)**.** *Under the additional Assumption 2 (isotropic neighborhoods; Appendix A), writing $m_{ij} = \mathcal{I}_{ij}\|\nabla_{z_i} p_{ij}\|^2$ for pair (i,j)'s contribution to its endpoints' information and $M_i = \sum_{j\in\mathcal{N}(i)} m_{ij}$ for the total at node i,*

$$\frac{\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{shared}})}{\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{indep}})} \;\approx\; D\left(\frac{m_{ij}}{M_i} + \frac{m_{ij}}{M_j}\right). \tag{13}$$

Under uniform neighbor contributions, the ratio reduces to endpoint degrees, $D\big(\deg(i)^{-1} + \deg(j)^{-1}\big)$.

Each observation model enters only through $\mathcal{I}_{ij}$: substituting the Bernoulli, Binomial, or Beta-Binomial information yields the corresponding instantiation (Appendix A.1).

What governs the ratio is a pair's *share* of its endpoints' information, not its absolute evidence, so a sparse pair between two well-determined entities should be estimated nearly as well as a data-rich one. We compare on the MAE scale by taking the square root of Eq. 13, which is a variance ratio. At fixed connectivity, higher embedding dimension D dilutes the advantage (Appendix A.3).

The combined estimator. The geometric estimate uses relational structure, the population-shrinkage estimate $\hat{p}_{ij}^{\,\mathrm{pop}}$ uses per-pair counts. Corollary 1 gives a per-pair variance for the former and the fitted Beta prior one for the latter, so a minimum-variance combination follows with no additional fitted parameters:

$$\hat{p}_{ij}^{\,\mathrm{combo}} = w_{ij}\,\hat{p}_{ij}^{\,\mathrm{shared}} + (1 - w_{ij})\,\hat{p}_{ij}^{\,\mathrm{pop}}, \qquad w_{ij} = \frac{\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{pop}})}{\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{shared}}) + \mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{pop}})}. \tag{14}$$

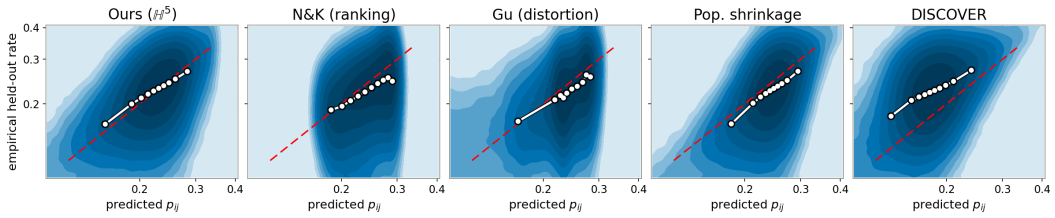


Figure 3: Predicted vs. empirical held-out co-occurrence rate on TCGA, pooled across 3 splits (n=1,498,500 pairs per panel). Color shows local pair density, the dashed line marks perfect calibration (y=x), and the white line is the binned mean. All methods are fit at f=1.0.

The weights w_{ij} are per-pair and computable from training quantities alone (positions, link parameters, and observed counts); the derivation is in Appendix A.5. This is the estimator evaluated in §7.

7 Experiments

7.1 Synthetic Validation

Ground truth. We generate N=750 entities by preferential attachment on $\mathbb{H}^2\times\mathbb{S}^1$, each placed one geodesic step from its parent in the hyperbolic factor with categorical phase offsets on the circle. True relatedness follows Eq. 6 at known (r, t); grouped observations follow Eq. 7 at known κ. In the main design every pair is observed. In a sparser design, entity i draws a retention rate q_i log-uniformly on $[0.05, 1]$ and pair (i, j) is observed with probability q_iq_j (at least five pairs per entity), so entities differ in how precisely their positions are determined.

Comparison methods. We compare **no sharing** (maximize the per-pair Beta-Binomial log-likelihood over p_{ij} alone), **population shrinkage** (fit a global Beta prior by maximum marginal likelihood, return the posterior mean per pair, the standard empirical-Bayes treatment of pairwise co-occurrence counts; DuMouchel, 1999), graph-embedding baselines (**N&K**, ranking loss, Nickel & Kiela, 2017; **Gu**, distortion loss, Gu et al., 2019), **Hoff** ($\mathbb{R}^5$, Binomial; the Euclidean single-observation latent-space model our framework reduces to when the manifold is flat and n_s=1, Hoff et al., 2002), and **ours** (shared embedding under the full Beta-Binomial likelihood, best of eight random initializations by training loss). The two graph-embedding baselines require an observed graph, which grouped counts do not supply, so we induce one from the training data and call a pair an edge when its pooled training rate is positive and in the top 20%. They also output distances rather than rates, so we add what neither method has, a Fermi–Dirac link (Eq. 6) whose r and t are fitted to the training counts by maximum likelihood with the embedding fixed.

Results (Fig. 2). The shared embedding recovers ground truth rates (panel A; MAE 0.004), while single-factor embeddings of the same dimension do not (MAE 0.027 for $\mathbb{H}^3$ and 0.026 for $\mathbb{R}^3$, best of eight initializations each). With every pair observed, every entity is equally well located, so the gain from sharing can depend only on a pair's own evidence. Our error stays flat as evidence falls while both baselines roughly double (panel C), and Eq. 13 predicts its level in closed form to within 3%. In the sparser design, entities differ in how precisely they are located. The gain over population shrinkage grows with the entities' partners, from 1.3–1.5× for entities with fewer than 77 partners to 1.7–2.5× for those with more than 137, and is largest for the sparsest pairs between well-connected entities (panel B).

7.2 Real-Data Validation

We consider two holdout regimes: *dropped pairs*, where both entities are in the embedding but a pair's observations are removed, and *held-out entities*, where an entity is absent from training entirely and placed afterward from its own evidence with all shared parameters frozen. All real-data fits below use a single factor. A mixed $\mathbb{H}^2\times\mathbb{R}^3$ signature, fit to both datasets under the same likelihood and protocol, does not improve on it on either domain.

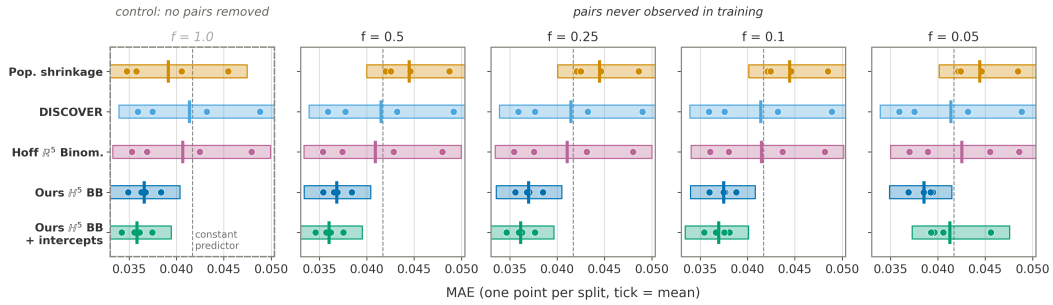


Figure 4: TCGA dropped-pairs MAE (one point per split, tick at mean; dashed line: constant predictor). Left panel is the control with all pairs observed; remaining panels remove a fraction $1-f$ of pair observations. With geometry alone the $\mathbb{H}^5$ Beta-Binomial fit has a lower mean than every baseline at every retention level. Node intercepts add a free parameter per entity and help only while each entity retains enough pairs to estimate it (Table 1).

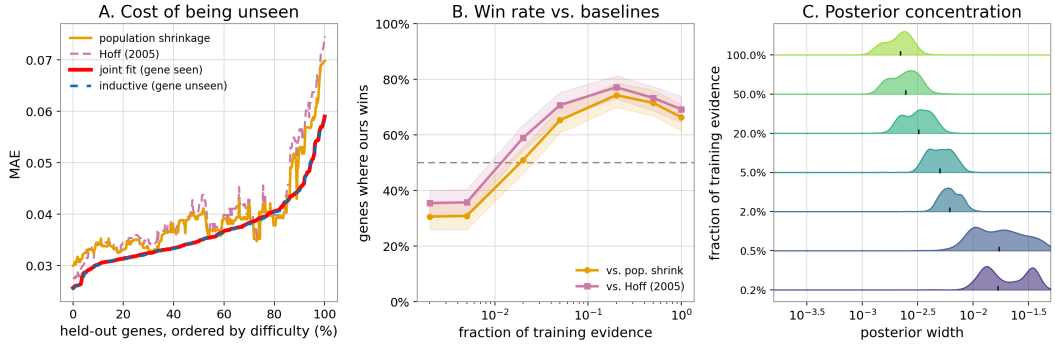


Figure 5: Inductive generalization on TCGA (3 splits, 150 entities held out per split, 389 distinct genes across splits; Table 3). **(A)** Rolling-median MAE across genes ordered by difficulty, with population shrinkage and the $\mathbb{R}^5$ Binomial baseline shown for reference: the inductive embedding tracks the joint fit. **(B)** Fraction of genes where the subsampled embedding beats each full-evidence baseline; the curves cross 50% near $f{=}2\%$. **(C)** Posterior width distributions at each subsampling fraction; posterior concentrates as evidence grows.

TCGA co-mutation. We validate on somatic mutation data from The Cancer Genome Atlas (TCGA) PanCancer Atlas (Weinstein et al., 2013; Hoadley et al., 2018; Cerami et al., 2012; Gao et al., 2013). The entity is a gene, the grouped observation a tumor sample, and the stratifying context is cancer cohort; mutation rates vary sharply across cohorts (Hoadley et al., 2018).

PPI protein-protein interaction. As a second domain we validate on protein-protein interaction data assembled from six independent yeast two-hybrid screens (Luck et al., 2020; Rolland et al., 2014; Yu et al., 2011; Venkatesan et al., 2009; Yang et al., 2016; Rual et al., 2005). Each gene is an entity, each screen a stratifying context, and PPI's candidate pairs are those whose two genes both appear in at least one screen.

Calibration (TCGA, Fig. 3). With every pair observed, the graph-embedding baselines are right on average once given the fitted link, but they compress their predictions into a narrow band. Held-out rates vary widely at any predicted value, and their MAE is within 0.001 of a constant predictor's (Table 1). Population shrinkage is well calibrated in this regime and is the strongest baseline. The experiment below removes the observations it depends on.

Dropped pairs (TCGA, Fig. 4). We remove a pair's observations while keeping both entities in the embedding, then predict the dropped pair from the positions its endpoints acquired elsewhere; the figure reports the geometric estimate, since dropped pairs have no counts to combine with. At every

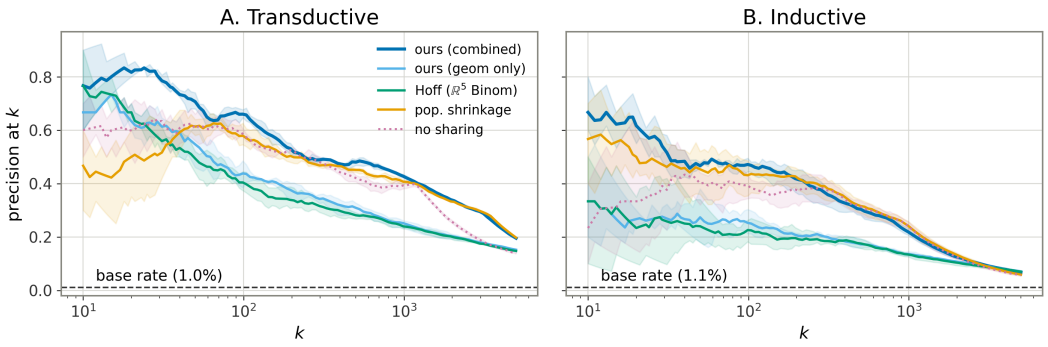


Figure 6: PPI precision at k (bands: 3 seeds; dashed: $\sim$1% base rate). **(A)** Transductive. **(B)** Inductive. The combined estimator lies above both estimates it combines at small k.

retention level the $\mathbb{H}^5$ Beta-Binomial fit has a lower mean MAE than every baseline. Population shrinkage steps from MAE 0.0391 in the control to 0.0445 the moment pairs are unobserved, and scores AUROC 0.500 at every retention level (Table 2), confirming it carries no relational information. The Beta-Binomial likelihood lowers MAE relative to the Binomial at every retention level (Fig. 4).

Flat and curved geometry. The dropped-pair gain comes from the shared positions and holds in flat and curved geometry alike: $\mathbb{R}^5$ and $\mathbb{H}^5$ agree within 0.0003 at every retention level (Table 1), consistent with Corollary 1, which involves the geometry only through its dimension. On PPI the hyperbolic fit adds 0.009 AUROC (Table 4). Curvature pays when it matches the generating geometry. On synthetic $\mathbb{H}^2\times\mathbb{S}^1$ data, the matched embedding is about six times more accurate than $\mathbb{H}^3$ or $\mathbb{R}^3$.

Held-out entities: inductive generalization (Fig. 5). A gene held out of training can be placed from its own evidence alone, with global parameters frozen. Across 3 splits the inductive embedding tracks the joint fit (Fig. 5A). With 5% of its training evidence, the embedding beats population shrinkage and Hoff's model, each using all the gene's data, on 65% and 71% of genes (Fig. 5B), and the posterior concentrates as evidence grows (Fig. 5C). On synthetic data with known rates, 80% credible intervals contain the true value for 79% of pairs (Appendix B.1). On PPI, the combined estimator recovers the calibration the geometry-only estimate loses to population shrinkage while keeping its discrimination advantage (Table 3).

Combined-estimator weights. The per-pair weights adapt to each domain. On TCGA, where the geometric estimate is already well-determined, $\bar{w} = 0.954$ and the combined estimator gains marginally. On PPI the weights split by interaction level, and the combined estimator improves on both the geometry-only estimate and population shrinkage (Table 4; Fig. 6).

8 Conclusion

We introduced a generative model for inferring latent geometric embeddings from grouped relational data, where relatedness is estimated rather than observed. A pair's gain depends not only on its own evidence but on how precisely its two entities are located by their other pairs, so that sparse pairs between well-located entities borrow strength that pair-level methods cannot. The same quantities yield a combined estimator with no fitted parameters, and the generative likelihood places entities never seen in training. On cancer co-mutation data, 5% of an entity's evidence suffices to outpredict full-evidence baselines on two thirds of genes. Two directions remain open. One is whether the joint fit attains the predicted variance when evidence is sparse and uneven, as it does when every pair is observed. The other is causal inference, which could proceed by conditioning the observation layer on covariates such as tumor mutation burden.

Reproducibility Statement

Code for the generative model, inference (MAP and RSGLD), and all comparison methods will be released upon acceptance. Synthetic data generation is fully specified in §7.1 and reproducible from the stated parameters. Real-data experiments use TCGA PanCancer Atlas data (publicly available via cBioPortal) and protein-protein interaction data assembled from six public yeast two-hybrid screens. The full theory derivation is in Appendix A.

AI Use Statement

We used generative AI tools for the following tasks requiring disclosure: generating synthetic datasets; assisting in the writing of proofs (symbolic differentiation and algebraic simplification in the derivation of Proposition 1); implementing methods (writing and editing code); and feedback on experimental design. We additionally used them for editing the paper to improve readability, drafting some passages that the authors then revised, creating figures, searching for literature alongside conventional search, and scheduling batch jobs.

All derivations were verified by hand, all generated code was read and tested, and every reported number traces to verified code. All references were checked against their sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Avanti Athreya, Donniell E. Fishkind, Minh Tang, Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018. URL http://jmlr.org/papers/v18/17-448.html.

Gregor Bachmann, Gary Bécigneul, and Octavian-Eugen Ganea. Constant curvature graph convolutional networks. In *International Conference on Machine Learning*, pp. 486–496, 2020.

Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013. ISSN 1467-9469. doi: 10.1111/sjos.12036. URL http://dx.doi.org/10.1111/sjos.12036.

Sander Canisius, John W.M. Martens, and Lodewyk F.A. Wessels. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biology*, 17(1):261, 2016.

Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012.

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

Ines Chami, Sami Abu-El-Haija, Bryan Perozzi, Christopher Ré, and Kevin Murphy. Machine learning on graphs: A model and comprehensive taxonomy. *Journal of Machine Learning Research*, 23 (89):1–64, 2022.

Alex Chen, Philippe Chlenski, Kenneth Munyuza, Antonio Khalil Moretti, Christian A. Naesseth, and Itsik Pe'er. Variational combinatorial sequential monte carlo for bayesian phylogenetics in hyperbolic space. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025. URL https://arxiv.org/abs/2501.17965.

P. Chlenski, Q. Chu, R. Khan, K. Du, A. Moretti, and I. Pe'er. Mixed-curvature decision trees and random forests. In *International Conference on Machine Learning*, 2025.

Philippe Chlenski, Ethan Turok, Antonio Moretti, and Itsik Pe'er. Fast hyperboloid decision tree algorithms. In *International Conference on Learning Representations*, 2024.

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

Alexei J. Drummond, Marc A. Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973, 2012.

William DuMouchel. Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. *The American Statistician*, 53(3):177–190, 1999. ISSN 00031305, 15372731. URL http://www.jstor.org/stable/2686093.

Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2285814.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1972–1982. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/franceschi19a.html.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pp. 1646–1655, 2018.

Jianjiong Gao, Bulent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):pl1, 2013.

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

Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A*, 170(2):301–354, 2007.

Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vesteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pp. 2688–2697, 2018.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, 1965.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Phys. Rev. E*, 82:036106, Sep 2010. doi: 10.1103/PhysRevE.82.036106. URL https://link.aps.org/doi/10.1103/PhysRevE.82.036106.

Pavel N. Krivitsky, Mark S. Handcock, Adrian E. Raftery, and Peter D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213, 2009. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2009.04.001. URL https://www.sciencedirect.com/science/article/pii/S0378873309000173.

Shane Lubold, Arun G Chandrasekhar, and Tyler H McCormick. Identifying the latent space geometry of network models through analysis of curvature. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):240–292, 04 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad002. URL https://doi.org/10.1093/jrsssb/qkad002.

Kerstin Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J Campos-Laborie, Benoit Charloteaux, et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman & Hall, 2nd edition, 1989.

Takahiro Mimori and Michiaki Hamada. Geophy: Differentiable phylogenetic inference via geometric gradients of tree topologies. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32634–32654, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/732c5757aa5577de9b103332cf7ac0bf-Abstract-Conference.html.

Antonio Khalil Moretti, Liyi Zhang, Christian A. Naesseth, Hadiah Venner, David M. Blei, and Itsik Peér. Variational combinatorial sequential monte carlo methods for bayesian phylogenetic inference. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161 of *Proceedings of Machine Learning Research*, pp. 971–981. PMLR, 2021. URL https://proceedings.mlr.press/v161/moretti21a.html.

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*, pp. 4693–4702, 2019.

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

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pp. 3779–3788, 2018.

Fragkiskos Papadopoulos, Constantinos Psomas, and Dmitri Krioukov. Network mapping by replaying hyperbolic growth. *IEEE/ACM Trans. Netw.*, 23(1):198–211, February 2015. ISSN 1063-6692. doi: 10.1109/TNET.2013.2294052. URL https://doi.org/10.1109/TNET.2013.2294052.

Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, et al. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, 2014.

Fredrik Ronquist, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard, and John P. Huelsenbeck. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542, 2012.

Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko NHirozawa, Gabriel F Berriz, Frederick D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E. Priebe. A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1446–1473, 09 2022. ISSN 1369-7412. doi: 10.1111/rssb.12509. URL https://doi.org/10.1111/rssb.12509.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *International Conference on Machine Learning*, pp. 4460–4469, 2018.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pp. 355–366, 2011.

Daniel K. Sewell and Yuguo Chen. Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657, 2015.

Anna L. Smith, Dena M. Asta, and Catherine A. Calder. The geometry of continuous latent space models for network data. *Statistical Science*, 34(3):428–453, 2019. doi: 10.1214/19-STS702. URL https://projecteuclid.org/journals/statistical-science/volume-34/issue-3/The-Geometry-of-Continuous-Latent-Space-Models-for-Network-Data/10.1214/19-STS702.full.

Kavitha Venkatesan, Jean-François Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozawa, Tong Hao, et al. An empirical framework for binary interactome mapping. *Nature Methods*, 6(1):83–90, 2009.

R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, 61(3):439–447, 1974.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R. Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, et al. The Cancer Genome Atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688, 2011.

Hanming Yang, Antonio Khalil Moretti, Sebastian Macaluso, Philippe Chlenski, Christian A. Naesseth, and Itsik Pe'er. Variational pseudo marginal methods for jet reconstruction in particle physics. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2406.03242. CoRR abs/2406.03242.

Xinping Yang, Jasmin Coulombe-Huntington, Shuli Kang, Gloria M Sheynkman, Tong Hao, Aaron Richardson, Song Sun, Fan Yang, Yun A Shen, Ryan R Murray, et al. Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, 164(4):805–817, 2016.

Haiyuan Yu, Leah Tardivo, Stanley Tam, Evan Weiner, Fana Gebreab, Changyu Fan, Nenad Svrzikapa, Tomoko Hirozane-Kishikawa, Edward Rietman, Xinping Yang, Julie Sahalie, Kourosh Salehi-Ashtiani, Tong Hao, Michael E. Cusick, David E. Hill, Frederick P. Roth, Pascal Braun, and Marc Vidal. Next-generation sequencing to generate interactome datasets. *Nature Methods*, 8(6):478–480, apr 2011. ISSN 1548-7105. doi: 10.1038/nmeth.1597.

A A VARIANCE COMPARISON FOR SHARED GEOMETRIC EMBEDDINGS

This appendix derives Proposition 1 and Corollary 1. Together they show that the shared embedding gains most over a per-pair estimator when a pair carries a small share of its endpoints' total information, the mechanism behind the sparse-pair results in §7.1 and §7.2. The argument is an asymptotic-variance comparison in the spirit of classical shrinkage estimation (Efron & Morris, 1975), adapted to pooling through a shared *low-dimensional geometric embedding* rather than toward a single scalar mean. We state which steps are exact and which rely on an approximation. We treat κ, r, and t as known, and state the argument for a single constant-curvature factor of dimension D.

A.1 PER-PAIR INFORMATION

Fix a pair (i, j) and write $p = p_{ij}$, $\alpha = p\kappa$, $\beta = (1-p)\kappa$, so $\alpha + \beta = \kappa$ is constant in p. From Eq. 8, the log-likelihood of one context's count k_s (out of n_s trials) is

$$\log P(k_s \mid n_s, p, \kappa) = \log \binom{n_s}{k_s} + \log\Gamma(k_s + \alpha) + \log\Gamma(n_s - k_s + \beta) - \log\Gamma(n_s + \alpha + \beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) + \log\Gamma(\alpha + \beta). \quad (15)$$

Since $\alpha + \beta = \kappa$ does not depend on p, the terms $\log\Gamma(\alpha + \beta)$ and $\log\Gamma(n_s + \alpha + \beta)$ are constant in p and vanish under d/dp. The remaining terms each depend on p through either $\alpha = p\kappa$ or $\beta = (1-p)\kappa$. Writing $\psi = (\log\Gamma)'$ for the digamma function and applying the chain rule term by term:

$$\frac{d}{dp}\log\Gamma(k_s + \alpha) = \psi(k_s + \alpha) \cdot \frac{d\alpha}{dp} = \kappa\,\psi(k_s + \alpha), \quad (16)$$

$$\frac{d}{dp}\log\Gamma(n_s - k_s + \beta) = \psi(n_s - k_s + \beta) \cdot \frac{d\beta}{dp} = -\kappa\,\psi(n_s - k_s + \beta), \quad (17)$$

$$\frac{d}{dp}\big[-\log\Gamma(\alpha)\big] = -\kappa\,\psi(\alpha), \quad (18)$$

$$\frac{d}{dp}\big[-\log\Gamma(\beta)\big] = \kappa\,\psi(\beta). \quad (19)$$

Collecting terms, the score is

$$s(k_s; p) = \kappa\Big\{\big[\psi(k_s + \alpha) - \psi(\alpha)\big] - \big[\psi(n_s - k_s + \beta) - \psi(\beta)\big]\Big\}. \quad (20)$$

The true Fisher information $\mathrm{Var}(s(K_s; p))$ requires an expectation of squared digamma differences under K_s's Beta-Binomial distribution, which has no closed form. We therefore substitute the standard *quasi-likelihood* information (Wedderburn, 1974; McCullagh & Nelder, 1989), a standard tool for models where the exact Fisher information is intractable but the first two moments are known:

$$\mathcal{I}(p; n_s) := \frac{\left(\frac{d}{dp}\mathbb{E}[K_s]\right)^2}{\mathrm{Var}(K_s)}. \quad (21)$$

Where the exact Fisher information measures how sharply the log-likelihood peaks around the true p, the quasi-likelihood information asks how precisely p can be pinned down from the mean–variance relationship alone: the numerator is the sensitivity of the expected count to p, the denominator is the noise. For the rest of the derivation, only a per-pair scalar measuring "how well can this pair's own data determine p_{ij}" is needed; the quasi-likelihood provides this, in closed form.

Beta-Binomial moments. We need $\mathbb{E}[K_s]$ and $\mathrm{Var}(K_s)$. Conditionally, $K_s \mid \theta_s \sim \mathrm{Binomial}(n_s, \theta_s)$, with $\theta_s \sim \mathrm{Beta}(\alpha, \beta)$. By iterated expectation, $\mathbb{E}[K_s] = n_s\,\mathbb{E}[\theta_s] = n_s p$. For the variance, the law of total variance gives

$$\begin{aligned}\mathrm{Var}(K_s) &= \mathbb{E}\big[\mathrm{Var}(K_s \mid \theta_s)\big] + \mathrm{Var}\big(\mathbb{E}[K_s \mid \theta_s]\big) \\ &= \mathbb{E}\big[n_s\,\theta_s(1-\theta_s)\big] + \mathrm{Var}\big(n_s\,\theta_s\big) \\ &= n_s\big[\mathbb{E}[\theta_s] - \mathbb{E}[\theta_s^2]\big] + n_s^2\,\mathrm{Var}(\theta_s). \end{aligned} \quad (22)$$

Using the Beta moments $\mathrm{Var}(\theta_s) = \frac{p(1-p)}{\kappa+1}$ and $\mathbb{E}[\theta_s^2] = \mathrm{Var}(\theta_s) + p^2 = \frac{p(1-p)}{\kappa+1} + p^2$, substituting into Eq. 22 and simplifying:

$$\mathrm{Var}(K_s) = n_s\, p(1-p)\left[1 + \frac{n_s - 1}{\kappa + 1}\right]. \tag{23}$$

The factor in brackets is the *variance inflation factor*: it equals 1 when $\kappa \to \infty$ (no overdispersion, plain Binomial) and grows linearly in n_s for finite κ.

Quasi-likelihood information. Substituting $\mathbb{E}[K_s] = n_s p$ and Eq. 23 into Eq. 21:

$$\mathcal{I}(p; n_s) = \frac{(n_s)^2}{n_s\, p(1-p)\left[1 + \frac{n_s-1}{\kappa+1}\right]} = \frac{n_s}{p(1-p)\left[1 + \frac{n_s-1}{\kappa+1}\right]} = \frac{\mathcal{I}_{\mathrm{Bin}}(p; n_s)}{1 + \frac{n_s-1}{\kappa+1}}, \tag{24}$$

i.e. the plain-Binomial information deflated by the variance inflation factor. Two immediate consequences: (i) as $n_s \to \infty$, $\mathcal{I}(p; n_s) \to (\kappa+1)/[p(1-p)]$, a *finite cap*: additional within-stratum data has strictly diminishing returns, the same phenomenon as the design effect in clustered-data and survey sampling (Kish, 1965); (ii) since strata are independent given p, informations add across the S strata of pair (i, j):

$$\mathcal{I}_{ij} := \sum_{s=1}^{S} \mathcal{I}(p_{ij}; n_{ijs}), \qquad \mathrm{Var}\big(\hat{p}_{ij}^{\,\mathrm{indep}}\big) \approx \mathcal{I}_{ij}^{-1}. \tag{25}$$

$\mathcal{I}_{ij}$ is the information available to an estimator of p_{ij} that uses only pair (i, j)'s own data (the *no sharing* baseline of §7.1).

A.2 Node-Level Information

Node i's position z_i enters the likelihood only through p_{ij}, for every j it is paired with. We now derive how the per-pair information $\mathcal{I}_{ij}$ about p_{ij} translates into information about the position z_i itself.

Fisher-information chain rule. Consider a single pair (i, j). The propensity p_{ij} is a scalar function of z_i (with z_j and the global parameters held fixed). For a scalar parameter p that is a differentiable function of a D-dimensional parameter z, the Fisher information transforms as

$$\mathcal{I}^{(z)} = \mathcal{I}^{(p)} \cdot \big(\nabla_z p\big)\big(\nabla_z p\big)^\top, \tag{26}$$

where $\mathcal{I}^{(p)}$ is the scalar Fisher information about p and $(\nabla_z p)(\nabla_z p)^\top$ is the $D \times D$ rank-one outer product of the gradient. (This is the multivariate delta method applied to the information: $\mathrm{Var}(\hat{z}) \approx (\mathcal{I}^{(z)})^{-1}$ and $\mathrm{Var}(\hat{p}) \approx (\nabla_z p)^\top \mathrm{Var}(\hat{z})(\nabla_z p)$, inverted.)

Since z_i is involved in every pair (i, j) for $j \in \mathcal{N}(i)$, and the pairs contribute independently to the likelihood given the positions, the total information about z_i is the sum over all such pairs:

$$\mathcal{I}_i := \sum_{j \in \mathcal{N}(i)} \mathcal{I}_{ij}\, \big(\nabla_{z_i} p_{ij}\big)\big(\nabla_{z_i} p_{ij}\big)^\top, \tag{27}$$

a $D \times D$ matrix summing one rank-one contribution per neighbor $j \in \mathcal{N}(i)$.

Gradient of p_{ij} with respect to z_i. We compute $\nabla_{z_i} p_{ij}$ by differentiating through the Fermi–Dirac link (Eq. 6). Writing $\sigma(x) = 1/(1+e^{-x})$ with $\sigma'(x) = \sigma(x)(1-\sigma(x))$, and $x_{ij} = (r - d_{ij}^2)/t$ where $d_{ij} = d(z_i, z_j)$:

$$\nabla_{z_i} p_{ij} = \sigma'(x_{ij}) \cdot \nabla_{z_i} x_{ij} = p_{ij}(1-p_{ij}) \cdot \frac{-\nabla_{z_i}(d_{ij}^2)}{t}. \tag{28}$$

The squared distance d_{ij}^2 is a function of z_i (with z_j fixed). Its gradient with respect to z_i is

$$\nabla_{z_i}(d_{ij}^2) = 2\, d_{ij}\, \nabla_{z_i}\, d_{ij}. \tag{29}$$

Unit-norm gradient of distance. On a Riemannian manifold, the gradient of the distance function $d(\cdot, y)$ with respect to its first argument at a point $x \neq y$ (away from the cut locus) is the unit tangent vector pointing away from y:

$$\nabla_x \, d(x, y) = -\frac{\mathrm{Log}_x(y)}{d(x, y)}, \qquad \|\nabla_x \, d(x, y)\| = 1. \tag{30}$$

This follows from the first variation of arc length (see e.g. do Carmo, 1992, Chapter 3): the derivative of $d(x, y)$ along a curve through x with velocity v equals $-\langle v, \dot{\gamma}(0)\rangle$, where $\dot{\gamma}(0)$ is the unit tangent to the minimizing geodesic from x to y. On $\mathbb{H}^D$ and $\mathbb{R}^D$, which are uniquely geodesic (no cut locus), this holds everywhere for $x \neq y$.

Combining. Substituting Eq. 30 and Eq. 29 into Eq. 28:

$$\nabla_{z_i} p_{ij} = p_{ij}(1 - p_{ij}) \cdot \frac{-2 \, d_{ij}}{t} \cdot \nabla_{z_i} \, d_{ij} = p_{ij}(1 - p_{ij}) \cdot \frac{2 \, d_{ij}}{t} \cdot \frac{\mathrm{Log}_{z_i}(z_j)}{d_{ij}}. \tag{31}$$

The d_{ij} factors simplify, and the remaining vector $\mathrm{Log}_{z_i}(z_j)/d_{ij}$ is unit-norm, giving:

$$\left\|\nabla_{z_i} p_{ij}\right\|^2 = \frac{4 d_{ij}^2 \, [p_{ij}(1 - p_{ij})]^2}{t^2}, \qquad u_{ij} := \frac{\nabla_{z_i} p_{ij}}{\|\nabla_{z_i} p_{ij}\|} \text{ points from } z_i \text{ toward } z_j. \tag{32}$$

Thus $\mathcal{I}_i$ (Eq. 27) is a sum, over i's neighbors, of rank-one matrices whose *directions* u_{ij} are purely geometric (which way each neighbor sits) and whose *magnitudes* $m_{ij} := \mathcal{I}_{ij} \|\nabla_{z_i} p_{ij}\|^2$ combine each pair's information with the link's local sensitivity.

Remark 1 (Identifiability requires directional spread, not just degree)**.** *If i's neighbors all lie in similar directions, $\mathcal{I}_i$ remains low-rank / ill-conditioned no matter how many neighbors i has. Position estimation is underdetermined in the orthogonal directions. Degree alone is not sufficient for good position estimation; the neighborhood must also be directionally diverse. This has no analogue in classical scalar shrinkage, where there is no notion of "direction" for a single pooled mean.*

Shared estimator variance via the delta method. The propensity $p_{ij} = p_{ij}(z_i, z_j)$ depends on both endpoints. Under the approximation that the estimation errors of z_i and z_j are independent (a leading-order, profile-style assumption: we treat other nodes' positions as fixed when computing the information about one node), the delta method gives Proposition 1 (Eq. 12).

Proof of Proposition 1. Under the endpoint-independence approximation, the estimation errors of z_i and z_j contribute independently. By the multivariate delta method, $\mathrm{Var}(\hat{p}_{ij})$ from endpoint i alone is $(\nabla_{z_i} p_{ij})^\top \mathrm{Var}(\hat{z}_i)(\nabla_{z_i} p_{ij}) = (\nabla_{z_i} p_{ij})^\top \mathcal{I}_i^{-1} (\nabla_{z_i} p_{ij})$, where $\mathcal{I}_i$ is the node-level Fisher information (Eq. 27). Summing the two independent endpoint contributions gives Eq. 12. □

This is the one non-trivial approximation in the derivation: it ignores that pair (i, j) contributes to both $\mathcal{I}_i$ and $\mathcal{I}_j$ simultaneously; a fully rigorous treatment would require a single joint information matrix over all $\{z_i\}_{i=1}^N$. Corollary 1 in the main text follows from Proposition 1 by adding one further, separately stated approximation (isotropic, homogeneous neighborhoods, Assumption 2 below) that collapses the matrix expression to a scalar.

A.3 The Scalar Form Under Isotropic Neighborhoods

Assumption 2 (Isotropic neighborhoods)**.** *The unit vectors $\{u_{ij}\}_{j \in \mathcal{N}(i)}$ are distributed with no preferred direction in $T_{z_i}\mathcal{M}$, and the per-neighbor magnitude $m_{ij} := \mathcal{I}_{ij} \|\nabla_{z_i} p_{ij}\|^2$ is uncorrelated with direction.*

Under Assumption 2, $\mathbb{E}[u_{ij} u_{ij}^\top] = \frac{1}{D} I_D$ by symmetry (the trace must equal $\|u_{ij}\|^2 = 1$, and no direction is preferred), so Eq. 27 concentrates around

$$\mathcal{I}_i \approx \frac{M_i}{D} I_D, \qquad M_i := \sum_{j \in \mathcal{N}(i)} m_{ij}. \tag{33}$$

Symmetry of gradient norms. Because $d(z_i, z_j) = d(z_j, z_i)$ and the Fermi–Dirac link (Eq. 6) is a function of d^2 alone, the chain rule through p_{ij} has identical structure from either endpoint. The gradient norm (Eq. 32) depends only on d_{ij} and p_{ij}, both of which are symmetric, so $\|\nabla_{z_i} p_{ij}\| = \|\nabla_{z_j} p_{ij}\|$. We write $g_{ij} := \|\nabla_{z_i} p_{ij}\|^2$ for this common squared norm.

Substituting into the shared variance. With $\mathcal{I}_i \approx \frac{M_i}{D} I_D$, the first term of Eq. 12 becomes

$$(\nabla_{z_i} p_{ij})^\top \mathcal{I}_i^{-1} (\nabla_{z_i} p_{ij}) = (\nabla_{z_i} p_{ij})^\top \cdot \frac{D}{M_i} I_D \cdot (\nabla_{z_i} p_{ij}) = \frac{D}{M_i} \|\nabla_{z_i} p_{ij}\|^2 = \frac{D\, g_{ij}}{M_i}.$$

The second term is identical with j replacing i (using the gradient norm symmetry above), giving:

$$\mathrm{Var}\big(\hat{p}_{ij}^{\,\mathrm{shared}}\big) \approx D\, g_{ij} \left(\frac{1}{M_i} + \frac{1}{M_j} \right). \tag{34}$$

Taking the ratio: information-share form. Corollary 1 follows from Eq. 34 in one step, with no further approximation. By Assumption 1 the independent estimator has $\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{indep}}) \approx \mathcal{I}_{ij}^{-1}$, and since $m_{ij} = \mathcal{I}_{ij}\, g_{ij}$ this is $\mathcal{I}_{ij}^{-1} = g_{ij}/m_{ij}$. Dividing:

$$\frac{\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{shared}})}{\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{indep}})} \;\approx\; \frac{D\, g_{ij} \big(\frac{1}{M_i} + \frac{1}{M_j} \big)}{g_{ij}/m_{ij}} \;=\; D \left(\frac{m_{ij}}{M_i} + \frac{m_{ij}}{M_j} \right). \tag{35}$$

The gradient norm g_{ij} cancels, leaving the pair's share of the information pinning down each endpoint. This is the form stated in Corollary 1 and used throughout; the degree reduction below is a further approximation, not an equivalent statement.

Network-homogeneity approximation. If we further treat the network as approximately homogeneous, each node's per-neighbor magnitude is roughly constant: $m_{ij} \approx \bar{g}$ for a network-wide average $\bar{g}$, so $M_i \approx \deg(i) \cdot \bar{g}$. Under this approximation, a typical pair's own per-pair information is also of order $\bar{g}$, giving $\mathcal{I}_{ij} \approx \bar{g}/g_{ij}$ (since $m_{ij} = \mathcal{I}_{ij} g_{ij}$) and hence $\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{indep}}) \approx \mathcal{I}_{ij}^{-1} \approx g_{ij}/\bar{g}$.

Taking the ratio. Dividing Eq. 34 by $\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{indep}}) \approx g_{ij}/\bar{g}$:

$$\frac{\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{shared}})}{\mathrm{Var}(\hat{p}_{ij}^{\,\mathrm{indep}})} \;\approx\; \frac{D\, g_{ij} \big(\frac{1}{\deg(i)\,\bar{g}} + \frac{1}{\deg(j)\,\bar{g}} \big)}{g_{ij}/\bar{g}} \;=\; D \left(\frac{1}{\deg(i)} + \frac{1}{\deg(j)} \right), \tag{36}$$

a degree-only reduction of Eq. 35. The pair-specific terms g_{ij} and the network-wide average $\bar{g}$ cancel, leaving a ratio that depends only on the embedding dimension D and the degrees of the two endpoints. This form rests on network homogeneity in addition to isotropy (Appendix A.4).

Consequences. Two follow directly. First, the shrinkage benefit should track the *degree* of the two entities, not only the pair's own sample size n_{ij}: a sparse pair between two well-connected entities should calibrate nearly as well as a data-rich pair, while a sparse pair between two poorly-connected entities should not. This is a finer claim than n_{ij}-only bucketing. Second, for fixed degree, higher embedding dimension D dilutes the benefit: the same amount of neighbor information is spread across more directions.

Remark 2 (Scope of the approximations)**.** *The per-pair quasi-likelihood substitution, Beta-Binomial variance formula, and design-effect observation are textbook results (Wedderburn, 1974; McCullagh & Nelder, 1989; Kish, 1965). The chain rule for Fisher information and the geometric fact about the gradient norm (§A.2) are exact. The two approximations specific to this derivation are endpoint independence (Proposition 1) and isotropic neighborhoods (Corollary 1). On fitted curved embeddings, neighbors rarely lie in uniformly distributed directions, so node-level information matrices are not scalar multiples of the identity (Appendix A.4). Endpoint independence is standard in latent-position models. The contribution of this appendix is the combination: a variance comparison for pooling through a shared, low-dimensional geometric embedding under a Beta-Binomial observation model. To our knowledge, this has not been stated in the shrinkage-estimation literature; the closest precedents are the latent-space model of Hoff et al. (2002) with its curved-geometry extensions (Smith et al., 2019; Lubold et al., 2023), and nonparametric graphon estimation. The curved extensions study which geometry best describes an observed network, one tie per pair; none of these estimates relatedness from overdispersed grouped counts or characterizes which pairs the shared positions help estimate.*

A.4 Matrix Form vs. Isotropic Reduction

The three forms of the variance comparison sit in a strict hierarchy of assumptions. Proposition 1 requires only endpoint independence. Corollary 1 adds Assumption 2, that each node's information matrix is a scalar multiple of the identity. The degree reduction of §A.3 adds network homogeneity on top of that. Each step buys a simpler expression at the cost of a stronger premise.

On a fitted curved embedding a node's neighbors rarely lie in uniformly distributed directions, so Assumption 2 holds only approximately and Proposition 1 is the general statement. The predicted ratio in Fig. 2C and the weights of §A.5 use Corollary 1, which needs only the magnitudes m_{ij} and their node totals M_i, not the per-node information matrix and its inverse. Replacing the Corollary 1 variance with the matrix form of Proposition 1 changes the weights little (correlation 0.998 on PPI), and the reported metrics are unchanged within split-to-split variation.

A.5 A Constructive Combined Estimator

Proposition 1 compares variances, but its inputs are quantities the fitted model already supplies at prediction time, namely positions, link parameters, and training counts. The same ingredients therefore yield a constructive result, a combined estimator that improves on the pure geometric estimate with no additional fitted parameters.

Population shrinkage (§7.1) is itself a shrinkage estimator. Reading the fitted concentration κ_g as $\mathcal{I}_{\text{prior}} := \kappa_g/[p_g(1-p_g)]$ pseudo-observations of the global rate p_g (the standard correspondence between a Beta prior and pseudo-data), its implicit weight on a pair's own information is $s_{ij} = \mathcal{I}_{ij}/(\mathcal{I}_{ij} + \mathcal{I}_{\text{prior}})$, so $\text{Var}(\hat{p}_{ij}^{\text{pop}}) \approx s_{ij}^2/\mathcal{I}_{ij}$, in the same units as $\text{Var}(\hat{p}_{ij}^{\text{indep}}) = 1/\mathcal{I}_{ij}$ (Eq. 25). Combining this with Corollary 1 in its information-share form (Eq. 13, rather than the degree reduction, which requires a further network-homogeneity assumption; Appendix A.4) gives $\text{Var}(\hat{p}_{ij}^{\text{shared}}) \approx D\, m_{ij}(1/M_i + 1/M_j)/\mathcal{I}_{ij}$, in the same units. The minimum-variance linear combination of the two estimates is then

$$w_{ij} = \frac{\text{Var}(\hat{p}_{ij}^{\text{pop}})}{\text{Var}(\hat{p}_{ij}^{\text{shared}}) + \text{Var}(\hat{p}_{ij}^{\text{pop}})}, \qquad \hat{p}_{ij}^{\text{combo}} = w_{ij}\,\hat{p}_{ij}^{\text{shared}} + (1 - w_{ij})\,\hat{p}_{ij}^{\text{pop}}. \tag{37}$$

A scale multiplier λ on $\mathcal{I}_{\text{prior}}$ interpolates between the pure geometric estimate ($\lambda \to 0$) and population shrinkage ($\lambda \to \infty$); $\lambda = 1$ uses the fitted κ_g as-is, with no additional parameter.

At $\lambda = 1$, the combined estimator's behaviour differs sharply between the two domains, and the weights explain why.

On TCGA, training counts are large (median n=326) and the geometric estimate is already well-determined, so $\bar{w} \approx 0.95$. The combined estimator places almost all of its weight on geometry and preserves the geometric estimate's accuracy, improving on it at most splits and never measurably degrading it. On a split where the geometric estimate loses to population shrinkage, no λ causes the combined estimate to surpass population shrinkage; it converges to population shrinkage in the limit, consistent with §A.3's ratio being a variance argument that does not bound bias from geometric misspecification.

On PPI, training counts are sparse (median n=3) and coverage varies across six screens. The weights adapt per pair: $w_{ij} \approx 0.90$ for low-interaction pairs, where population shrinkage has little signal beyond the global rate, dropping to $w_{ij} \approx 0.25$ for high-interaction pairs, where the per-pair counts are informative. This adaptive weighting improves on both of the estimates it combines, leading the geometry-only estimate and population shrinkage on area under the precision–recall curve and on precision at 100 (Table 4).

Selecting λ from the held-out metric being reported would be circular; selecting it by training likelihood is invalid for a different reason, since population shrinkage's own prediction is constructed from the exact per-pair counts that likelihood evaluates, which favors it independent of which estimator actually generalizes. We report only the zero-parameter case ($\lambda = 1$) throughout.

B Additional Results

B.1 Posterior Coverage on Synthetic Data

Fig. 5B,C shows that accuracy and posterior width improve with evidence, but not that the intervals are correctly sized. We check the latter directly on synthetic data, where the true p_{ij} is known. Using a purely hyperbolic configuration (N=250, $\mathbb{H}^5$, κ=8, 5 seeds; 3 chains, 4,000 burn-in steps, 200 samples thinned by 20, step size 3×10^{-5}), we form a central 80% credible interval for every pair from the pooled RSGLD samples and ask how often it contains the true value.

Stationarity check. A chain joins the pool if, over its post-burn-in trace, the log-posterior has no residual trend (second and first halves differ in mean by less than the trace's standard deviation) and the sampled positions stay near the mode (mean geodesic distance from the MAP estimate below a characteristic length of the configuration: twice the generator's inter-node step here, the prior scale γ on real data). Chains failing either are discarded; no run reported here lost all of them.

Sampling dominates the cost and does not depend on the level, so we draw once per seed and re-reduce the same pooled samples at every nominal level from 0.50 to 0.95. Across 31,125 pairs per seed, empirical coverage tracks nominal to within 1.3 percentage points everywhere, 0.4957 at nominal 0.50, 0.7897 at 0.80, 0.9375 at 0.95, with an across-seed standard deviation of about 1.2 points, i.e. comparable to the deviation itself. Coverage is also stable across evidence: splitting pairs into quartiles by their own evidence, per-quartile coverage stays within 0.765 to 0.816 at every seed, so the intervals are not merely correct on average while being overconfident for data-poor pairs and underconfident for data-rich ones. The posterior is therefore usable as a statement of uncertainty and not only as a route to a point estimate, which is the property a point-estimate method has no analogue of.

B.2 TCGA Dropped Pairs

AUROC threshold on TCGA. On TCGA each pair has a median of 80 held-out observations pooled across the 10 cohorts (median 326 in training), so the empirical rate is continuous. To compute AUROC we binarize at a threshold of 0.20: a pair is labeled positive if its empirical rate exceeds the threshold. Calibrated methods (AUROC ≈ 0.75) are not separated by this metric because they cluster near the true rates; MAE is therefore primary for TCGA.

	f=1.0	Dropped pairs (never observed in training)			
Method	retained	f=0.5	f=0.25	f=0.1	f=0.05
Best constant (median held-out rate)	0.0416	0.0416	0.0416	0.0417	0.0416
Population shrinkage	0.0391 ± 0.0043	0.0445 ± 0.0026	0.0445 ± 0.0026	0.0444 ± 0.0025	0.0444 ± 0.0025
Conditional independence	0.1290 ± 0.0102	0.1290 ± 0.0104	0.1290 ± 0.0104	0.1290 ± 0.0103	0.1290 ± 0.0103
DISCOVER, shrunk	0.0414 ± 0.0051	0.0415 ± 0.0052	0.0414 ± 0.0051	0.0414 ± 0.0051	0.0414 ± 0.0051
$\mathbb{R}^5$, Binomial (Hoff)	0.0406 ± 0.0050	0.0409 ± 0.0049	0.0410 ± 0.0049	0.0415 ± 0.0048	0.0425 ± 0.0047
$\mathbb{H}^5$, Binomial	0.0406 ± 0.0050	0.0409 ± 0.0049	0.0411 ± 0.0049	0.0415 ± 0.0048	0.0425 ± 0.0047
$\mathbb{R}^5$, Beta-Binomial	0.0366 ± 0.0013	0.0368 ± 0.0011	0.0370 ± 0.0011	0.0376 ± 0.0009	0.0388 ± 0.0010
$\mathbb{H}^5$, BB, geometry only	0.0366 ± 0.0012	0.0368 ± 0.0011	0.0370 ± 0.0010	0.0375 ± 0.0010	0.0385 ± 0.0010
$\mathbb{H}^5$, BB + intercepts	0.0358 ± 0.0012	0.0360 ± 0.0011	0.0361 ± 0.0011	0.0369 ± 0.0010	0.0413 ± 0.0025
Intercept only (no geometry)[b]	0.0397 ± 0.0012	0.0399 ± 0.0010	0.0399 ± 0.0011	0.0400 ± 0.0011	0.0402 ± 0.0011
N&K, ranking loss[a]	0.0423 ± 0.0045		—		
Gu, distortion ($\mathbb{H}^5$)[a]	0.0408 ± 0.0047		—		

Table 1: TCGA dropped-pair MAE across retention fractions (mean±std, 4 splits, N=1,000). The control column (f=1.0) retains all pairs; subsequent columns evaluate only pairs never observed in training. Embedding-method error rises slowly as pairs are dropped, while population shrinkage jumps to 0.0445 as soon as any pairs are unobserved. [a]Full-graph methods that require all pairs (3 splits). [b]$p_{ij} = \sigma((a_i + a_j)/t)$, same Beta-Binomial likelihood, no manifold.

Method	f=1.0 retained	Dropped pairs (never observed in training) f=0.5	f=0.25	f=0.1	f=0.05
Population shrinkage	0.744 ± 0.015	0.500 ± 0.000	0.500 ± 0.000	0.500 ± 0.000	0.500 ± 0.000
Conditional independence	0.671 ± 0.010	0.670 ± 0.012	0.670 ± 0.011	0.671 ± 0.010	0.671 ± 0.010
DISCOVER, shrunk	0.727 ± 0.014	0.727 ± 0.014	0.727 ± 0.014	0.727 ± 0.014	0.727 ± 0.014
$\mathbb{R}^5$, Binomial (Hoff)	0.751 ± 0.015	0.749 ± 0.016	0.747 ± 0.016	0.741 ± 0.016	0.728 ± 0.020
$\mathbb{H}^5$, Binomial	0.752 ± 0.015	0.749 ± 0.016	0.747 ± 0.016	0.741 ± 0.016	0.727 ± 0.021
$\mathbb{R}^5$, Beta-Binomial	0.747 ± 0.015	0.744 ± 0.016	0.742 ± 0.016	0.734 ± 0.015	0.721 ± 0.018
$\mathbb{H}^5$, BB, geometry only	0.747 ± 0.015	0.745 ± 0.015	0.742 ± 0.015	0.736 ± 0.015	0.723 ± 0.017
$\mathbb{H}^5$, BB + intercepts	0.754 ± 0.015	0.753 ± 0.015	0.751 ± 0.016	0.744 ± 0.018	0.712 ± 0.026
N&K, ranking loss[a]	0.710 ± 0.011		—		
Gu, distortion ($\mathbb{H}^5$)[a]	0.739 ± 0.012		—		

Table 2: TCGA dropped-pair AUROC across retention fractions (mean±std, 4 splits, N=1,000; threshold 0.20). Population shrinkage collapses to chance (0.500) on dropped pairs because all such pairs receive the same global estimate. Embedding methods retain discriminative power. [a]Full-graph methods (3 splits).

B.3 Held-Out Entities

Inductive generalization (Table 3). Table 3 reports held-out accuracy when 150 of 1000 entities are withheld entirely from training and placed inductively from their own evidence with all shared parameters frozen. On TCGA, where MAE is primary, the inductive embedding matches the joint fit and beats both non-relational baselines and the $\mathbb{R}^5$ Binomial (Hoff) baseline. On PPI, the combined inductive embedding has the highest AUROC (.834 vs. .810 for Hoff) and log-likelihood (-0.0411 vs. -0.0415 for population shrinkage), and lower MAE than population shrinkage (0.0131 vs. 0.0156). The combined estimator improves over the geometry-only inductive embedding on both AUROC and MAE, confirming that the variance-weighted blend (Eq. 14) helps in the inductive regime as well.

Method	TCGA (inductive) MAE ↓	log-lik ↑	PPI (inductive) MAE ↓	AUROC ↑	log-lik ↑
No sharing	0.0375 ± 0.0014	-0.0334 ± 0.0013	0.1414 ± 0.0006	$.761 \pm .009$	-0.1657 ± 0.0011
Population shrinkage	0.0391 ± 0.0052	-0.0338 ± 0.0024	0.0156 ± 0.0002	$.763 \pm .010$	-0.0415 ± 0.0008
$\mathbb{R}^5$, Binomial (Hoff)	0.0399 ± 0.0057	-0.0341 ± 0.0026	0.0178 ± 0.0002	$.810 \pm .013$	-0.0446 ± 0.0005
Ours (inductive)	0.0364 ± 0.0016	-0.0331 ± 0.0013	0.0131 ± 0.0001	$.834 \pm .004$	-0.0411 ± 0.0002

Table 3: Inductive held-out metrics (mean±std, 3 splits, 150 of 1000 entities held out). Each held-out entity is placed from its own evidence alone with all shared parameters frozen. PPI reports the combined estimator (Eq. 14); TCGA reports geometry only, since $\bar{w} \approx 0.95$ makes the combined estimate indistinguishable at this precision. TCGA: MAE is primary. PPI: AUROC is the most informative metric given class imbalance (0.9% positive rate); all PPI metrics are macro-averaged per gene.

B.4 PPI

Method	MAE ↓	log-lik ↑	AUROC ↑	ρ_s ↑	AUPRC ↑	P@100 ↑
Best constant predictor[†]	0.0089 ± 0.0001	-0.0498 ± 0.0003	.500	—	.010	.033
No sharing	0.1398 ± 0.0002	-0.1632 ± 0.0003	$.760 \pm .003$	$.098 \pm .001$	$.117 \pm .001$	$.580 \pm .014$
Population shrinkage	0.0152 ± 0.0000	-0.0395 ± 0.0002	$.763 \pm .003$	$.100 \pm .001$	$.158 \pm .001$	$.580 \pm .046$
N&K, ranking loss	0.0167 ± 0.0000	-0.0431 ± 0.0002	$.809 \pm .007$	$.108 \pm .003$	$.087 \pm .005$	$.480 \pm .052$
Gu, distortion[§]	0.0168 ± 0.0000	-0.0431 ± 0.0002	$.836 \pm .003$	$.118 \pm .001$	$.056 \pm .002$	$.130 \pm .010$
$\mathbb{R}^5$, Binomial[¶]	0.0161 ± 0.0001	-0.0406 ± 0.0003	$.882 \pm .003$	$.134 \pm .002$	$.120 \pm .003$	$.403 \pm .042$
$\mathbb{R}^5$, Beta-Binomial[‡]	0.0161 ± 0.0000	-0.0405 ± 0.0003	$.882 \pm .003$	$.134 \pm .001$	$.120 \pm .003$	$.393 \pm .032$
$\mathbb{H}^5$, Binomial	0.0162 ± 0.0002	-0.0394 ± 0.0002	$.892 \pm .003$	$.137 \pm .001$	$.127 \pm .004$	$.440 \pm .030$
$\mathbb{H}^5$, BB, geometry only (ablation)[‡]	0.0161 ± 0.0001	-0.0395 ± 0.0003	$.891 \pm .003$	$.137 \pm .002$	$.126 \pm .005$	$.437 \pm .045$
Ours ($\mathbb{H}^5$, BB, combined)[‡]	0.0124 ± 0.0000	-0.0379 ± 0.0002	$.891 \pm .004$	$.137 \pm .002$	$.198 \pm .003$	$.660 \pm .020$

Table 4: PPI test metrics on retained pairs (mean±std, 3 splits, N=1,000, n_s=1). Given the $\sim$1% positive rate, the ranking metrics are more informative than MAE. Ranking-metric intervals are across-seed standard deviations and are unavailable for the constant-predictor row. [†]Best constant per metric (§7.2). [§]Gu's $\mathbb{H}^2 \times \mathbb{R}^3$ scores MAE 0.0168 ± 0.0000, AUROC .835, AUPRC .048, P@100 .097. [¶]The $\mathbb{R}^5$ Binomial cell recovers the latent-space model of Hoff et al. (2002) extended to repeated observations; each further row adds one generalization. [‡]With n_s=1 the Beta-Binomial marginal (Eq. 8) is Bernoulli and κ drops out of the likelihood, so these are Bernoulli fits and the Binomial/Beta-Binomial pairs differ only by optimizer noise. PPI therefore does not exercise the overdispersion layer. The "geometry only" row evaluates the geometric estimate $\hat{p}_{ij}^{\text{shared}}$ alone, without the combined weighting of Eq. 14. On PPI 98% of candidate pairs are never observed interacting, so the top-20% rule of §7.1 is vacuous and the induced graph the two graph-embedding baselines receive is the set of pairs observed interacting at least once (1.7% of candidates).

C Computational Cost

Per-iteration complexity. Each gradient step evaluates the Beta-Binomial log-likelihood (Eq. 8) for every observed pair-context cell. For P active pairs and S contexts this is $O(PS)$; the Beta function evaluation is $O(1)$ per cell (three `lgamma` calls), so the per-iteration cost is the same order as a plain Binomial likelihood. The product-distance computation (Eq. 5) adds $O(PLD)$ per step, where L is the number of factors and D the factor dimension.

MAP vs. RSGLD. MAP estimation uses Adam in the spatial chart and converges in $\sim$5 000–9 000 steps on our datasets. RSGLD (§5) adds one parallel-transport operation per entity per step ($O(D)$ for the Lorentz factor) and runs multiple chains; wall-clock cost is roughly 3–5$\times$ that of MAP alone, depending on the number of chains and thinning interval.

Inductive embedding. Embedding a new entity (§5) freezes all global parameters and optimizes a single z_{new} from that entity's evidence. Cost scales with the new entity's own pair count, not with the size of the training set.

Wall-clock times. On a single CPU node (N=1,000 genes), MAP fitting takes roughly 22 minutes on TCGA and 35 on PPI; population shrinkage is instantaneous. All methods are well within single-node feasibility at these scales.

D PRODUCT-MANIFOLD GEOMETRY

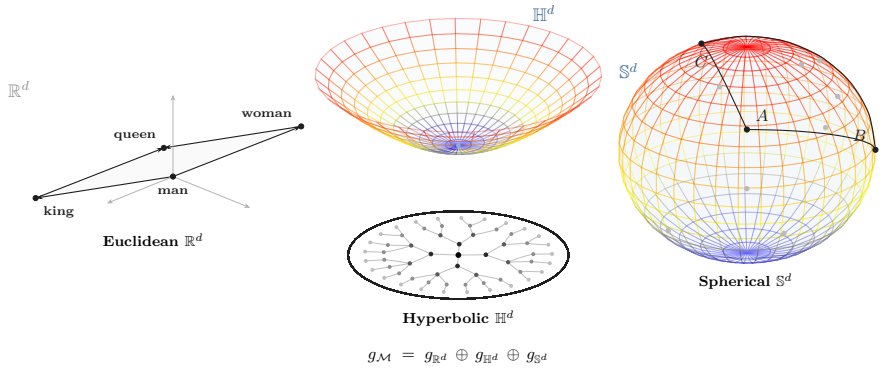


Figure 7: Different data types require different geometries. Euclidean space $\mathbb{R}^d$ captures semantic similarity via vector arithmetic. Hyperbolic space $\mathbb{H}^d$ represents tree-structured hierarchies with exponentially growing volume. Spherical space $\mathbb{S}^d$ encodes directional and compositional data. Product manifolds $\mathcal{M} = \mathbb{R}^{d_1} \times \mathbb{H}^{d_2} \times \mathbb{S}^{d_3}$ with block-diagonal metric $g_{\mathcal{M}} = g_{\mathbb{R}^d} \oplus g_{\mathbb{H}^d} \oplus g_{\mathbb{S}^d}$ provide the representational flexibility required for heterogeneous structured data.

E EXTENDED RELATED WORK

Dynamic and multiplex latent space models. Recent extensions of latent space models to dynamic or multiplex networks (Sewell & Chen, 2015) share our interest in heterogeneous evidence, and work in Euclidean geometry without modeling overdispersion.

Graph structure learning from node features. A separate line of work infers graph structure from node-level features rather than observed edges, from the graphical lasso (Friedman et al., 2008) to neural adjacency learning (Kipf et al., 2018; Franceschi et al., 2019); we share the motivation of not assuming the graph is given, but infer relatedness from grouped co-occurrence counts rather than node features.

Statistical mechanics of hyperbolic random graphs. The Fermi–Dirac link (Eq. 6) originates in the statistical mechanics of random graphs, not in statistical estimation. Krioukov et al. (2010) derive $\sigma((R - x)/T)$ as the connection probability of an equilibrium ensemble on the hyperbolic disk, identifying distance with energy and (R, T) with chemical potential and temperature, in order to explain why real networks have heavy-tailed degree distributions and strong clustering. The object is the ensemble and the aggregate statistics it reproduces; the associated inference problem, recovering coordinates from an observed adjacency matrix (Papadopoulos et al., 2015), takes the graph as given. Beyond the functional form of the link, the settings differ. We estimate a graded per-pair propensity, with a posterior, from overdispersed grouped counts. The link reaches us through its adoption as a decoder in graph embedding (Chami et al., 2019).

Co-occurrence testing in cancer genomics. The applied literature on somatic co-mutation asks a hypothesis-testing question rather than an estimation one: is a gene pair's observed co-occurrence incompatible with independence? DISCOVER (Canisius et al., 2016) supplies the null we compare against, replacing a fixed marginal-frequency assumption with one that conditions on per-tumor mutation burden and per-gene rates. We take it as the baseline of record (§7.2) because it is the strongest available account of what marginal structure alone explains. Its output is a test statistic rather than a rate, so we shrink its per-pair estimates to place them on the same footing as ours; our target is a calibrated propensity for every pair, including pairs never observed together, which a test of independence does not provide.

Bayesian phylogenetics and hyperbolic tree inference. Bayesian phylogenetic inference recovers latent branching structure from grouped biological evidence, placing a posterior over tree topologies

and branch lengths (Ronquist et al., 2012; Drummond et al., 2012; Moretti et al., 2021), and recent work represents phylogenies as points in hyperbolic space (Mimori & Hamada, 2023; Chen et al., 2025). The same formulation extends beyond biology: Yang et al. (2024) infer a posterior over latent tree structure in particle-physics jet reconstruction under a different generative model, via variational pseudo-marginal combinatorial SMC. Our setting is structurally parallel, and the difference is in the support of the posterior rather than whether one is formed: these methods place mass over discrete topologies, so relatedness between two taxa is mediated by a tree path, whereas we place mass over continuous positions and read off a propensity for every pair directly. This accommodates the non-treelike structure typical of co-mutation and protein-interaction data, which no topology in the support represents.

Downstream learning in product manifolds. An embedding of the kind fit here is a set of points in mixed curvature, and predictive models consuming it must operate in that geometry rather than in a tangent-space approximation. Decision-tree and random-forest algorithms have been extended to the hyperboloid (Chlenski et al., 2024) and to product manifolds (Chlenski et al., 2025), and provide one route from the positions inferred in §5 to supervised prediction.