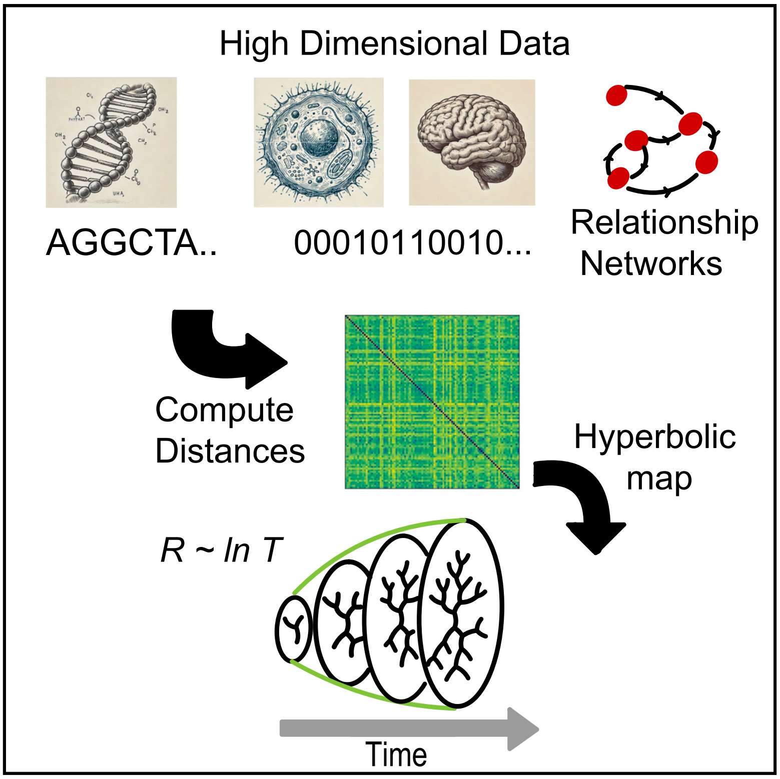
iScience

Article

Adaptive data embedding for curved spaces

Graphical abstract



Highlights

- We developed a Bayesian model for embedding high dimensional data in hyperbolic space
- The model infers embedding coordinates, dimension, and geometric curvature
- We detect logarithmic growth in the geometry of viral evolution in COVID-19
- We geometrically quantify the impact of vaccination through a contraction of the space

Authors

Anoop Praturu, Tatyana O. Sharpee

Correspondence

sharpee@salk.edu

In brief

Complex systems; Computational mathematics; Computing methodology

Praturu & Sharpee, 2024, iScience *27*, 111266
December 20, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.111266

CellPress
OPEN ACCESS

Article

Adaptive data embedding for curved spaces

Anoop Praturu[1,2] and Tatyana O. Sharpee[1,2,3,*]
[1]Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA, USA
[2]Department of Physics, University of California, San Diego, La Jolla, CA, USA
[3]Lead contact
*Correspondence: sharpee@salk.edu

https://doi.org/10.1016/j.isci.2024.111266

SUMMARY

Recent studies have demonstrated the significance of hyperbolic geometry in uncovering low-dimensional structure within complex hierarchical systems. We developed a Bayesian formulation of multi-dimensional scaling (MDS) for embedding data in hyperbolic spaces that allows for a principled determination of manifold parameters such as curvature and dimension. We show that only a small amount of data are needed to constrain the manifold, the optimization is robust against false minima, and the model is able to correctly discern between Hyperbolic and Euclidean data. Application of the method to COVID sequences revealed that viral evolution leaves the dimensionality of the space unchanged but produces a logarithmic increase in curvature, indicating a constant rate of information acquisition optimized under selective pressures. The algorithm also detected a contraction in curvature after the introduction of vaccines. The ability to discern subtle changes and structural shifts showcases the utility of this approach in understanding complex data dynamics.

INTRODUCTION

Hyperbolic geometry has gained traction recently as a powerful framework for understanding complex hierarchies in both machine learning and the basic sciences. Hyperbolic space can informally be thought of as the continuous analog of a tree, and so the exponential expansion of hyperbolic spaces allows them to capture hierarchical structure with only a few degrees of freedom. This has spurred a variety of techniques for embedding taxonomies, networks, and continuous datasets in these spaces.[1–3] For example[4] used hyperbolic embeddings to show that volatile metabolites from plants and animals conform to a low-dimensional hyperbolic geometry. It has also been shown that real world networks such as the internet possess a latent hyperbolic geometry that allows for efficient communication,[2] and[5] has proposed a general framework for understanding how scale-free network topologies arise from networks being embedded in hyperbolic spaces.

Hierarchical structures are typically understood in the form of graphs, so previous representation learning studies in hyperbolic space have focused on embedding explicit networks or taxonomies,[1,3,6,7] where links between nodes determine their geometric similarity. However in many cases explicit hierarchical relationships are not known beforehand, and often the hierarchy cannot be decomposed cleanly into a tree-like graph.[8] Instead, data typically have continuous relationships, more akin to a distance or similarity, than a binary connection. Even in studies that have worked with data in this form,[9,10] there is no clear prescription for determining the curvature or dimension of the underlying space. Both are important geometric parameters for interpreting continuous maps obtained from discrete data. For example, dimensionality can be used to derive a minimal set of independent parameters to describe variations in the data, and curvature can act as a continuous indicator of how hierarchical the data are. This emphasizes the need for an embedding framework that can explicitly fit for the proper curvature and dimension of the hyperbolic space. In particular, complex systems typically have many degrees of freedom, but display a large scale coherence and organization that suggests the dynamics have a reduced “effective” dimension. We seek a systematic treatment of complex systems that allows us to infer their low dimensional structure, and hierarchical connections within it. Applied to COVID19 data, this method produced a geometric insight into constraints in viral evolution.

In this study we formulate the hyperbolic embedding problem within a Bayesian framework for multi-dimensional scaling. Previous studies have investigated Bayesian MDS,[11] but restricted themselves to Euclidean space. While embedding problems are typically stated as the task of minimizing some stress function, we instead formulate the equivalent maximum likelihood problem and re-interpret the stress of classical MDS[12] as a probability distribution. This allows us to incorporate hyperparameters, such as curvature, directly into the model by introducing their prior distributions. We also leverage the “Occam’s Razor” property of Bayesian statistics,[13] and give a simple criteria for unambiguously determining dimension based on the evidence integral.

RESULTS

Hyperbolic geometry

Distances and scaling

Hyperbolic geometry refers to geometric spaces of uniform *negative* curvature. While spaces of zero curvature like a flat

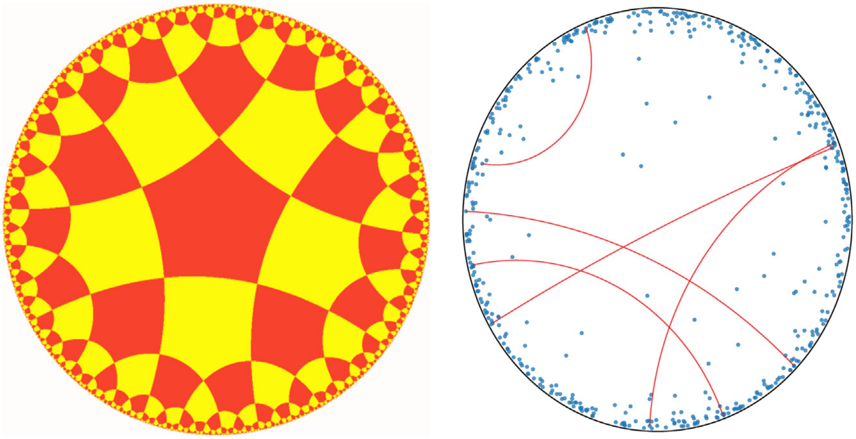


Figure 1. Geometry in the Poincare plane

Left: a tesselation of hyperbolic space in the Poincare plane using pentagons. Note that despite the distortions, each pentagon has the same area and the edges are connected by straight lines. Right: a random sample of points in hyperbolic space, with geodesics plotted between a small subsample of points.

plane are straightforward to visualize, hyperbolic spaces are difficult to represent without distortion due to their exponential expansion. Concretely, the area enclosed by a circle in the space with radius r grows as $\sim e^{\zeta r}$, where $K = -\zeta^2$ is the sectional curvature of the space (as opposed to the familiar πr^2 for flat spaces). This exponential expansion is analogous to the exponential expansion of trees, and is what allows hyperbolic space to model complex hierarchical structures with only a few dimensions. We can visualize the space in the "Poincare" ball where the circle of radius 1 represents the "boundary at infinity." In the tiling of the Poincare plane shown in Figure 1A we see that the tiles get compressed as they stretch out to infinity, but this is just a distortion of the representation. Each of these tiles represents a region of equal area. Note that the *number* of tiles grows exponentially as you move outwards. In the right panel we show a random sample of points in the Poincare ball and some geodesics (shortest paths) connecting them. Notice that the shortest path between two points at large radii curves inwards toward the center. This reflects another familiar property of trees: paths connecting nodes in a tree pass through their nearest common ancestor, traversing the hierarchy up to a higher depth. As such, hyperbolic space encodes hierarchical depth in the radial coordinate.

We now turn to some mathematical details relevant for the present work. In the "native" coordinate representation of hyperbolic spaces the radial coordinate r of a point is equal to its distance from the origin. In this representation we can compute the distance l between 2 points with angular separation $\Delta\theta$ by the hyperbolic law of cosines as follows:

$$\cosh(\zeta l) = \cosh(\zeta r_1)\cosh(\zeta r_2) - \sinh(\zeta r_1)\sinh(\zeta r_2)\cos(\Delta\theta) \tag{Equation 1}$$

where ζ defines the sectional curvature K of the space as $K = -\zeta^2$. This is a direct analog of the familiar Euclidean law of cosines, and as $\zeta \to 0$ this reduces to the Euclidean law, as expected.

Although the straightforward scale invariance of flat spaces is lost, hyperbolic spaces still possess a more subtle form of scale invariance that our MDS algorithm will exploit. Intuitively, spaces exhibit different levels of curvature based on the scale on which you are viewing them (the Earth looks flat from the surface but round from space). We can thus modulate the strength of curvature by changing the scale on which we view the space. Mathematically, we can consider a joint rescaling of the coordinates $r \to \lambda r$ and curvature $K \to \lambda^{-2} K$. By Equation 1 the distance must then be rescaled as $l \to \lambda l$. Thus, unlike Euclidean spaces that can be rescaled simply by scaling the coordinates, scaling hyperbolic spaces must also be accompanied by a rescaling of the curvature. Furthermore, we see that up to an overall scaling of distances, a hyperbolic space with unit curvature and maximum radius R_{max} is equivalent to a space with unit radius and curvature $K = -R^2_{max}$. This allows us to modulate the curvature of our space at fixed radius simply by rescaling our distance matrix, a fact we will exploit in when constructing our Bayesian model in order to adaptively fit for the curvature of our embedding space.

Embedding coordinates

There are many equivalent coordinate systems to describe hyperbolic spaces.[5] Although "compact" projections such as the Poincare or Beltrami-Klein representation can be intuitive for visualization, we follow Nickel and Kiela[6] who found that the Lorentz model behaved significantly better computationally. In this model a D dimensional hyperbolic space is represented by its embedding in a $D+1$ dimensional space Minkowski space using the following constraint equation in the $D+1$ dimensional space:

$$x_0^2 - x_1^2 - x_2^2 \ldots - x_D^2 = 1. \tag{Equation 2}$$

Following conventions from physics we denote points in the $D+1$ space as $x^\alpha = (x_0, \vec{x})$, where $\vec{x}$ is a D dimensional vector referred to as the spacelike component, and x_0 is referred to as the time-like component. In these coordinates the hyperbolic distance between any two points x^α, y^β satisfying the constraints is computed as follows:

$$d_{xy} = \text{arcosh}(x_0 y_0 - \vec{x} \cdot \vec{y}). \tag{Equation 3}$$

Computationally, we take the D space-like components $\vec{x}$ of the coordinates as our free parameters, and compute the time-like component according to the constraint $x_0 = \sqrt{1 + \vec{x} \cdot \vec{x}}$.

Note that these coordinates are for a space with unit curvature $K = -1$. As we will see in the next section, our embedding model will fit for the maximum radius of the distribution of points in this space. Once the embedding is complete, we can rescale distances and coordinates and reinterpret the model as having a different curvature, but for computational simplicity it is preferred to perform the embedding itself at fixed curvature.

A Bayesian model for MDS

We now turn to describing our Bayesian model for hyperbolic MDS, and inferring dimension. We demonstrate that only a small number of points are required to correctly infer the curvature of the space, and our modeling of embedding uncertainty makes the optimization robust against false minima. Finally, we present an iterative algorithm for effectively scaling the optimizer up to large datasets.

The likelihood function

Given a matrix δ_{ij} of distances (dissimilarities) between data points, MDS seeks an embedding of points $\{\overrightarrow{r}_n\}$ in a geometric space whose distance matrix d_{ij} matches the dissimilarity matrix as closely as possible.[12] This is formulated by defining a *stress* function that is minimized when the distances matrices are exactly equal.

$$S\left(\left\{\overrightarrow{r}_n\right\}\right) = \sum_{i<j}(d_{ij} - \delta_{ij})^2. \qquad \text{(Equation 4)}$$

As has been pointed out in previous Bayesian studies,[11] minimizing Equation 4 is equivalent to finding the maximum likelihood of an associated Gaussian likelihood function. Taking this perspective, we seek to construct a generative stochastic model for the data such that the posterior distribution $P(\{\overrightarrow{r}_n\}|\delta_{ij})$ is maximized exactly when the stress is minimized.

By analogy with generative models for linear regression, we model our data dissimilarities as being generated directly from an underlying geometric model by some stochastic process that introduces white noise to the system. Thus we write:

$$\delta_{ij} = \frac{d_{ij}}{\lambda} + \epsilon_{ij}, \qquad \text{(Equation 5)}$$

where the $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij})$ are independent and normally distributed random variables, but with possibly differing variances, and λ is a global scale parameter. Without loss of generality we normalize $\boldsymbol{\delta}$ to have a maximum value of 2 (i.e., unit radius), so that we can view λ as setting the maximum radius of the embedding with unit curvature $K = -1$. Equivalently, as discussed in the hyperbolic geometry section we can interpret λ as setting the curvature of the interior of the unit sphere to be $-\lambda^2$. From the form of the distribution for ϵ_{ij} we can write the conditional distribution $P(\delta_{ij}|\overrightarrow{r}_i, \overrightarrow{r}_j) \sim \mathcal{N}(d_{ij}/\lambda, \sigma_{ij})$. Taking the product over all pairs of points gives the likelihood of the parameters given the dissimilarity matrix:

$$\mathcal{L}\left(\left\{\overrightarrow{r}_n\right\}, \lambda, \{\sigma_{ij}\}\right) \equiv \prod_{i<j} P\left(\delta_{ij}\middle|\overrightarrow{r}_i, \overrightarrow{r}_j\right) = \prod_{i<j}\frac{1}{\sqrt{2\pi\sigma_{ij}^2}}e^{-\frac{1}{2\sigma_{ij}^2}\left(d_{ij}/\lambda - \delta_{ij}\right)^2}. \qquad \text{(Equation 6)}$$

For given values of the parameters λ and constant $\{\sigma_{ij}\}$ the maximum likelihood solution of Equation 6 is equivalent to the minimum of Equation 4. However it is unclear what the optimal values of these parameters should be, and in the case of λ, the actual value of the parameter has geometric implications for the interpretation of the model. Therefore, we take a Bayesian approach and fit for all parameters simultaneously by introducing priors over $\{\overrightarrow{r}_n\}$, λ, and $\{\sigma_{ij}\}$ to compute the posterior:

$$P\left(\left\{\overrightarrow{r}_n\right\}, \lambda, \{\sigma_{ij}\}\middle|\boldsymbol{\delta}\right) \propto \mathcal{L}\left(\left\{\overrightarrow{r}_n\right\}, \lambda, \{\sigma_{ij}\}\right) P\left(\left\{\overrightarrow{r}_n\right\}\right) P(\lambda)\, P(\{\sigma_{ij}\}). \qquad \text{(Equation 7)}$$

The posterior distribution

We start by choosing prior distributions for our parameters that appropriately regularize them without being too restrictive. For our embedding coordinates we would like something like a harmonic oscillator potential that possesses spherical symmetry and prevents points from escaping to extremely large radii. Though we could implement this directly with a normal prior on the radial coordinates, this is a complicated function of the Lorentzian coordinates and could impair the speed and stability of the code. Instead, we use the fact that λ controls the size of the space, so putting a normal prior on λ will have the same effect and is significantly simpler. Note that the log likelihood scales with the number of points as $\sim N(N - 1)/2$, while $\lambda \sim N^0$. Thus, as the number of points in our embedding increases the prior on lambda becomes negligible relative to the likelihood. To remedy this, we multiply the log-prior on λ by $N(N-1)/2$ so that it scales with the likelihood. With this, we can simply leave a flat prior on the Lorentzian coordinates. Although there is a Jacobian distortion of the flat prior when transforming back to the native hyperbolic space, the effect is negligible and the embedding results are unaffected by it.

For the embedding variances, we reduce the number of parameters by introducing an uncertainty σ_n associated to each data point $\overrightarrow{r}_n$. We then compute the uncertainty of the distance between points i and j as follows:

$$\sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2. \qquad \text{(Equation 8)}$$

From a physical perspective, minimizing the stress is equivalent to finding the lowest energy configuration of a collection of points fully connected by springs of equilibrium lengths given by δ_{ij} and stiffnesses $k_{ij} = \sigma_{ij}^{-2}$. Thus the interpretation of Equation 8 is that each point has a characteristic stiffness σ_i^{-2}, and each pair of points is connected by their two springs in series. Physically, this model allows subsets of points that are well fit relative to each other to condense into high-stiffness/low-uncertainty clusters, while poorly fit points have low stiffness and can still easily explore the space. Not only does this help the optimizer (Figure 2, left panel), but it also allows us to identify if points have gotten caught in a false minima and help guide them out of it (Figure 2, middle and right panels). This allows us to iteratively handle problems with false minima when scaling up to large datasets. We complete our model by putting an inverse-gamma prior on each σ_i, as a standard semi-informative prior for Gaussian variances. We found that our results are not sensitive to the choice of the prior on σ, so long as it is not too restrictive.

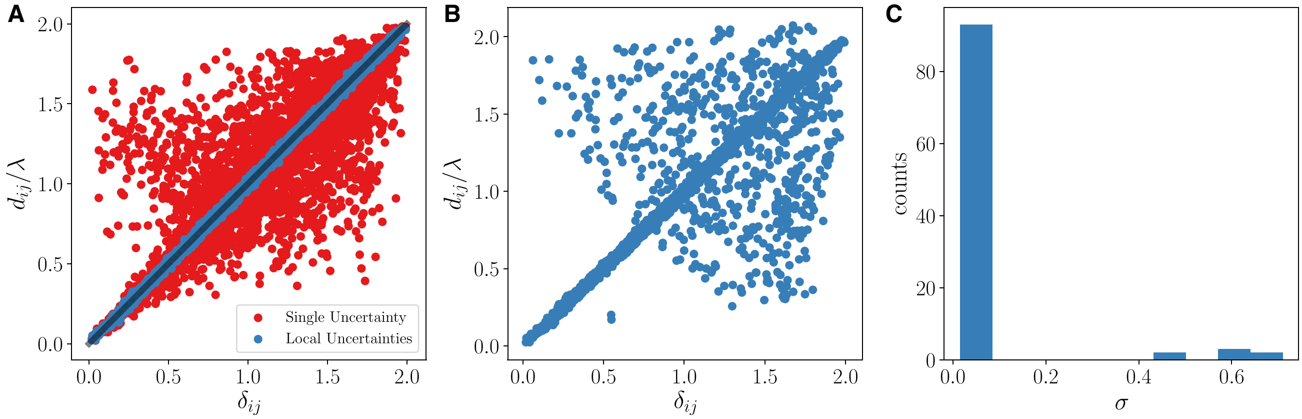


Figure 2. Uncertainty modeling

(A) "Shepard diagrams," plot of d_{ij}/λ versus δ_{ij}, for an embedding in 2D with a single global uncertainty σ in red, and an embedding with individual uncertainties σ_n for each point in blue. The optimal solution should coincide with the line of slope 1 show in gray.

(B) An embedding with local uncertainties caught in a false minimum. Most points are well fit to each other, with a few poorly fit points responsible for the observed scatter.

(C) Distribution of σ values for the false minimum embedding. We can clearly identify a stiff and loose cluster of points. We can randomize the coordinates of points with high σ and re-run the optimizer until all points are at the optimal solution.

Putting all this together, we can write the posterior distribution as follows:

$$-\ln P\left(\left\{\vec{r}_n\right\},\lambda,\{\sigma_n\}\middle|\boldsymbol{\delta}\right)\propto\frac{1}{2}\sum_{i<j}\left(\frac{(d_{ij}/\lambda-\delta_{ij})^2}{\sigma_i^2+\sigma_j^2}+\ln\left(2\pi\left(\sigma_i^2+\sigma_j^2\right)\right)\right)+\frac{N(N-1)}{4\sigma_\lambda^2}\lambda^2+\sum_i\left((a+1)\ln\sigma_i+\frac{b}{\sigma_i}\right).$$

(Equation 9)

This represents the objective function that we seek to minimize by our embedding. The crucial point is that the embedding distance matrix d_{ij} is computed with respect to a hyperbolic metric, and the embedding coordinates $\vec{r}_i$ are the Lorentzian coordinates discussed earlier. We chose $\sigma_\lambda = 10$ and $a = 2$, $b = 0.5$, though we have confirmed that the embedding results are insensitive to the choice of hyper-parameters. We minimize Equation 9 using an L-BFGS algorithm implemented in the Stan statistical package,[14] distributed under a BSD license. We deal with the singularity at $\lambda = 0$ by imposing $\lambda > 0.001$. We confirmed that this value is small enough that the hyperbolic law of cosines with the corresponding curvature gives indistinguishable results from the Euclidean law of cosines, thus this lower bound is small enough to cover the Euclidean case. Coordinate parameters are initialized with a uniform distribution on the interval $[-2,2]$, and the remaining positively constrained parameters are initialized with a uniform distribution on $(0,2]$. We compared this to initializing scale and uncertainty parameters at various fixed values, but found no discernible difference in performance.

Synthetic data results

We first test our method on synthetic distance matrices produced by randomly generating points uniformly in hyperbolic spaces out to some maximum radius. We add noise of magnitude $0.05R_{max}$ to each distance matrix to simulate a more realistic dataset.

To test our method's ability to fit for the underlying curvature of the space we generate points in 3 dimensional spaces with $K = -1$ out to varying maximum radii. We then rescale all of the distance matrices so that their maximum distance is 2 (i.e., radius of 1), and we are thus ignorant as to the true radius of the data. With this rescaling the value of λ predicted by the simulation is exactly the predicted R_{max}, and we can compute the model curvature as $K_{model} = -\lambda^2$ (recall the discussion in the hyperbolic geometry section). Figure 3A demonstrates that we are able to effectively fit for the correct curvature with as few as $25 - 50$ points. In orange we show the results of embedding data generated in a flat space with noise. Our model correctly predicts curvature close to 0, and the embedding matrix matches δ_{ij} almost perfectly. Thus our method *subsumes* traditional MDS methods that only operate in flat spaces. We also show in Figure 3B how the variance in λ_{model} with respect to different random seeds converges as a function of the number of points.

One advantage of a Bayesian approach to MDS is that we can employ Bayesian model selection techniques to determine the underlying dimension of the space. Let $\boldsymbol{\theta}\equiv(\{\vec{\boldsymbol{r}}_{\boldsymbol{n}}\},\lambda,\{\sigma_{\boldsymbol{n}}\})$ denote the set of all parameters. Naively, one could compute the likelihood $\mathcal{L}_D(\boldsymbol{\theta})$ of an embedding in dimension D and choose the dimension that maximizes this value[12]; however, this does not account for volume effects introduced by having a different number of parameters in embeddings of different dimension. Instead one would like to compare models by computing the evidence,[13] which integrates out all parameters and defines the Bayesian information criteria (BIC)[15] as follows:

$$P(\boldsymbol{\delta}|D) = \int P(\boldsymbol{\delta}|\boldsymbol{\theta},D)P(\boldsymbol{\theta}|D)d\boldsymbol{\theta}\approx e^{-\mathrm{BIC}/2},\ \mathrm{BIC} = k\ln(n) - 2\ln(P(\boldsymbol{\delta}|\boldsymbol{\theta}_{MP})),$$

(Equation 10)

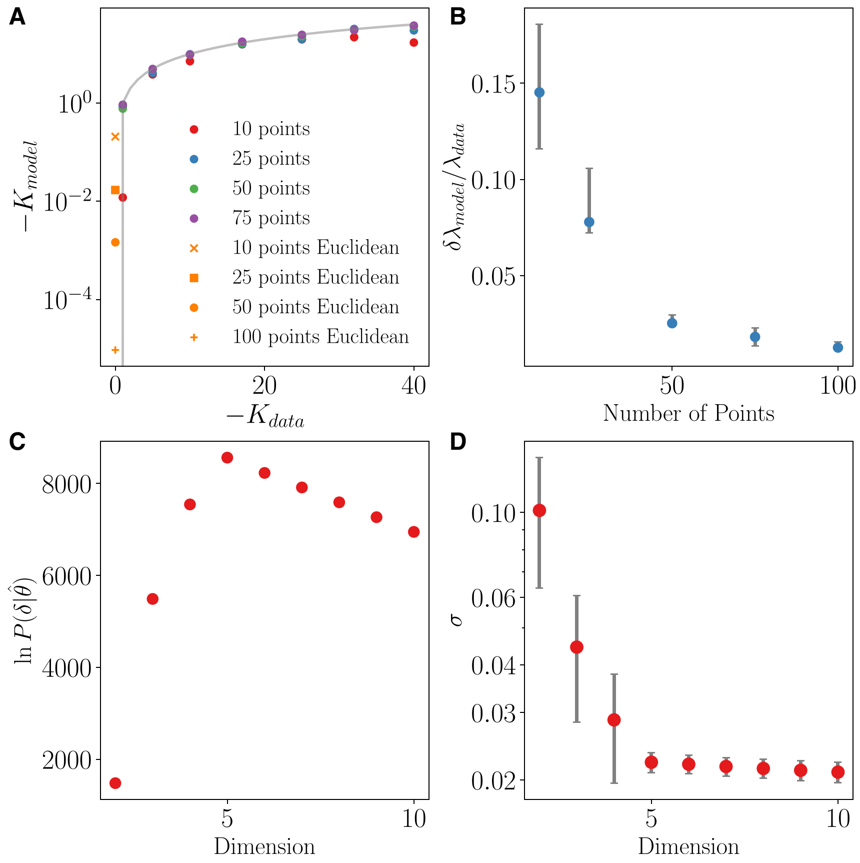


Figure 3. Curvature and Dimension

(A) Predicted curvatures K_{model} plotted against actual curvatures K_{data} for synthetic data. $K_{model} = K_{data}$ line is shown in gray. Predicted curvatures for Euclidean data are shown in orange.

(B) Fractional error in λ_{model} ($\lambda_{data} = 5$) as a function of the number of points. Errors bars encompass the full range of values with respect to simulating multiple times with different random seeds.

(C) Bayesian information criteria versus embedding dimension evaluated on a synthetic dataset of 100 points and true dimension $D = 5$.

(D) Mean σ for the same embeddings. Error bars encompass the full range of σ values for all points in the embedding.

where k is the number of parameters in the model, n is the number of observations, and $\boldsymbol{\theta_{MP}}$ are the values of the parameters that maximize the posterior. This approximate expression for the evidence integral is obtained by evaluating the integral according to the saddle point method. Since the posterior is invariant to rigid rotations of the configuration of points there are $D(D-1)/2$ "redundant" transformations that do not change the probability distribution. We exclude these rotational degrees of freedom when computing the number of model parameters so $k = N(D+1)+1 - D(D-1)/2$.

To test our method's ability to infer the underlying dimension of the space we generate 100 points in a 5 dimensional hyperbolic space, and embed the resultant normalized distance matrix across a range of dimensions. We compute the evidence for all embeddings and plot the result in Figure 3C. For $D<5$ the evidence decreases due to the poor quality of fit, while for $D>5$ the evidence decreases since the quality of fit remains constant but the number of parameters increases. Penalizing additional parameters that do not aid the model is one of the principle features of using the evidence for model selection. Thus we are able to correctly identify $D = 5$ from the maximum of the evidence. When we do not know the ground truth dimension beforehand, it is not as straightforward to select the range of dimensions over which to embed data to construct the BIC curve. For large, costly embeddings, with potentially large dimension, one can typically consider a coarse grained range of dimensions, embedding data in different dimensions in steps of 5 or 10 to build an estimate of the BIC curve over large changes in D. We can use this curve to find the small subrange over which the minimum occurs and sample only this small range densely.

Alternatively, we can analyze the distributions of the σ values of the embeddings, shown in Figure 3D. There is a clear "elbow" at $D = 5$ where the mean and standard deviation of the distribution drops dramatically. Adding more dimensions gives little to no improvement so we can still identify the correct dimension as $D = 5$. When working with real datasets, however, this elbow can often be less clear so we use the BIC going forward.

We ran two further synthetic tests to demonstrate the robust capability of the method. First, we embedded binary trees of various depths to elaborate on the claim that hyperbolic spaces are the continuous analog of trees. The input distances for these embeddings are the path distances on the tree graphs. In the left panel of Figure 4, we show how the fitted scale parameters λ scales with tree depth. We see a linear relationship between λ and depth, which we expect since tree depth is the graph analog of hyperbolic radius. In the middle panel, we show the Shepard diagrams for the embedding of a tree of depth 6 in both hyperbolic and Euclidean spaces of the same dimension. The hyperbolic fit is extremely tight with little to no distortion, as expected. The Euclidean embedding, by contrast, has significantly more scatter and scale dependent distortion, thus further demonstrating how explicit hierarchies can be well fit in a hyperbolic space, while failing to be fit into a Euclidean space.

To show the effectiveness of the algorithm beyond the case of uniform distributions we introduce the following model for testing correlated distributions of points. Consider a multivariate Gaussian distribution in $D = 5$ with a covariance matrix C_{ij} that is a unit diagonal along the first three dimensions: $C_{ij} = \delta_{ij}$ for i, $j \leq 3$. We introduce correlations along the final two dimensions by setting $C_{45} = C_{54} = b$ and $C_{44} = C_{55} = \sqrt{1+b^2}$. In this model, b parameterizes the strength of correlations while enforcing that C_{ij} has unit determinant. If we consider the $D = 5$

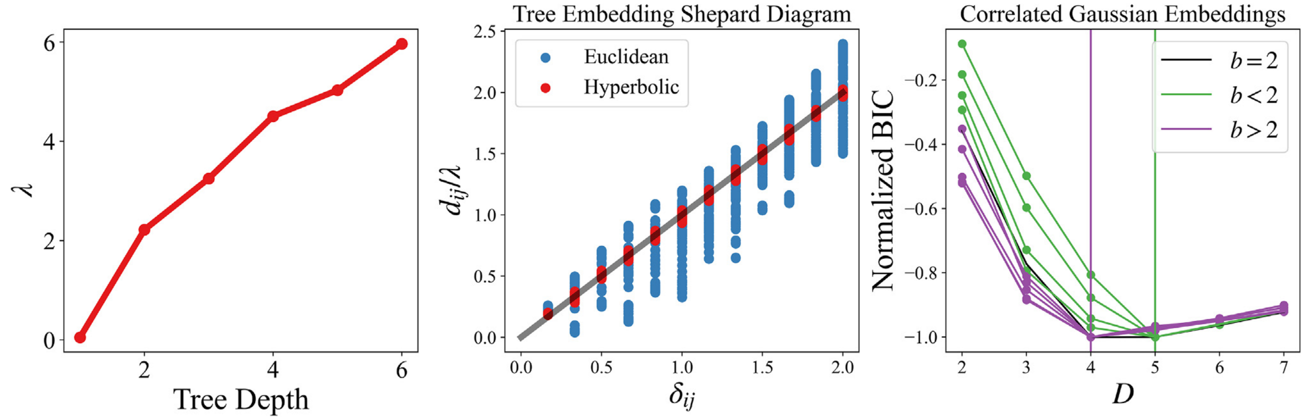


Figure 4. Nonuniform embeddings

Left: scale parameter λ as a function of tree depth for hyperbolic embeddings of binary trees. Middle: Shepard diagrams for both Euclidean (blue) and hyperbolic (red) embeddings of a binary tree with a depth of 6. The quality of the hyperbolic embedding is far superior due to the exponential expansion of the curved space. Right: BIC curves for increasingly correlated distributions of points generated in a $D = 5$ hyperbolic space. There is a transition when the correlations become strong enough that one of the dimensions becomes redundant and BIC selects a lower dimensional description.

Euclidean space as the tangent space at the origin of a 5 dimensional hyperbolic space, we can map correlated distributions of points into the hyperbolic space using the exponential map.[6] For our experiment we generated correlated distributions of points in a $D = 5$ hyperbolic space over a range of b values from 0 (no correlation) to 4 (high correlation). We added a small amount of Gaussian noise to the resultant distance matrices and embedded them over a range of dimensions to see how well the embeddings would perform in the presence of correlations. We found that despite correlations, the embedding quality was still excellent across the range of b values. When fitting for the dimensionality of the data using BIC we saw a crossover around $b = 2$. When the correlations become very strong, relative to the added noise, the distribution of points has effectively lowered its dimension since the low degree of variance along the correlated dimension is indistinguishable from the noise present in the system. In the right panel of Figure 4, we correctly see that the minimum of the BIC curves transitions down to $D = 4$ at the critical correlation value.

Techniques for embedding large datasets

The highly non-linear nature of hyperbolic spaces that endows them with the geometric power to capture complex hierarchical relationships also makes them very difficult to deal with numerically. We describe here an iterative procedure for scaling up the BHMDS algorithm to embed arbitrarily large datasets.

The primary difficulty with large datasets is that for random initial conditions the gradient of each contribution to the cost function tends to be very large. Summed over $\sim N^2$ points this can easily lead to an overflow. Our solution is to allow the simulation to find the “optimal” initial condition so that initial gradients are minimal. To do this, suppose that a subset $N_{seed} < N$ points have been embedded in a D dimensional hyperbolic space and we would like to add one more point (or a batch of points, see the following text) to the embedding. Since $N > N_{seed} > D$ the distances between the new point and the existing N_{seed} points are in theory enough to constrain the position of the new point. So, we freeze the N_{seed} points in place and find the position of only the new point based on its distances to the N_{seed} frozen points. Since the data distance matrices are noisy we expect this to only be approximate, so we think of this procedure as finding the optimal initial condition for the new point. We finish by unfreezing and fully coupling all points as if it were standard MDS and continue optimizing the cost function until convergence.

The essence of our iterative algorithm is to scale this procedure up by adding points in batches of 100 at a time, instead of one at a time. We do this as follows.

- Start with random subsample of N_{seed} points out of the full N points and embed them according to the standard BHMDS algorithm. We typically take $N_{seed} = 300$, but any number small enough that it can be manageably embedded will work.
- Select 100 new points that have not been embedded yet. With the N_{seed} points frozen in place we “initialize” each of the 100 new points individually in the manner described previously.
- Unfreeze all points and optimize the full cost function of $N_{seed} + 100$ points until converged.
- Update $N_{seed} = N_{seed} + 100$ and continue adding points in batches of 100 until $N_{seed} = N$.

The procedure is flexible in the explicit choices of 300 points for the initial embedding and adding points in batches of 100. We found success embedding up to $5,000$ points with this method, though we have not yet fully probed the upper limit of what this algorithm can embed.

In Figure 5 we show the experimentally obtained computational complexity of both the standard BHMDS algorithm and the large scale embedding method described previously. The standard approach exhibits an N^2 scaling as expected, since each iteration of the optimizer must loop over all $\sim N^2$ pairwise distances. The limiting factor in this case is stability; as mentioned previously, the large gradients that arise in the large

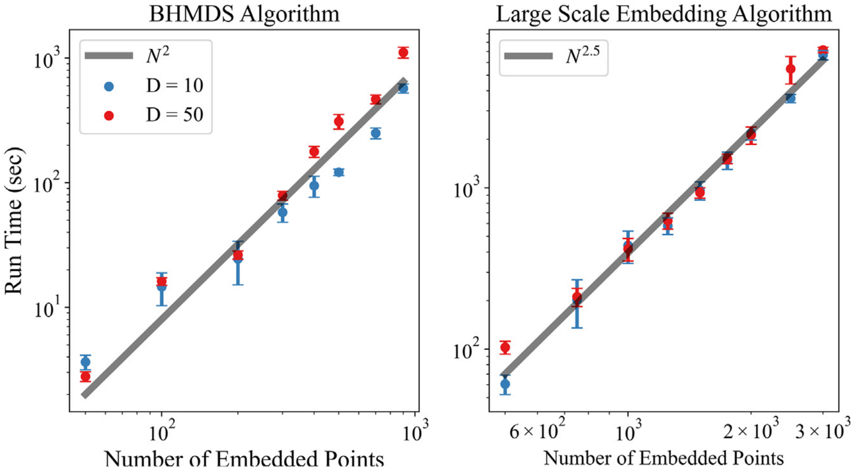


Figure 5. Complexity plots

Left: computational complexity of the BHMDS algorithm for embeddings in dimensions 10 and 50. Run times have the expected N^2 dependence. Right: complexity of the large scale embedding algorithm. The overhead induced by the multiple stages of optimization results in a scaling well fit by $N^{2.5}$ in both cases. The increase in computational cost is compensated for by increased stability that allows the embeddings to work with larger N. Error bars show standard deviation of run times with respect to multiple trials with different randomly generated data.

N limit cause the optimizer to crash often. The large scale embedding method exhibits a steeper scaling, well fit by $N^{2.5}$. This additional cost makes sense due to the overhead induced by having to perform optimization in multiple stages. This additional cost is compensated for by improved stability: the large scale embedding technique allows us to push to much higher ranges of data where the limiting factor is no longer stability.

Illustrative examples

We now consider two example problems with real world data from broadly different fields to illustrate the power and versatility of the Bayesian hyperbolic MDS. First we consider the WordNet hypernym tree, to demonstrate the advantages of the proposal hyperbolic embedding compared to previous hyperbolic and Euclidean embedding methods. Our second example studies the hierarchical nature of viral evolution from a geometric perspective, and enables us to elucidate a constraint on the dynamics of viral evolution in geometric terms.

The geometry of WordNet

WordNet is a massive lexical database encoding the semantic relationships between words. The inherently hierarchical nature of language has made this an excellent candidate dataset for previous hyperbolic studies.[1,6] WordNet thus provides the opportunity to directly compare our embedding method to previous works. We work with WordNet's hypernym tree of "is-a" relationships. For direct comparison with other works we look specifically at the "mammal" subtree, consisting of 1,170 words and generate a distance matrix by computing the graph distance between words in the hypernym tree. We can now seek a low dimensional representation of data based on their distance in order to significantly compress the $\sim N^2$ bits needed to represent the full adjacency matrix.

We show the results of our embedding in Figure 6. Computing the BIC gives the explicit prediction for a three dimensional model for the WordNet graph. Note this improvement over previous methods[1] that cannot unambiguously select a single optimal dimension, and instead must simultaneously analyze results over a range of dimensions, and could not prevent overfitting provided by selecting a higher dimensional embedding. In the left panel we plot the Shepard diagrams for three dimensional hyperbolic and Euclidean embeddings. Not only does our Bayesian model predict a strong curvature of $\lambda = 6.8$, but by direct comparison we see that the hyperbolic embedding is able to far better fit the data with the same number of degrees of freedom. Note how although information about the network connections are only implicitly given to the algorithm through the distance matrix, we can see from the visualization of the embedding in the middle panel that the network topology conforms to the geometry extremely well. In the right hand panel, we show how the radial coordinate of the embedding encodes the hierarchical structure in the data. We define the "specificity" of a word as the number of levels of hierarchy it is removed from the root hypernym, and show the very clear trend of increasing specificity with radius. The example words shown in the plot make clear why we call this quantity "specificity." As the radius increases the words represented transition from very broad categories to very specific examples. This allows us to assign a single scalar value to each word to quantify its linguistic specifying power without needing knowledge of the entire network topology.

The geometry of viral evolution

For our second example we analyze the hierarchical nature of viral evolution through the lens of hyperbolic geometry. The conception of evolution as a vast branching tree was immortalized early on by Darwin with his depiction of a "Tree of Life."[8] Based on this analogy, we theorize that hyperbolic geometry can be used to effectively map out viral evolution. To study this quantitatively we use the database of COVID-19 gene sequences provided by the NCBI.[16] We seek to geometrically quantify the pace, progress, and dynamics of evolution by analyzing structure found in large scale embeddings of COVID-19 sequences, as well as comparing the geometry recovered from embeddings of sequences collected over different timescales.

To compile our dataset we take a random sample of 1,000 COVID sequences sampled uniformly in time between January 1st, 2020, and October 1st, 2021. We measure the distance between sequences by counting the number of nucleotide positions in which two gene sequences *disagree*, also known as the Hamming distance. We embed this distance matrix over a

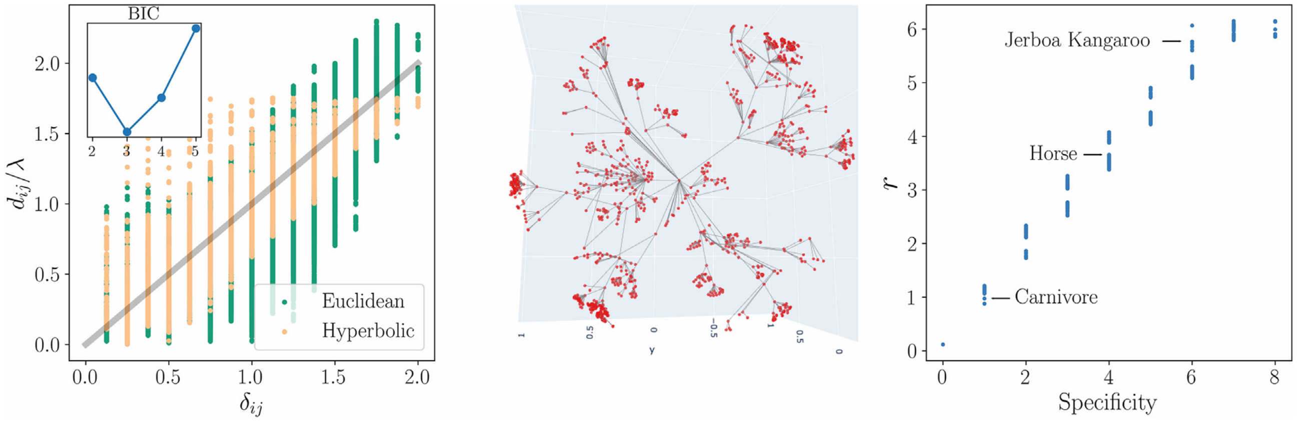


Figure 6. Embedding the WordNet graph in hyperbolic space

Left: Shepard diagrams for hyperbolic and Euclidean embeddings of the WordNet mammal subtree. In the inset we plot Bayesian information criteria for the hyperbolic embeddings to determine the optimal embedding dimension of $D = 3$. Middle: visualization of the optimal hyperbolic embedding in the 3D Poincare model. Each red point represents a single word in the network, and the connections are shown in gray. Right: specificity vs. radius, with select example words shown. Words transition from broad categories to specific labels as radius increases.

range of dimensions using the large scale embedding algorithm. A BIC analysis (Figure 7, left panel) predicts a significant compression of the $\sim 10^4$ nucleotide sequences down to a 40-dimensional space. The resultant embeddings are strongly hyperbolic, with a predicted maximum radius of $\lambda = 7.95$ in $D = 40$. We can also confirm this by comparing the Shepard diagrams of hyperbolic and Euclidean embeddings in the optimally predicted dimension. From the inset of the left panel we can clearly see the hyperbolic embedding gives a better fit to the data than Euclidean embedding. A temporal hierarchy is immediately revealed by the embedding: in the middle panel we see that the hyperbolic embedding radius scales with the date that the sequences were collected. This suggests that we are seeing an evolutionary hierarchy unfolding in time in the hyperbolic space.

We also seek to move beyond the analyses of individual embeddings, and ask what questions can be answered by comparing the geometry of multiple embeddings. To test the hypothesis that hyperbolic geometry encodes information about the hierarchical depth of evolution we analyze samples of viruses taken over time windows of varying length. We hypothesize that since longer time windows allow for more mutations to push the evolutionary tree further down the hierarchy, we expect to see embeddings over longer time windows to have stronger hyperbolic curvature. To test this, we take samples of 500 COVID sequences sampled uniformly starting from March 2020 over time windows of lengths 1 week, 1 month, 3 months, and 6 months. We show the results of these multiple embeddings in the right panel of Figure 7. In the inset panel we show the BIC curves for fitting the optimal dimension for each time window

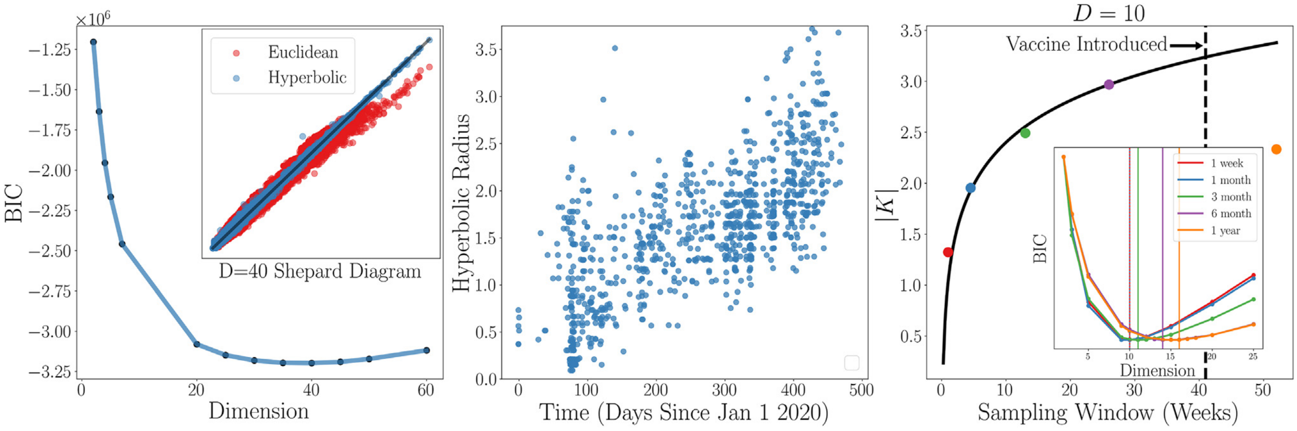


Figure 7. Viral evolution

Left: BIC curve and Shepard diagrams for large embedding of $N = 1,000$ COVID sequences. BIC predicts an optimal dimension of $D = 40$, and from the shepard diagrams the hyperbolic embedding fits the data in $D = 40$ much better than Euclidean embeddings. Middle: hyperbolic radius increases with time of collection date. The evolutionary hierarchy unfolding in time is sorted along the radial axis in the hyperbolic embedding. Right: fitted curvatures and dimensions for embeddings of $N = 500$ points sampled over time windows of increasing length. Curvature grows logarithmically with time window length, while dimension stays roughly constant.

embedding. While the dimension only weakly increased with time window length (and remained within the error bars of the BIC estimate), the main effect of increasing the sampling window size was on curvature (middle panel). Importantly, the curvature increase was logarithmic (in orange we show the best fit logarithmic curve). This is interesting because a logarithmic dependence on time describes the maximum entropy rate of by a discrete Poisson process.[17,18] This suggests that viral evolution is following a strategy that maximizes information acquisition in time. Of course, viral evolution is not a fully random process, being subject to natural selection. Instead, these results indicate under natural selection (prior to vaccine introduction) follows the maximally random process with a rescaled time constant that quantifies selective pressure on viral evolution.

DISCUSSION

We have presented a Bayesian method for embedding data in hyperbolic spaces, with an improved approach to uncertainty modeling, as well as probabilistic techniques for inferring the curvature and dimension of the underlying space. We established through tests on synthetic datasets that the method is both accurate and efficient: the algorithm consistently reconstructs the data in space with high fidelity, and can correctly infer the geometric parameters of the space with very little data. We emphasize the ability of our model to both fit data to geometry, through MDS embedding, and the ability to fit geometry to data, through the Bayesian inference of geometric hyper-parameters. On real datasets from complex systems, the Bayesian hyperbolic method show vast improvements over Euclidean embeddings and uncover insights about the hierarchical nature of the data. We also emphasize that embeddings allow us to infer the underlying hierarchy in the data in a continuous manner, and thus can afford more power and flexibility than discrete hierarchical clustering algorithms.

We also mention Liu et al.,[19] who simultaneously released an alternative approach to Bayesian MDS in hyperbolic space. Although similar in spirit, our approach differs in a variety of ways. For example, in the study by Liu et al.,[19] they impose explicit priors on the hyperbolic coordinates, while we only regularize the scale parameter itself. They also model the embedding with a single global uncertainty parameter σ, while we take a more granular approach and allow each point to have its own uncertainty. Another notable difference is that our approach is based on optimization of the exact posterior, while theirs is based on sampling of an approximate posterior. Finally, our method includes the fitting of a scale parameter, which as we discussed allows us to effectively fit for the curvature of the embedding space and subsumes traditional Euclidean MDS.

Of notable significance are our findings concerning the hyperbolic geometry of viral evolution. We found that with time, the latent manifold maintained its dimensionality while its size (relative to inverse curvature) increased logarithmically with time. This type of dynamics is what is expected for a maximum entropy Poisson process with a constant rate. The rate is presumably set by selection pressure. In future studies, it will be important to test the manifold properties against different types of viruses and other pathogens. If it can be established more broadly that the geometry of pathogenic mutations follows a low-dimensional hyperbolic geometry, then this finding could serve as an organizing principle for testing the optimality in the immune system and its function. Notably, upon vaccine's introduction, the size of the hyperbolic embedding decreased abruptly, indicating reduced complexity. This reduction allows for quantifying the vaccine's effectiveness in slowing down viral evolution. These results offer a tangible method for assessing the impact of interventions such as vaccines on viral evolutionary dynamics.

Limitations of the study

When using hyperbolic geometry to model hierarchical data we make the implicit assumption that the underlying hierarchy can be approximated by a uniform b-ary tree (some degree of loops can be tolerated[5]). Real systems are of course much more complex, with branching factors varying with both depth and direction. A proper model of such systems must allow for dynamically varying curvature and its accompanying geometric complexities. While hyperbolic geometry is certainly a better model for hierarchical data than Euclidean embeddings, one must exercise caution when interpreting to what degree they have captured the hierarchical structure in the data. Additionally, the non-linearities induced by curvature that endow hyperbolic spaces with their enhanced modeling capacity also induce severe computational complexity. The resultant optimization problems are much harder and scaling to large datasets $\sim 10,000$ poses a significant computational challenge that will require new algorithmic techniques.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Tatyana O. Sharpee (sharpee@snl.salk.edu).

Material availability

N/A.

Data and code availability

- The WordNet dataset is publicly available for download at https://wordnet.princeton.edu/. COVID sequence datasets are publicly available for download at https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/.
- All original code for implementation of the BHMDS algorithm can be found at https://github.com/sharpee/BayesianHMDS.
- Any additional information required to analyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The authors thank Prof. Shenshen Wang for discussions on viral evolution and Milo Jolis for help in computing distance matrices between viral sequences. This work was funded by AHA-Allen Initiative in Brain Health and Cognitive Impairment award made jointly through the American Heart Association and the Paul G. Allen Frontiers Group (19PABH134610000); National Science Foundation (NSF) grants IIS-1724421 and PHY-2413080; the NSF Next Generation Networks for Neuroscience Program (award 2014217); National Institutes of Health grants U19NS112959 and P30AG068635, and the Edwin K. Hunter Chair in Neurobiology.

AUTHOR CONTRIBUTIONS

The mathematical model for the study was devised by A.P., who also ran simulations present in the work. The connections between viral evolution and information acquisition were elucidated by T.O.S. The specific synthetic experiments and datasets analyzed were decided jointly by A.P. and T.O.S. Both authors contributed equally to the writing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS

Received: November 27, 2023
Revised: June 3, 2024
Accepted: October 23, 2024
Published: October 28, 2024

REFERENCES

1. Nickel, M., and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. Adv. Neural Inf. Process. Syst. *30*.
2. Boguná, M., Papadopoulos, F., and Krioukov, D. (2010). Sustaining the internet with hyperbolic mapping. Nat. Commun. *1*, 1–8.
3. De Sa, C., Gu, A., Ré, C., and Sala, F. (2018). Representation tradeoffs for hyperbolic embeddings. Preprint at arXiv. https://doi.org/10.48550/arXiv.1804.03329.
4. Zhou, Y., Smith, B.H., and Sharpee, T.O. (2018). Hyperbolic geometry of the olfactory space. Sci. Adv. *4*, eaaq1458.
5. Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguñá, M. (2010). Hyperbolic geometry of complex networks. Phys. Rev. *82*, 036106.
6. Nickel, M., and Kiela, D. (2018). Learning continuous hierarchies in the lorentz model of hyperbolic geometry. Preprint at arXiv. https://doi.org/10.48550/arXiv.1806.03417.
7. Chamberlain, B.P., Clough, J., and Deisenroth, M.P. (2017). Neural embeddings of graphs in hyperbolic space. Preprint at arXiv. https://doi.org/10.48550/arXiv.1705.10359.
8. Simon, H.A. (1962). The architecture of complexity. Proc. Am. Phil. Soc. *106*, 467–482.
9. Klimovskaia, A., Lopez-Paz, D., Bottou, L., and Nickel, M. (2019). Poincaré maps for analyzing complex hierarchies in single-cell data. Preprint at bioRxiv. https://doi.org/10.1101/689547.
10. Cvetkovski, A., and Crovella, M. (2011). Multidimensional scaling in the poincaré disk. CoRR abs/1105.5332 (05).
11. Oh, M.-S., and Raftery, A.E. (2001). Bayesian multidimensional scaling and choice of dimension. J. Am. Stat. Assoc. *96*, 1031–1044.
12. Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika *29*, 1–27.
13. Mackay, D. (1992). Bayesian Methods for Adaptive Models. Doctoral Dissertation (California Institute of Technology).
14. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. J. Stat. Softw. *76*, 1.
15. Schwarz, G. (1978). Estimating the dimension of a model. Ann. Statist. *6*, 461–464.
16. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. (2021). Database resources of the national center for biotechnology information. Nucleic Acids Res. *50*, D20–D26.
17. Bialek, W. (2012). Biophysics: Searching for Principles (Princeton University Press).
18. Zhang, H., Rich, P.D., Lee, A.K., and Sharpee, T.O. (2023). Hippocampal spatial representations exhibit a hyperbolic geometry that expands with experience. Nat. Neurosci. *26*, 131–139.
19. Liu, B., Lubold, S., Raftery, A.E., and McCormick, T.H. (2024). Bayesian hyperbolic multidimensional scaling. J. Comput. Graph Stat. *33*, 869–882.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
WordNet	https://wordnet.princeton.edu/	
Covid Sequence Datasets	https://www.ncbi.nlm.nih.gov/labs/virus/vssi//	
Software and algorithms		
BHMDS Algorithm	https://github.com/sharpee/BayesianHMDS	

METHOD DETAILS

All simulations and generation of synthetic datasets were done with varying random seeds so we could ensure robustness of the method to statistical fluctuations. Error bars when reported show the full range of values of a given variable, except in the complexity scaling plots where the error bars correspond to the standard deviation with respect to multiple trials with differing random seeds.