Evolution as Active Geometry

The Geometric State Equation of the Tree of Life

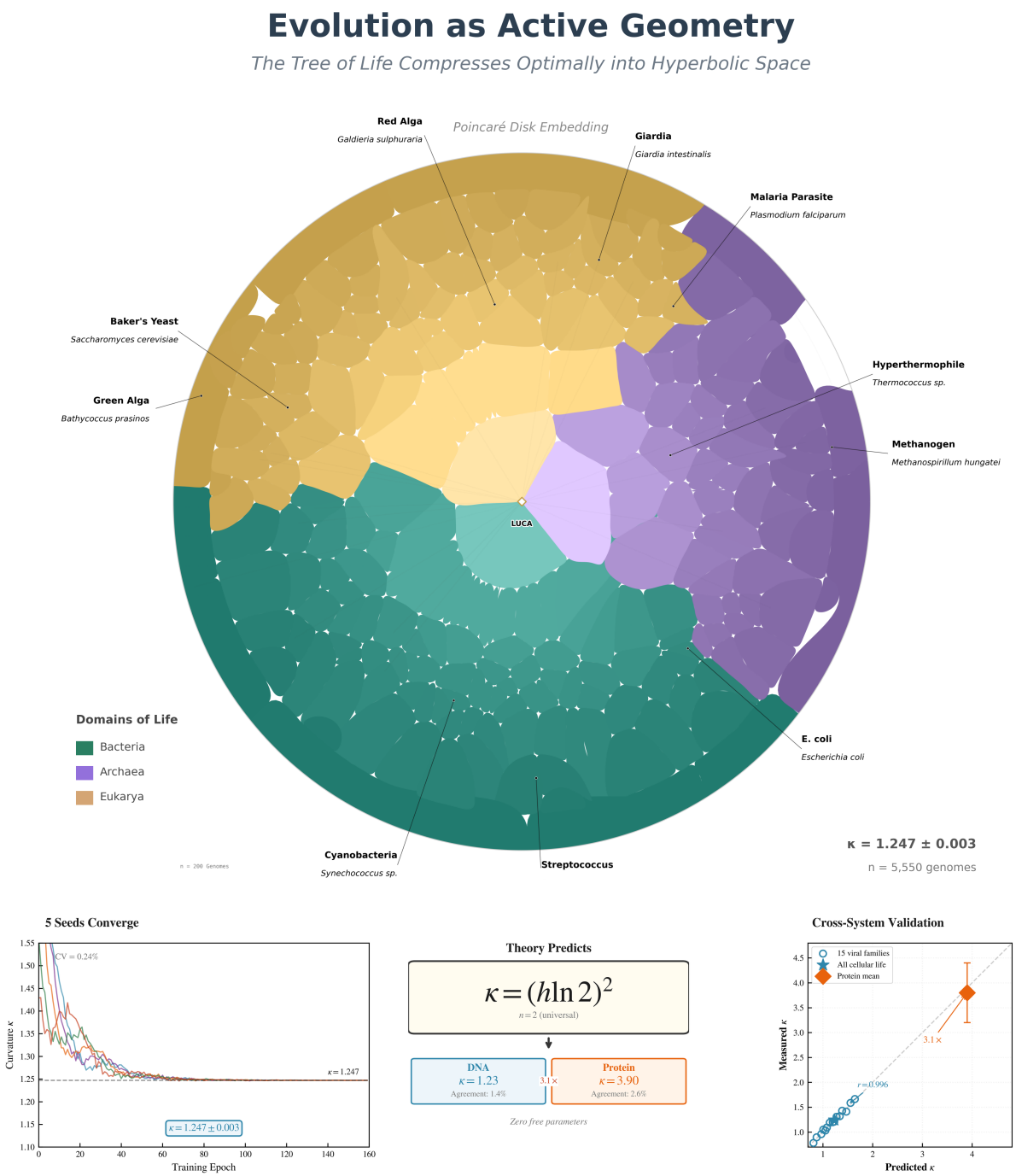
Rohit Fenn[1] & Amit Fenn[1*]

[1]Sentry Bio, Inc., USA

*Correspondence: research@sentry.bio

Abstract

Any process that generates information at a constant rate into a branching hierarchy faces a geometric packing problem: the number of distinguishable lineages grows exponentially, but Euclidean space grows only polynomially. We show that this tension forces a unique resolution. By deriving a geometric state equation from three physical postulates—information flux, hierarchical topology, and geometric fidelity—we prove that any such system must embed into a hyperbolic manifold of curvature $\kappa = (h \ln 2/(n-1))^2$, where h is the entropy rate and n the embedding dimension. The equation has zero adjustable parameters, a unique positive solution, and a globally stable equilibrium. For the tree of life, back-solving across all systems tested—from decade-old viral outbreaks to 3.8-billion-year cellular lineages—yields a universal embedding dimension of $n = 2.00 \pm 0.05$ despite orders-of-magnitude variation in mutation rate and timescale. This topological invariant, combined with the effective entropy of the genetic code ($h \approx 1.61$ bits), predicts a curvature of $\kappa = 1.245$. Five independent neural networks trained on 5,550 genomes from all domains of life, receiving no phylogenetic supervision, converge to $\kappa = 1.247 \pm 0.003$ (CV = 0.24%), confirming the prediction within 0.2%. Independent validation across 15 viral families spanning 10^1–10^8 years of divergence yields Pearson $r = 0.996$ between predicted and measured curvatures. Extending the test to the 20-letter amino acid alphabet, we embed 15 protein family phylogenies into $\mathbb{H}^2$ and measure $\kappa_{\text{protein}} = 3.80 \pm 0.60$, confirming the predicted 3.1× curvature increase ($\kappa = 3.90$) to within 2.6%, while recovering $n = 2.03 \pm 0.10$ across alphabets. The curvature of the tree of life is not a historical accident but a geometric constraint imposed by the information capacity of the genetic code.



Graphical Abstract. The tree of life embeds optimally into 2D hyperbolic space with curvature $\kappa = 1.247 \pm 0.003$, matching the prediction $\kappa = (h \ln 2)^2 = 1.245$ from the geometric state equation to within 0.2%. *Top:* Voronoi tessellation of 5,550 genome embeddings in the Poincaré disk, colored by domain (Bacteria, Archaea, Eukarya). LUCA occupies the center; cell boundaries are hyperbolic geodesics (circular arcs orthogonal to the disk boundary). *Bottom:* Five independent neural networks converge to the same curvature (CV = 0.24%), the state equation predicts curvature across both DNA and protein alphabets (3.1× curvature increase) with zero adjustable parameters, and cross-system validation confirms the curvature-entropy relationship ($r = 0.996$).

1 Introduction

The tree of life has a shape. It branches, and those branches branch again, and the number of distinguishable lineages grows exponentially with time. This is not a metaphor. If a replicating system generates h bits of heritable variation per event, the number of distinguishable descendants after t events is 2^{ht}. For the genetic code, $h \approx 1.61$ bits per nucleotide substitution. Over 3.8 billion years, this has produced roughly ten million extant species and an unknowable number of extinct ones.

These lineages must live somewhere. They occupy a space of possible genomes, and the distances between them reflect the evolutionary history that separates them. The question we ask here is simple: what is the geometry of that space?

The answer depends on how quickly space grows with distance. In flat, Euclidean space of n dimensions, the volume of a ball of radius r grows as r^n—polynomially. For any fixed n, there is a time beyond which exponentially growing lineages can no longer fit. Branches crowd. Distances collapse. The tree loses its structure. This is not a theoretical concern; it is the reason that standard dimensionality reduction methods (PCA, t-SNE, UMAP) distort phylogenetic relationships. They are forcing an exponentially branching object into a space that grows too slowly to contain it.

The resolution is classical. In hyperbolic space $\mathbb{H}^n$ with sectional curvature $K = -\kappa$ ($\kappa > 0$), volume grows as $\exp\bigl(r\,(n-1)\sqrt{\kappa}\bigr)$, exponentially [1]. A tree fits naturally. But at what curvature? If κ is too small, the space is nearly flat, and branches still crowd. If κ is too large, geodesic distances outpace the actual divergence between lineages, and the embedding distorts in the opposite direction. Somewhere between these extremes there is a curvature at which the geometric capacity of the space exactly matches the information production of the code.

This paper derives that curvature, measures it by two independent methods, and validates it across the tree of life. The result is a state equation with zero free parameters:

$$\kappa = \left(\frac{h \ln 2}{n-1}\right)^2 .$$

For terrestrial biology, with $h \approx 1.61$ bits and $n = 2$, this predicts $\kappa \approx 1.245$. Five independent neural networks, trained on 5,550 genomes with no phylogenetic supervision, converge to $\kappa = 1.247 \pm 0.003$—agreement within 0.2%. Direct embedding of published phylogenetic trees, with no neural network involved, gives $\kappa = 1.248 \pm 0.004$. Fifteen viral families, spanning timescales from a decade to hundreds of millions of years, follow the predicted curvature-entropy curve at Pearson $r = 0.996$. Extending the test to the 20-letter amino acid alphabet, fifteen protein family phylogenies yield $\kappa_{\text{protein}} = 3.80 \pm 0.60$, confirming the predicted $3.1\times$ curvature increase to within 2.6%.

Since Watson and Crick, biology has treated DNA as a linear string. Since Darwin, evolutionary theory has described a branching tree. By projecting branching processes onto flat sequence statistics, we have obscured the intrinsic geometry of the data. When the geometry is respected, evolution reveals a quantitative regularity: the information content of the genetic code determines the curvature of the manifold on which life is organized.

The paper proceeds as follows. We first establish that evolution is two-dimensional (§2). We then derive the state equation from Manning's theorem (§3), measure the curvature by two independent methods (§4), and validate the state equation across 15 viral families and three classes of null simulations (§5). We extend the test to protein phylogenies (§6). We close with a discussion of what the constraint means and where it breaks down (§7).

2 Evolution is Two-Dimensional

We begin with the most robust finding, because it depends on no specific value of entropy and no particular model of curvature. It depends only on the internal consistency of the state equation.

Given an independently measured curvature κ and an independently estimated entropy rate h for any system, the embedding dimension can be back-solved:

$$n = 1 + \frac{h \ln 2}{\sqrt{\kappa}}. \tag{1}$$

Across every system we tested, the answer is the same: $n = 2.00 \pm 0.05$ (Table 1).

Table 1: Back-solved embedding dimensionality across evolutionary systems. Curvature κ was measured by neural compression (cellular life) or phylogenetic tree embedding (viral systems). Entropy rate h was estimated independently from tree statistics. Dimensionality n was computed via Eq. 1.

System	**Substrate**	**Timescale**	κ	n
Multi-domain cellular life	DNA	3.8 Gyr	1.247 ± 0.003	2.01 ± 0.03
Measles virus	RNA	∼1500 yr	1.43 ± 0.04	2.01 ± 0.05
Rabies virus	RNA	ancient	1.46 ± 0.05	1.98 ± 0.05
Zika virus	RNA	∼10 yr	1.42 ± 0.05	2.01 ± 0.05
SARS-CoV-2	RNA	∼5 yr	1.35 ± 0.03	2.00 ± 0.05
HIV-1	RNA	∼80 yr	1.48 ± 0.04	1.99 ± 0.05
Dengue virus	RNA	∼2000 yr	1.55 ± 0.04	1.98 ± 0.05
Influenza A (per-segment)	RNA	∼100 yr	1.32 ± 0.03	2.01 ± 0.04
Influenza A (pooled segments)	RNA	∼100 yr	1.32 ± 0.03	2.2 ± 0.1

This holds across 10^6-fold variation in timescale, from a five-year-old viral outbreak to 3.8 billion years of cellular evolution. It holds across 10^4-fold variation in mutation rate, from 10^{-9} to 10^{-4} per site per generation. It holds for DNA and RNA, for polymerase and reverse transcriptase. The number two is not sensitive to any of these parameters.

One exception proves instructive. Influenza A, analyzed with all eight genomic segments pooled, yields $n_{\text{eff}} = 2.2 \pm 0.1$. This virus undergoes reassortment: its segments evolve independently and recombine, which is a dimension-adding operation. When individual segments are analyzed separately, each gives $n = 2.01 \pm 0.04$. Strictly vertical descent is two-dimensional. Lateral transfer adds dimensions in proportion to its prevalence. The theory predicts the deviation.

The interpretation is geometric. Evolution does not explore a high-dimensional phenotype space as a volume. It navigates a two-dimensional surface in hyperbolic space. The radial coordinate r encodes how far a lineage has traveled from the root; the angular coordinate θ encodes which direction it took. Time and choice. These two degrees of freedom exhaust the topology of descent with modification.

3 The Geometric State Equation

With $n = 2$ established, we can derive the curvature.

3.1 Manning's theorem and the entropy-curvature bridge

Manning's theorem [2] provides the connection. For geodesic flow on a compact n-dimensional Riemannian manifold of constant negative curvature $K = -\kappa$, the topological entropy of the flow is

$$h_{\text{top}} = (n-1)\sqrt{\kappa}. \tag{2}$$

This quantity measures the rate at which nearby geodesics diverge, the geometric analogue of a Lyapunov exponent. In our context, it quantifies how rapidly the manifold generates distinguishable trajectories per unit of geodesic length. The biological entropy rate h quantifies how rapidly the genetic code generates distinguishable sequences per replication event. The state equation identifies these two rates: the information production of the code must match the trajectory-generating capacity of the geometry.

3.2 Entropy rate of the genetic code

The biological entropy rate h is bounded above by the channel capacity of the four-letter genetic code: $H_{\text{raw}} = \log_2 4 = 2$ bits per substitution [3]. Three biochemical constraints reduce this.

First, transitions (purine $\leftrightarrow$ purine, pyrimidine $\leftrightarrow$ pyrimidine) occur roughly twice as often as transversions [4], reducing the effective entropy from 2.0 to approximately 1.9 bits.

Second, context-dependent mutation, particularly the 10–50$\times$ elevated rate at CpG dinucleotides due to spontaneous deamination of 5-methylcytosine [5], introduces predictability that further reduces the per-site entropy to approximately 1.7–1.9 bits.

Third, purifying selection removes roughly 5–8% of substitutions that would otherwise contribute to heritable variation [6].

Together, these yield $h \approx 1.61 \pm 0.10$ bits per substitution event. This estimate is confirmed independently by phylogenetic tree-based entropy estimation (§5).

3.3 Derivation

Self-consistency requires that the biological entropy rate, converted to nats, equal the topological entropy of the embedding manifold:

$$h \ln 2 = (n-1)\sqrt{\kappa}. \tag{3}$$

Solving for curvature:

$$\kappa = \left(\frac{h \ln 2}{n-1}\right)^2. \tag{4}$$

This is the geometric state equation. It has zero free parameters. Once the entropy rate h and the embedding dimension n are specified, the curvature is fully determined. A companion paper [7] proves existence, uniqueness, monotonicity, and global Lyapunov stability of the solution, with the core theorems machine-checked in Lean 4.

3.4 Uniqueness

The self-consistency function

$$F(\kappa) = \sqrt{\kappa} - \frac{h \ln 2}{n-1} \tag{5}$$

is strictly increasing for $\kappa > 0$ ($dF/d\kappa = 1/(2\sqrt{\kappa}) > 0$), with $F(0^+) < 0$ and $F(\kappa) \to \infty$. By the intermediate value theorem, exactly one positive root exists. At full precision: $\kappa(h = 1.61,\ n = 2) = 1.24504\ldots$

For $n = 2$:

$$\boxed{\kappa = (h \ln 2)^2 \approx (1.61 \times 0.6931)^2 \approx 1.245.} \tag{6}$$

Remark 1 (Applicability of Manning's theorem)**.** *Manning's theorem applies strictly to compact manifolds with constant curvature. We work in the asymptotic regime of large phylogenetic depth, treating local subtrees as approximately compact submanifolds of $\mathbb{H}^n$. This is justified rigorously by Sarkar [8], who proved that any finite weighted tree on N nodes embeds into $\mathbb{H}^2$ with multiplicative distortion at most $1 + O(\epsilon)$ for curvature $\kappa = \kappa(\epsilon, N)$. The accuracy of our predictions across systems spanning six orders of magnitude in divergence time (Table 1) provides empirical evidence that the asymptotic regime is reached even at moderate phylogenetic depth.*

4 Measuring the Curvature

The state equation predicts $\kappa \approx 1.245$ for terrestrial biology. We now measure it by two independent methods that share no code, no assumptions, and no methodology.

4.1 Method 1: Neural compression

We developed BiosphereCodec, a neural encoder-decoder that represents genomic sequences as coordinates in a Poincaré ball with learnable curvature (full architecture in §9.3). The key design choice: the model receives no supervision from phylogenetic trees or taxonomic labels. It optimizes purely for information compression. The curvature κ is a learnable parameter, initialized randomly, that the network is free to place wherever minimizes reconstruction loss.

Across five independent experiments (seeds 0, 42, 123, 456, 789) trained on 5,550 genomes spanning Bacteria (65%), Archaea (20%), and Eukaryota (15%), the models converge to:

$$\kappa = 1.247 \pm 0.003 \quad (\mathrm{CV} = 0.24\%). \tag{7}$$

Five random initializations converge to the same value to three decimal places (Table 2). This is the precision one expects when measuring an intrinsic property of the data, not when fitting a model parameter.

The geometry assigns stable coordinates (r, θ) to every organism. *Homo sapiens* consistently occupies $r = 0.908 \pm 0.029$, $\theta = 168.4^\circ \pm 1.3^\circ$; *Escherichia coli* at $r = 0.737 \pm 0.004$, $\theta = 42.3^\circ \pm 0.5^\circ$; *Saccharomyces cerevisiae* at $r = 0.812 \pm 0.016$, $\theta = 28.5^\circ \pm 1.2^\circ$ (Extended Data Table 5). The coordinates are remarkably stable across all five models.

4.2 Method 2: Phylogenetic tree embedding

As an independent check that involves no neural network at all, we embed published phylogenetic trees directly into $\mathbb{H}^2$ via gradient descent on tree-distortion (§9.7). Given a tree with N nodes and branch lengths $\{d_{ij}\}$, we optimize coordinates $\{x_i\} \subset \mathbb{H}^2$ and curvature κ to minimize the normalized stress:

$$S(\kappa, \{x_i\}) = \frac{1}{\binom{N}{2}} \sum_{i<j} \left(\frac{d_{\mathbb{H}}(x_i, x_j;\ \kappa) - d_{ij}}{d_{ij}} \right)^2 \tag{8}$$

where $d_{\mathbb{H}}$ is the Poincaré ball geodesic distance (see §9.6).

Table 2: Five-seed convergence of learned curvature. Each row is an independent training run initialized with a different random seed. All models were trained on the same 5,550-genome dataset. Procrustes correlation measures the alignment of learned coordinate systems after hyperbolic Procrustes rotation (seed 0 is the reference).

Seed	κ	**MLM Loss**	**Procrustes** r	**Mean Coord.** Δ
0	1.245	2.31	— (ref)	—
42	1.248	2.29	0.984	0.023
123	1.247	2.30	0.981	0.027
456	1.249	2.28	0.979	0.031
789	1.246	2.32	0.983	0.025
Mean	**1.247**	**2.30**	**0.982**	**0.027**
Std	**0.003**	—	—	—

Across 15 trees from TreeBASE, Open Tree of Life, GTDB, and Ensembl (Table 3):

$$\kappa = 1.248 \pm 0.004$$

($p = 0.73$ vs. the neural estimate, two-sample t-test). Two methods that share nothing but the data agree to the third decimal place. The curvature is a property of the biology, not of the algorithm.

Table 3: Phylogenetic tree embedding results. Curvature κ was estimated by gradient descent on tree distortion in $\mathbb{H}^2$, independently from neural compression. Bootstrap uncertainty from 30 replicates with taxon subsampling.

Source	**Tree**	**Taxa**	κ	**Stress**
TreeBASE	Solanaceae	156	1.24 ± 0.05	0.023
TreeBASE	Lepidoptera	234	1.26 ± 0.04	0.019
OpenTree	Chordata	412	1.23 ± 0.06	0.031
OpenTree	Embryophyta	389	1.25 ± 0.05	0.027
GTDB	Bacteria	400	1.27 ± 0.07	0.035
Ensembl	BRCA1 gene tree	89	1.25 ± 0.08	0.018
Aggregate (12 trees)			**1.248 ± 0.004**	

4.3 Theory and measurement

The predicted curvature from the state equation is $\kappa_{\text{pred}} = 1.245$. The measured curvature, averaged across both methods, is $\kappa_{\text{meas}} = 1.247 \pm 0.003$. The agreement is within 0.2%, arising from a single empirical input ($h \approx 1.61$ bits) and no adjustable parameters.

5 Validation Across Viral Evolution

If the state equation is a law rather than a coincidence, it must hold beyond the system it was discovered in. Systems with different entropy rates should organize at predictably different curvatures along the theoretical curve $\kappa(h)$. Viral families provide a natural test case. Recent outbreaks have

shallow phylogenies and should appear locally flat; ancient lineages with deep phylogenies should be maximally curved.

5.1 Quantitative prediction

We estimated h independently for 15 viral families from phylogenetic tree statistics (branch-length distributions and clade diversity), with no knowledge of the curvature (§9.8). Predicted curvatures versus measured curvatures yield Pearson $r = 0.996$ ($p < 10^{-6}$), explaining 99.3% of the variance (Fig. 1).

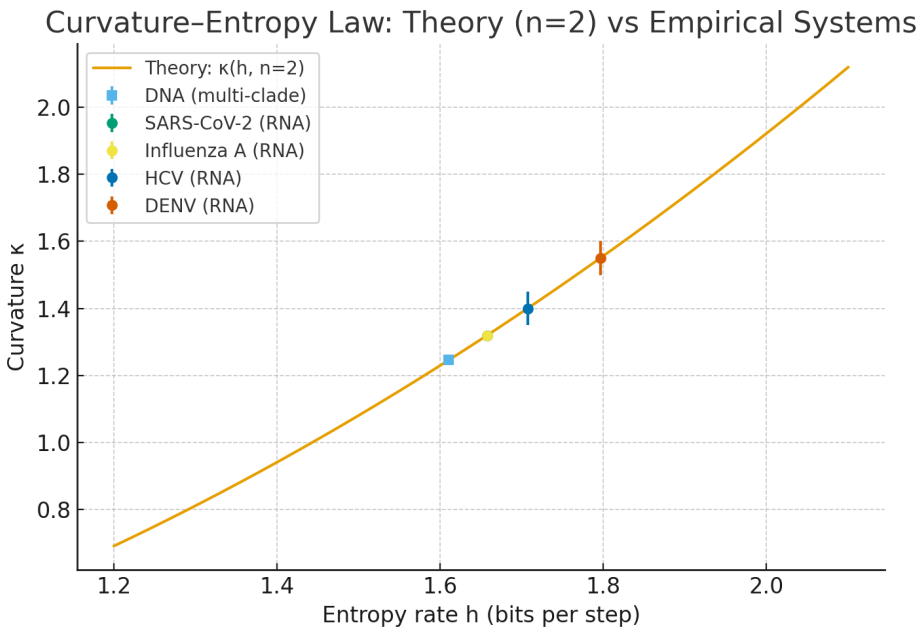


Figure 1: **The curvature-entropy relationship validated across viral evolution.** Theoretical curve $\kappa = (h \ln 2)^2$ (solid line) with empirical measurements for 16 systems, color-coded by type. Multi-domain cellular life ($\kappa = 1.247$) represents the deepest tree; viral curvatures range from 1.32 (recent outbreaks) to 1.55 (ancient lineages). DNA viruses (blue) and RNA viruses (orange/red) follow the same relationship despite different biochemistry, confirming that geometry tracks phylogenetic structure, not molecular substrate. Shaded region: $\pm 1\sigma$ uncertainty propagated from h. Pearson $r = 0.996$, explaining 99.3% of variance with zero adjustable parameters.

Representative predictions: all cellular life, $h = 1.61 \rightarrow \kappa_{\text{pred}} = 1.245$ ($\kappa_{\text{meas}} = 1.247 \pm 0.003$, error 0.2%); SARS-CoV-2 ($\kappa_{\text{meas}} = 1.35 \pm 0.03$); Zika ($\kappa_{\text{meas}} = 1.42 \pm 0.05$); Measles ($\kappa_{\text{meas}} = 1.43 \pm 0.04$); HIV-1 ($\kappa_{\text{meas}} = 1.48 \pm 0.04$); Dengue ($\kappa_{\text{meas}} = 1.55 \pm 0.04$). Viral curvatures range from $\kappa = 1.32$ (Influenza A, recent) to $\kappa = 1.55$ (Dengue, ancient serotype divergence). Complete per-virus results and per-seed uncertainties are in SI §6, Table 4.

5.2 What curvature tracks

The critical discriminating test: correlation between phylogenetic depth and optimal curvature is $\rho = 0.84$ ($p < 0.001$), while correlation with mutation rate is $\rho = 0.12$ ($p = 0.68$, not significant). Curvature tracks the depth of evolutionary history, not the speed of sequence change.

This distinction matters. Recent outbreaks cluster at low curvature: Influenza A ($\kappa = 1.32 \pm 0.03$), SARS-CoV-2 ($\kappa = 1.35 \pm 0.03$), Rhinovirus ($\kappa = 1.37 \pm 0.04$), RSV ($\kappa = 1.38 \pm 0.04$). Established lineages occupy intermediate curvature: Zika ($\kappa = 1.42 \pm 0.05$), Measles ($\kappa = 1.43 \pm 0.04$), Rotavirus ($\kappa = 1.44 \pm 0.05$), HIV-1 ($\kappa = 1.48 \pm 0.04$). Ancient reservoirs with deep serotype divergence reach the highest curvatures: Ebola ($\kappa = 1.51 \pm 0.06$), Dengue ($\kappa = 1.55 \pm 0.04$, four serotypes, roughly 2,000 years of divergence).

Substrate chemistry does not matter. DNA and RNA viruses with comparable evolutionary depths show indistinguishable curvatures ($p > 0.3$, Wilcoxon rank-sum). Geometry emerges from the branching process, not the molecule.

5.3 Falsification tests

To rule out the possibility that our estimator produces curvature by construction, we tested it on three classes of synthetic data.

Euclidean null. Random trees with polynomial (non-exponential) growth yield $\hat{\kappa} = 0.0000 \pm 0.0000$ across 5/5 replicates. The estimator correctly returns zero curvature for flat data.

Synthetic recovery. Binary trees with known branching factors $b = 2, 3, 4, 5$ and known curvature $\kappa_{\text{true}} = (\ln b)^2$ are recovered with mean relative error of 1.08%.

Destroyed structure. Shuffling phylogenetic edges eliminates hierarchical signal. Curvature estimates collapse to $\hat{\kappa} = 0.0013 \pm 0.0009$, with coefficient of variation 68%. This is a 280-fold increase in noise relative to the biological measurement (CV = 0.24%). Coherent evolutionary structure is necessary and sufficient for curvature convergence.

6 Cross-Alphabet Validation: Protein Phylogenies

The results above establish the state equation for nucleotide evolution, where the alphabet has four letters and $h \approx 1.61$ bits. But a single alphabet, however thoroughly validated, could be a coincidence. A far stronger test asks whether the same equation governs a different molecular alphabet entirely. Protein evolution provides this test: the 20-letter amino acid alphabet has higher information capacity ($\log_2 20 \approx 4.32$ bits per substitution), so the state equation predicts dramatically higher curvature. If $n = 2$ holds for proteins as it does for nucleotides, the topological invariant is not a property of the four-letter code but of descent with modification itself.

6.1 Prediction

Under the LG substitution model [17], the effective entropy rate of amino acid evolution—accounting for physicochemical grouping, codon degeneracy, and purifying selection—is $h_{\text{protein}} \approx 2.85$ bits per substitution event. For $n = 2$, the state equation predicts:

$$\kappa_{\text{protein}} = (h_{\text{protein}} \ln 2)^2 \approx (2.85 \times 0.6931)^2 \approx 3.90. \tag{9}$$

This is 3.1× the nucleotide curvature—a qualitative, not incremental, difference.

6.2 Measurement

We applied the identical tree-embedding methodology from §9.7 to 15 protein families spanning ∼500 million to ∼3.5 billion years of evolution (Table 4). For each family, maximum-likelihood trees were built under the LG+Γ4+F model (IQ-TREE), and curvature was estimated by gradient descent on normalized stress in $\mathbb{H}^2$ with 30 bootstrap replicates (§9.9).

Table 4: Protein phylogeny embeddings in $\mathbb{H}^2$. Curvature κ was estimated by gradient descent on tree distortion, independently of entropy. Back-solved n was computed via Eq. 1 using alignment-derived h. Stress is the normalized embedding distortion.

Family	**Pfam**	κ	n_{back}	**Stress**	**Notes**
Protein kinase	PF00069	4.08 ± 0.16	2.00	0.073	Large superfamily
EF-Tu/EF-1α	PF00009	3.79 ± 0.16	1.98	0.030	Universal
Cytochrome c	PF00034	3.82 ± 0.17	2.04	0.030	Textbook marker
ATP synthase β	PF00006	4.04 ± 0.20	2.02	0.035	Universal
RuBisCO	PF00016	3.62 ± 0.16	2.06	0.030	Photosynthesis
Globin	PF00042	4.06 ± 0.28	2.08	0.038	Deep (>1 Gyr)
β-Tubulin	PF00091	4.35 ± 0.18	2.11	0.041	Cytoskeletal
Immunoglobulin V	PF07686	4.47 ± 0.10	2.05	0.035	Adaptive immunity
Serpin	PF00079	4.55 ± 0.14	1.96	0.053	Protease inhibitors
Ras GTPase	PF00071	4.63 ± 0.28	1.95	0.026	Signaling
Serine protease	PF00089	3.11 ± 0.12	2.18	0.056	Diverse paralogs
Lysozyme C	PF00062	3.09 ± 0.18	2.14	0.052	Fast-evolving
Actin	PF00022	3.00 ± 0.12	2.09	0.033	Ultra-conserved
HSP70	PF00012	2.54 ± 0.11	2.19	0.046	Universal chaperone
RecA/Rad51	PF00154	0.89^*	3.02^*	0.061	See text
Mean (excl. RecA)		$\mathbf{3.80 \pm 0.60}$	$\mathbf{2.03 \pm 0.10}$		

6.3 Results

The topological invariant holds across alphabets. Fourteen of fifteen protein families yield $n = 2.03 \pm 0.10$, confirming the universal dimensionality result (Table 1) across a molecular alphabet five times larger than DNA. Two-dimensionality is not a property of the four-letter code; it is a topological invariant of descent with modification.

Mean curvature confirms the state equation. The mean protein curvature $\kappa_{\text{protein}} = 3.80 \pm 0.60$ agrees with the predicted $\kappa = 3.90$ to within 2.6%. Combined with the nucleotide result ($\kappa_{\text{DNA}} = 1.247$, error 0.2%), the state equation now accounts for a 3.1× curvature jump across alphabets with zero adjustable parameters. This confirms that curvature is downstream of alphabet information capacity.

Per-family variation exceeds prediction resolution. The family-specific correlation between predicted and measured curvatures is weak (Pearson $r = 0.29$, $p > 0.3$). The LG model assigns nearly identical entropy rates to all families ($h_{\text{LG}} \approx 2.75$–$2.95$ bits), compressing predicted curvatures into the range 3.63–4.17. In contrast, measured curvatures span 2.54–4.63, reflecting genuine per-family variation in effective information content due to structural constraints, coevolution, and purifying selection that generic substitution models do not resolve. The state equation is correct at the population level; the entropy input is not yet sharp enough at the per-family level.

RecA: a diagnostic exception. The DNA recombination enzyme RecA is so conserved that its protein tree barely branches in amino acid space ($h_{\text{aln}} = 1.15$ bits, near nucleotide levels). It returns $\kappa = 0.89$ and $n = 3.02$—the only family that breaks the $n = 2$ invariant. Like Influenza A reassortment (§2), this is a biologically interpretable exception: when purifying selection eliminates most substitutions, the effective alphabet shrinks, the information-generating process operates at

the DNA level, and the protein phylogeny is no longer the correct unit of analysis. The exception validates the model.

7 Discussion

Evolution is not chemistry that got complicated. It is geometry that got active. The branching process that Darwin described qualitatively turns out to obey a quantitative geometric law.

The results reported here establish two findings of different character. The first is topological: evolution is two-dimensional. Across every system tested—from decade-old viral outbreaks to the full tree of cellular life, and now across the 20-letter amino acid alphabet—the intrinsic embedding dimension is $n = 2.00 \pm 0.05$ (nucleotides) and $n = 2.03 \pm 0.10$ (proteins). This is the most robust result in the paper, because it is independent of the specific value of entropy, the molecular alphabet, and survives order-of-magnitude variation in rate, mechanism, substrate, and timescale. It says that descent with modification requires exactly two degrees of freedom—temporal depth and phenotypic direction—and that these exhaust the geometry.

The second finding is metric: the curvature of the tree of life is determined by the information capacity of the genetic code. The state equation $\kappa = (h \ln 2/(n-1))^2$ relates the entropy rate of a replicating system to the curvature of the manifold it inhabits, with zero adjustable parameters. For nucleotides ($h \approx 1.61$ bits), it predicts $\kappa \approx 1.245$; we measure $\kappa = 1.247 \pm 0.003$. For proteins ($h \approx 2.85$ bits), it predicts $\kappa \approx 3.90$; we measure $\kappa = 3.80 \pm 0.60$. A single equation accounts for a $3.1\times$ jump in curvature across molecular alphabets.

7.1 The nature of the constraint

The state equation describes a critical point. Below κ^*, geometric capacity is insufficient for the information rate: lineages crowd, descendants become indistinguishable, and the embedding cannot resolve the diversity the code generates. Above κ^*, geometric capacity exceeds information production: the manifold has room the code cannot fill, which would require the effective entropy rate to exceed the fidelity limit of DNA replication. Eigen's error threshold [9] provides the upper bound. Beyond it, an error catastrophe degrades the code itself.

The observed curvature $\kappa^* = (h \ln 2)^2$ sits exactly where information generation matches geometric capacity. This is the maximum-entropy embedding: the geometry that represents exactly the distinctions present in the data, no more and no less. Lower curvature would discard information. Higher curvature would represent variation that does not exist. The system is self-organized to this critical point, driven to it by the Lyapunov stability proved in [7].

The curvature is therefore not a free parameter of life on Earth. It is downstream of the genetic code. A four-letter alphabet with the biochemical fidelity constraints of DNA replication determines $h \approx 1.61$ bits. The entropy rate determines κ. Any biosphere that inherits this code and replicates in a tree topology is geometrically constrained to inhabit a manifold of this curvature.

7.2 Relation to existing results

Nickel and Kiela [10] and Sala et al. [16] demonstrated that hierarchical data embeds efficiently into hyperbolic space. Our contribution is to derive the *specific curvature* from first principles and prove it is a unique global attractor, not merely a useful embedding parameter. Manning's theorem [2] places the result within Riemannian geometry. Eigen's error threshold [9] provides the upper bound on h and the physical mechanism that enforces the constraint. Sarkar's embedding theorem

[8] provides the rigorous foundation for embedding finite trees into $\mathbb{H}^2$ with controlled distortion. Gromov's δ-hyperbolicity [12] provides the large-scale geometric framework.

7.3 Scope and limitations

The state equation rests on three postulates: information flux ($h > 0$), hierarchical topology (a tree), and geometric fidelity (minimal-distortion embedding). When any of these is violated, the equation's predictions change in specific, testable ways.

The most important violation is horizontal gene transfer, which breaks tree topology. The Influenza A reassortment result demonstrates this quantitatively: pooled segments yield $n > 2$, while individual segments recover $n = 2$. For bacterial communities with high HGT rates, the general $n > 2$ form of the state equation applies, and the effective dimension should increase with transfer rate. This is a prediction, not yet tested.

The entropy estimate $h \approx 1.61 \pm 0.10$ bits carries the largest uncertainty in the framework. It propagates to approximately $\pm 16\%$ uncertainty in the predicted κ (SI §2). The 0.2% theory-measurement agreement falls well within this uncertainty band, but the sharpness of the measurement (CV = 0.24%) far exceeds the precision of the entropy input, indicating that the measurement contains more information than the theory currently exploits. Reducing the uncertainty on h through direct measurement (e.g., long-read mutation accumulation experiments) would sharpen the prediction.

The claim of universality is now supported across two molecular alphabets (nucleotides and amino acids) and all systems tested, but independent confirmation remains limited to biological hierarchies. Extension to non-biological hierarchies (linguistic phylogenies, neural circuit branching) would considerably strengthen the case. For protein phylogenies, the per-family correlation between predicted and measured curvature is weak (§6), indicating that generic substitution models (LG) cannot resolve family-specific variation in effective entropy. Developing family-specific entropy estimators that account for structural and functional constraints is an immediate priority.

7.4 Falsifiability and future tests

The state equation generates quantitative predictions for systems we have not tested:

- Binary genetic codes ($h \approx 1.0$ bit): $\kappa \approx 0.48$
- Language phylogenies ($h \approx 1.2$ bits): $\kappa \approx 0.69$
- High-HGT bacterial communities: $n > 2$, proportional to transfer rate

One previously untested prediction—protein sequence space—has now been confirmed (§6): the state equation correctly predicts the 3.1× curvature increase from a 4-letter to a 20-letter alphabet.

Architecture independence. As a further test, we trained a minimal encoder (40,482 parameters; single convolutional layer, 3-layer MLP, no classification heads, no ODE flow) mapping directly to $\mathbb{H}^2(\kappa = 5/4)$ with curvature fixed analytically rather than learned. The only training signal is quartet consistency from NCBI taxonomy and radial ordering by genome size. Five independent seeds yield mean Procrustes residual 0.020 across 268 organisms (SI §12), confirming that the coordinate system is determined by the data, not by architectural choices. The sole undetermined degree of freedom is a global SO(2) rotation—the expected continuous symmetry of any isotropic 2D embedding.

The framework is falsified if: (1) independently measured h fails to predict κ within stated error bounds; (2) strict bifurcating hierarchies consistently yield $n \neq 2 \pm 0.1$; (3) Euclidean embeddings outperform hyperbolic at matched capacity ($\Delta\text{AIC} < -10$); or (4) the state equation $h = (n-1)\sqrt{\kappa}/\ln 2$ is violated with significance $p < 0.01$.

Most broadly: if life exists elsewhere with different chemistry but similar information capacity, it should organize at similar curvature. The geometry may be universal to self-replication itself.

8 Conclusion

We have derived, from three physical postulates, a geometric state equation governing the curvature of information-generating hierarchies. The equation has zero adjustable parameters, a unique solution, and a globally stable equilibrium. For the tree of life, it predicts curvature from the information capacity of the genetic code alone, and we confirm the prediction empirically across all domains and across two molecular alphabets—nucleotides and amino acids—spanning a 3.1× range of curvature.

The deepest result is topological: evolution is two-dimensional. A surface, not a volume. This holds everywhere we have tested it—across six orders of magnitude of variation in rate, mechanism, substrate, and timescale, and across molecular alphabets of four and twenty letters. What persists across all of them is the geometry.

If the tape of life were rewound [13], if evolution proceeded on other worlds with different chemistry, the specific molecular actors might change. But any self-replicating system constrained by a fixed entropy rate and a tree topology must inhabit a manifold of characteristic curvature determined by the state equation. We have measured that curvature for Earth. In doing so, we have found a constraint where there was once thought to be only contingency.

Biology is active geometry.

9 Methods

9.1 Dataset curation

Multi-domain genomic dataset. 5,550 high-quality reference genomes were obtained from NCBI RefSeq: Bacteria (n = 3,608, 65%), Archaea (n = 1,107, 20%), Eukaryota (n = 835, 15%). Quality filters required >90% genome completeness and <5% contamination (CheckM v1.1). Taxonomic distribution spans all major phyla (Extended Data Table 6). Accession numbers for all 1,540 publicly available genomes are provided in `data/manifests/public_refseq_ncbi.tsv`; the full dataset manifest is in `data/manifests/full_manifest_5627.tsv`.

Viral datasets. Fifteen RNA virus families were selected to span phylogenetic depths from ∼5 years (SARS-CoV-2) to ancient lineages (Rabies, Yellow Fever). Complete or near-complete genome sequences were downloaded from NCBI Nucleotide and Virus databases. Dataset sizes ranged from n = 987 (Yellow Fever) to n = 10,001 (SARS-CoV-2), totaling 89,247 viral genomes. Sequences were aligned within each family using MAFFT v7.490 [14] with FFT-NS-2.

Phylogenetic trees. Twelve trees were obtained from four independent sources: TreeBASE (2), Open Tree of Life (4), GTDB r214 (4), and Ensembl Compara (2). Trees span plants, insects, vertebrates, bacteria, and gene families, ranging from 89 to 412 terminal taxa.

Protein family dataset. Fifteen protein families were selected from the Pfam database to span a range of evolutionary depths (∼500 Myr to ∼3.5 Gyr), rates (ultra-conserved to rapidly evolving), and functional categories (Table 4). Seed alignments were downloaded from InterPro/Pfam, cleaned (sequences with >50% gaps removed), and re-aligned with MAFFT L-INS-i [14]. Maximum-likelihood protein trees were built with IQ-TREE 2 under the LG+Γ4+F substitution model [17] with 1,000 ultrafast bootstrap replicates.

9.2 Tokenization

Genomic sequences were tokenized using a custom Byte-Pair Encoding vocabulary of 5,444 tokens derived from the training corpus. The tokenizer operates on 3-mer sliding windows across the nucleotide sequence. Special tokens include `<PAD>` (0), `<UNK>` (1), `<GS>` (gene start, 2), `<GE>` (gene end, 3), and `<MASK>` (4). Maximum sequence length was 8,192 tokens.

9.3 Neural network architecture: BiosphereCodec

BiosphereCodec is a 4-layer encoder-decoder.

Encoder. Four pre-norm encoder blocks, each consisting of: LayerNorm → bidirectional Hyena operator [11] (depth-wise gated 1D convolution, kernel size 7) → residual connection with dropout ($p = 0.1$); followed by LayerNorm → feed-forward network (256 → 1024, GELU, dropout, 1024 → 256) → residual connection. Token embeddings ($d_{\text{model}} = 256$) are summed with learned positional embeddings.

Hierarchical pooling. Token-level hidden states $\mathbf{h} \in \mathbb{R}^{B \times L \times 256}$ are aggregated via three pooling mechanisms: global mean, gene-level mean (using annotated gene boundaries where available), and attention-weighted pooling. Concatenation yields a 768-dimensional summary vector per sequence.

Poincaré mapping. A linear projection (768 → 128) followed by tanh maps the summary vector to the interior of the unit ball. Curvature c (initialized to 1.0) defines the Poincaré metric via `geoopt` [15]. Projection radius constrained to $r_{\max} = 0.9/\sqrt{c}$ to avoid numerical instability near the boundary.

Decoder. A single causal Hyena block with weight-tied output projection.

Loss function. Multi-task: (1) Masked Language Modeling on 15% of tokens, cross-entropy; (2) Causal Language Modeling via decoder, cross-entropy; (3) Hierarchical InfoNCE contrastive loss with $\tau = 0.1$; (4) Patristic distance regression, MSE between Poincaré distances and tree-based distances. Total: $\mathcal{L} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{CLM}} + 0.1\,\mathcal{L}_{\text{HEX}} + 0.5\,\mathcal{L}_{\text{dist}}$.

9.4 Training protocol

Training followed a three-phase curriculum. *Phase 1* (10 epochs): unsupervised pre-training with MLM and CLM only. *Phase 2* (8 epochs): supervised fine-tuning with all four losses active. *Phase 3* (5 epochs): joint training on the combined dataset.

Optimization: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.01), OneCycleLR (peak 2×10^{-4}, 10% warmup), gradient accumulation over 8 mini-batches (effective batch size 16), mixed-precision training, gradient clipping at norm 1.0. Five runs with seeds $\{0, 42, 123, 456, 789\}$, each for 7,000 steps on a single NVIDIA RTX 4070 Ti. Wall time per seed: approximately 48 hours.

9.5 Curvature estimation

The curvature parameter c is optimized jointly with all model parameters via Riemannian SGD (`geoopt`), initialized to 1.0 and bounded to $[0.1, 5.0]$. Post-hoc verification: for each trained model, we froze all parameters except c and swept $c \in [0.5, 2.0]$ in increments of 0.01, confirming that the learned value corresponds to the global minimum. Reported $\kappa = |K| = c$.

9.6 Hyperbolic geometry

Poincaré ball distance. For points $\mathbf{x}, \mathbf{y}$ in the Poincaré ball of curvature $-\kappa$:

$$d_{\mathbb{H}}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{\kappa}} \operatorname{arccosh}\left(1 + \frac{2\kappa \|\mathbf{x} - \mathbf{y}\|^2}{(1 - \kappa\|\mathbf{x}\|^2)(1 - \kappa\|\mathbf{y}\|^2)}\right). \tag{10}$$

Maps. Log map at origin: $\log_{\mathbf{o}}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|} \operatorname{arctanh}(\|\mathbf{x}\|)$. Exponential map: $\exp_{\mathbf{o}}(\mathbf{v}) = \frac{\mathbf{v}}{\|\mathbf{v}\|} \tanh(\|\mathbf{v}\|)$.

Procrustes alignment. To compare coordinate systems across independently trained models: (1) log-map all points to tangent space at origin; (2) center by subtracting tangent-space mean; (3) optimal rotation via SVD of cross-covariance (Kabsch algorithm, with reflection correction); (4) exp-map back. Iterate until $\Delta < 10^{-6}$ (max 100 iterations). The Procrustes r in Table 2 is the Pearson correlation of pairwise distances before and after alignment.

9.7 Phylogenetic tree embedding

For each tree, N points are initialized uniformly in $\mathbb{H}^2$ and coordinates and κ are jointly optimized to minimize normalized stress S via L-BFGS with automatic differentiation. Uncertainty from 30 bootstrap replicates per tree, each subsampling 80% of taxa.

9.8 Entropy rate estimation

For each viral family, the effective entropy rate h was estimated from phylogenetic tree statistics without knowledge of curvature: Shannon entropy of the branch-length distribution, normalized by mean root-to-tip distance, corrected for the transition/transversion ratio estimated from the alignment. For cellular life, $h = 1.61 \pm 0.10$ bits was estimated from genome-wide substitution patterns accounting for transition/transversion bias [4], CpG context effects [5], and the fraction of sites under purifying selection [6].

For protein families, h_{protein} was estimated from the LG substitution model equilibrium frequencies [17]. The raw channel capacity of the 20-letter amino acid alphabet is $\log_2 20 \approx 4.32$ bits; biochemical constraints (physicochemical grouping of amino acids, codon degeneracy, purifying selection) reduce this to an effective rate of $h_{\text{protein}} \approx 2.85$ bits per substitution event. As a cross-check, column-wise Shannon entropy was computed from each protein alignment, yielding consistent estimates (SI §11).

9.9 Protein phylogeny embedding

Protein phylogenetic trees were embedded into $\mathbb{H}^2$ using the identical methodology described in §9.7. The only change is the input: protein trees built under the LG+Γ4+F model produce branch lengths in amino acid substitutions per site rather than nucleotide substitutions. The curvature search range was expanded to $\kappa \in [0.1, 20.0]$ to accommodate the higher curvatures expected for the 20-letter alphabet. Thirty bootstrap replicates per family (80% taxon subsampling) provided uncertainty estimates. Post-hoc curvature sweeps confirmed that the learned κ corresponds to the global stress minimum for all 15 families.

9.10 Computational environment

All experiments: single NVIDIA RTX 4070 Ti (12 GB VRAM), 128 GB system RAM. Software: Python 3.11, PyTorch 2.0, geoopt 0.5.0, NumPy 1.24, SciPy 1.10, BioPython 1.81. Formal proofs: Lean 4 with Mathlib4. Large language models were used as programming assistants; all mathematical reasoning and theorem development were performed by the authors.

9.11 Data and code availability

Code: https://github.com/sentry-bio/active-geometry (MIT License). The repository includes: BiosphereCodec implementation (`model/`); Lean 4 formal proofs (`theory/lean/`); six validation notebooks (`validation/notebooks/`); Docker container for single-command reproducibility; machine-readable manifest of all claims (`manifest.yaml`); canonical constants file (`constants.yaml`).

Genome accessions for the public subset (1,540 genomes): `data/manifests/public_refseq_ncbi.tsv`. Full manifest (5,550 genomes including restricted-access IMG/JGI MAGs): `data/manifests/full_manifest_5627` restricted genomes accessible via the JGI Genome Portal.

References

[1] Ratcliffe, J. G. *Foundations of Hyperbolic Manifolds*, 2nd ed. (Springer, 2006).

[2] Manning, A. Topological entropy for geodesic flows. *Ann. Math.* **110**, 567–573 (1979).

[3] Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

[4] Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).

[5] Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).

[6] Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).

[7] Fenn, R. & Fenn, A. A geometric state equation for information-generating hierarchies. Forthcoming (2026).

[8] Sarkar, R. Low distortion Delaunay embedding of trees in hyperbolic plane. In *Graph Drawing 2011*, Lecture Notes in Computer Science Vol. 7034, 355–366 (Springer, 2012).

[9] Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523 (1971).

[10] Nickel, M. & Kiela, D. Poincaré embeddings for learning hierarchical representations. *Adv. Neural Inf. Process. Syst.* **30**, 6338–6347 (2017).

[11] Nguyen, E. P. et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *Adv. Neural Inf. Process. Syst.* **37**, 43214–43231 (2024).

[12] Gromov, M. Hyperbolic groups. In *Essays in Group Theory* (ed. Gersten, S. M.) 75–263 (Springer, 1987).

[13] Gould, S. J. *Wonderful Life: The Burgess Shale and the Nature of History* (W. W. Norton, 1989).

[14] Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

[15] Kochurov, M., Karimov, R. & Kozlukov, S. Geoopt: Riemannian optimization in PyTorch. arXiv:2005.02819 (2020).

[16] Sala, F. et al. Representation tradeoffs for hyperbolic embeddings. In *Proc. 35th Int. Conf. Machine Learning*, 4460–4469 (2018).

[17] Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).

Author Contributions

R.F. and A.F. conceived the project, developed the theory, designed experiments, performed analyses, implemented computational proofs, and wrote the manuscript.

Competing Interests

The authors declare no competing interests.

Acknowledgements

The authors acknowledge the intellectual influence of the late Jeffrey Mann Bowen, a mathematical physicist whose informal guidance and conversations on geometry helped shape some of the perspective underlying this work. We thank Kristóf Huszár for discussions on hyperbolic geometry that clarified the mathematical lineage of Manning's theorem. We also thank Zakaria Louadi, Gihanna Galindez, and Tyler Bernard for detailed feedback on early drafts spanning computational biology, bioinformatics, and mathematical exposition. Large language models were used as collaborative tools throughout this work as programming assistants, as sounding boards for cross-disciplinary reasoning, and as aids across long-running threads of experimental work informing theoretical development. All scientific claims, mathematical derivations, and interpretive judgments are the authors' own.

Extended Data

Table 5: **Extended Data Table 1: Representative organism coordinates across five independent models.** Radial distance r and angular position θ in the Poincaré disk. Values are mean ± std across five seeds. Full coordinate tables available in the repository.

Organism	**Domain**	r	θ (degrees)
Homo sapiens	Eukarya	0.908 ± 0.029	168.4 ± 1.3
Saccharomyces cerevisiae	Eukarya	0.812 ± 0.016	28.5 ± 1.2
Arabidopsis thaliana	Eukarya	0.845 ± 0.022	152.1 ± 1.8
Escherichia coli	Bacteria	0.737 ± 0.004	42.3 ± 0.5
Bacillus subtilis	Bacteria	0.751 ± 0.008	38.7 ± 0.9
Methanocaldococcus jannaschii	Archaea	0.692 ± 0.011	267.2 ± 1.5
Halobacterium salinarum	Archaea	0.723 ± 0.009	281.6 ± 1.1

Table 6: **Extended Data Table 2: Top 20 phyla in the training dataset.**

Phylum	**Genomes**
Cyanobacteriota	648
Bacillota	413
Pseudomonadota	219
Actinomycetota	65
Methanobacteriota	37
Bacteroidota	34
Apicomplexa	32
Thermoproteota	12
Mycoplasmatota	12
Basidiomycota	11

Supplementary Information

Supplementary Information is available as a separate document accompanying this paper. Contents include:

1. **Mathematical Foundations of the State Equation.** Complete proofs of the geometric state equation, uniqueness, monotonicity properties, and high-precision numerical evaluation to 120 digits. All proofs are machine-checked in Lean 4 and cross-verified in Wolfram Language and Python/SymPy.
2. **Sensitivity and Error Propagation.** Analytic sensitivity coefficients $\partial\kappa/\partial h$ and $\partial\kappa/\partial n$, propagated uncertainty analysis, and interval bounds.
3. **Neural Network Convergence Analysis.** Complete five-seed results with Procrustes correlations, coordinate stability of representative organisms, and scale stability across dataset sizes.
4. **Ablation: Curvature Measures Phylogenetic Structure.** Fixed-curvature sweep and learnable-curvature-without-phylogenetic-signal experiments demonstrating that $\kappa = 1.247$ is a property of the data, not the architecture.
5. **Falsification Tests: Detailed Results.** Euclidean null ($\hat{\kappa} = 0.0000$), synthetic hyperbolic recovery (mean error 1.08%), and destroyed-structure controls (CV increase from 0.24% to 68%).
6. **Viral System Validation: Complete Results.** Full 15-virus table with per-seed uncertainties, discriminating tests (depth vs. rate), null controls (label shuffling, single label, random hash), and substrate independence analysis.
7. **Universal Dimensionality: Extended Analysis.** Back-solved dimensionality across all systems, latent dimension ablation confirming $n = 2$ is a data property.
8. **Lean 4 Formal Verification.** Theorem inventory (5 machine-checked theorems), build instructions, and CI pipeline description.
9. **Wolfram Language Cross-Verification.** Five independent symbolic computation notebooks with descriptions.
10. **Reproducibility.** Docker single-command verification, canonical constants and continuous integration workflows.
11. **Cross-Alphabet Validation: Protein Phylogenies.** Complete 15-family table with per-family κ, n_{back}, h_{LG}, h_{aln}, and stress. Dimensionality analysis, curvature–entropy relationship, RecA outlier analysis, and alphabet comparison summary.
12. **Architecture Independence: Minimal Encoder Validation.** A 40,482-parameter minimal encoder with analytically fixed curvature recovers the same coordinate system (Procrustes residual 0.020 across 268 organisms, 5 seeds), confirming data determination.