CURVATURE PREVENTS LOSSLESS LAYER FUSION IN FRÉCHET-MEAN NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Layer fusion is exact for networks of Euclidean weighted averages: a hierarchy of averages reduces to one average with path-product weights. We ask whether the same compression is valid for intrinsic networks that aggregate manifold-valued representations with weighted Fréchet means. For a canonical depth-two, three-input block, we show that the Euclidean path-product weights are the only possible fixed fused weights and derive the leading fusion error. For inputs of radius ε around p, the leading possible error is cubic in ε and linear in the Riemann curvature tensor; it is nonzero on suitable probes whenever $R_p \neq 0$. Consequently, this block admits an exact input-independent one-layer replacement on a neighborhood if and only if that neighborhood is flat. We extend the calculation to arbitrary finite Fréchet-mean trees via a bottom-up cubic recursion, yielding an architecture-level certificate for when depth is locally observable. Controlled experiments on constant-curvature spaces validate the predicted scaling and coefficient over $366{,}996$ source-block evaluations and $1{,}792$ random trees. A 90-configuration teacher–student study further separates three notions often conflated in compression: forced path-product fusion, a globally learned fixed surrogate, and input-conditioned fusion. Learned fixed weights reduce the reported average error but remain inexact on the curved runs; a per-input oracle attains near-zero error on the tested space-form inputs, without implying pointwise representability on general manifolds. On BCI Competition IV-2a, three frozen replacements retain the nested SPD teacher's subject-mean cross-session balanced accuracy to within 0.18 percentage points despite nonzero representation gaps; the task-specific FBCSP baseline remains stronger. These results identify curvature as a concrete barrier to lossless static compression while separating geometric fidelity from downstream accuracy.

1 INTRODUCTION

Manifold-valued representations arise whenever a model predicts or propagates objects with non-Euclidean constraints: covariance matrices, rotations, subspaces, shapes, and hyperbolic embeddings are standard examples (Huang & Van Gool, 2017; Brooks et al., 2019; Nickel & Kiela, 2017; Ganea et al., 2018; Chami et al., 2019). An intrinsic way to aggregate such representations is the weighted Fréchet mean, the minimizer of a weighted sum of squared geodesic distances. Fréchet means therefore appear as differentiable pooling, normalization, and message aggregation operators in geometric networks (Chakraborty et al., 2022; Lou et al., 2020).

This paper studies a model-compression question that is elementary in Euclidean space but unresolved for intrinsic networks:

> *Can several consecutive Fréchet-mean layers be replaced exactly by one Fréchet-mean layer with input-independent weights?*

The replacement is attractive because a Fréchet mean generally requires an iterative geometric solver and implicit differentiation. Fusing a hierarchy of such nodes into one node reduces solver calls, graph depth, and stored intermediates. In Euclidean space this transformation is exact: distributing the node weights through the computation tree yields one convex combination. The same argument cannot simply be reused on a curved manifold because intrinsic means are not affine or associative. Generic non-associativity alone, however, is not enough to answer the compression question: the

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

fused layer may choose weights different from the internal node weights, and those weights may be learned from a data distribution.

We give a local, architecture-specific characterization and connect it to observable learning behavior. Our canonical block is

$$F_{\alpha,\beta,\gamma}(A, B, C) := (A\#_\alpha B)\#_\gamma(B\#_\beta C), \tag{1}$$

where $x\#_t y$ is binary geodesic interpolation, equivalently the binary weighted Fréchet mean. This is the smallest shared-input block that captures the fusion question posed for ManifoldNet (Chakraborty et al., 2022). We compare it with a single three-input mean $\mathcal{B}_w(A, B, C)$ whose weight vector w is fixed before the input is observed.

Our contributions are:

- **An irreducible local fusion error.** We identify the only possible fixed fused weights exactly and derive the complete leading error for arbitrary tangent inputs. Its first possible term is cubic in the feature radius and linear in curvature; nonzero-curvature points admit probes on which that term is nonzero. Thus the block in Equation (1) is exactly fusible on a neighborhood if and only if that neighborhood is flat.
- **A certificate for larger aggregation networks.** For any finite tree of weighted Fréchet means, a bottom-up recursion computes its effective linear weights and its cubic curvature response. Comparing this response with the fused layer gives an inexpensive, architecture-level non-fusibility certificate and a small-radius error predictor.
- **A separation of compression objectives.** A globally learned fixed surrogate, a surrogate fitted to one triangle, and an input-conditioned surrogate solve different optimization problems. A 90-configuration teacher–student experiment shows that globally learning fixed weights improves average risk but retains a nonzero reported residual, while a per-input oracle attains near-zero error on the tested space-form examples. Pointwise exact representability on a general manifold requires an additional feasibility condition, not merely a change of quantifiers.
- **Mechanism-level validation.** Controlled experiments across seven constant curvatures, 34 source architectures, and $1{,}792$ random aggregation trees verify the predicted radius, curvature, parameter, and depth dependence. High-precision positive and flat controls distinguish geometric signal from solver error.
- **A downstream SPD benchmark.** On all nine official cross-session splits of BCI Competition IV-2a, frozen path-product, learned-fixed, and input-adaptive replacements preserve the nested teacher's subject-mean balanced accuracy to within 0.18 percentage points. Their nonzero AIRM gaps coexist with nearly unchanged task scores, while a task-specific FBCSP baseline remains stronger; this separates compression fidelity from task-model selection.

The main message is that depth in an intrinsic averaging network is not merely a redundant parameterization. Curvature makes the bracketing of aggregation operations locally observable. Static layer fusion therefore incurs a predictable representation error unless the relevant feature region is flat or the architecture belongs to a special exactly collapsing family.

Related work. Fréchet means originate in intrinsic statistics (Fréchet, 1948; Karcher, 1977; Pennec, 2006; Afsari, 2011). Recent work makes them differentiable and usable inside learned systems (Lou et al., 2020), while manifold architectures employ related intrinsic maps for SPD, hyperbolic, and general manifold-valued data (Huang & Van Gool, 2017; Brooks et al., 2019; Chakraborty et al., 2022; Ganea et al., 2018; Chami et al., 2019). Our focus differs from approximation of the mean solver itself: both the deep and fused networks use exact squared-distance Fréchet means, and we ask whether their represented functions coincide. Prior metric-space results show that barycenters need not be associative (Ginestet et al., 2012), but do not identify the uniquely admissible fused weights or quantify the local architecture error. The curvature expansion below supplies both, and turns a geometric obstruction into a testable statement about model compression.

Table 1: Three distinct compression objectives. Only the first is lossless static layer fusion.

Objective	Mathematical form	Dependence of surrogate weights
Exact fixed fusion	$\exists w_{\text{fix}}\ \forall x:\ F(x) = \mathcal{B}_{w_{\text{fix}}}(x)$	Architecture parameters only; fixed at test time
Learned fixed surrogate	$\min_{w_{\text{fix}}}\ \mathbb{E}\, d^2(F, \mathcal{B}_{w_{\text{fix}}})$	Training data; fixed at test time
Input-conditioned exact target	$\forall x\ \exists w(x):\ F(x) = \mathcal{B}_{w(x)}(x)$	Current input

2 Fréchet aggregation and the fusion objective

2.1 Intrinsic mean layers

Let $(\mathcal{M}, g)$ be a smooth connected Riemannian manifold of dimension $d \geq 2$, and let $w \in \Delta_{m-1} := \{w_i \geq 0 : \sum_i w_i = 1\}$. A weighted Fréchet-mean layer maps m manifold-valued inputs to

$$\mathcal{B}_w(x_1, \ldots, x_m) := \arg\min_{z \in U} \frac{1}{2} \sum_{i=1}^{m} w_i\, \mathrm{d}(z, x_i)^2. \tag{2}$$

The restriction to U makes the local uniqueness hypothesis explicit and avoids cut-locus and global-minimizer ambiguity. Uniqueness alone does not imply differentiability. The smooth dependence used here is established near the diagonal: the coordinate Hessian of the mean objective is the identity when all inputs equal p, so the parameter-dependent implicit-function theorem applies, uniformly over the closed weight simplex for each fixed arity (theorem A.4).

Definition 2.1 (Local mean domain)**.** A relatively compact, strongly convex normal ball $U \subset \mathcal{M}$ is a *local mean domain* if every weighted squared-distance objective in Equation (2), of any finite arity and with inputs in U, has one minimizer in U, and that minimizer is the unique zero in U of the Karcher field $z \mapsto \sum_i w_i \operatorname{Log}_z x_i$.

Sufficiently small balls satisfy this condition (Karcher, 1977; Kendall, 1990; Groisser, 2004; Afsari, 2011). For two inputs, $x \#_t y := \operatorname{Exp}_x(t \operatorname{Log}_x y) = \mathcal{B}_{(1-t,t)}(x, y)$. A network of mean layers is therefore a computation tree whose internal nodes aggregate their children; shared external inputs can occur at several leaves. Because each node maps U back to U, any fixed finite tree is well defined on U.

2.2 What counts as a fused layer?

We distinguish three objectives that coincide only in special cases. Let $x = (A, B, C)$ and let $F(x)$ denote the depth-two block in Equation (1).

The first row asks whether a compiler or deployment system can replace the block by one static layer without changing its function. The second is a distillation problem: it may lower average error by moving away from the Euclidean fused weights, but nonzero population risk is still compatible with an exact impossibility result. The third is an exact target for an attention-like model with a different hypothesis class and computational graph, not a universal existence theorem. For a specified target $y = F(x)$ in U, it is attainable precisely when zero belongs to the convex hull of the vectors $\operatorname{Log}_y x_i$ in $T_y\mathcal{M}$ (proposition F.3). This condition can fail even for arbitrarily concentrated SPD inputs (proposition F.5). Fitting weights after seeing one test example therefore proves neither static fusibility nor pointwise feasibility outside the evaluated configurations.

For interior parameters $(\alpha, \beta, \gamma) \in (0, 1)^3$, the Euclidean path-product calculation gives

$$\omega^0 = \big((1-\gamma)(1-\alpha),\ (1-\gamma)\alpha + \gamma(1-\beta),\ \gamma\beta\big). \tag{3}$$

Throughout, write $g_{\alpha,\beta,\gamma} := \gamma(1-\gamma)(1-\alpha)\beta$. These weights are not merely a convenient baseline. Move one external input along a radial geodesic through p and keep the other two equal to p. Every node then lies on the same geodesic, where intrinsic means reduce exactly to scalar weighted averages. Repeating this probe for A, B, and C forces the three coordinates in Equation (3). Thus any fixed

weight $w \neq \omega^0$ disagrees with the block already at first order. Curvature enters only after this identifiable linear response has been matched.

2.3 Local probes and evaluation metric

Fix $p \in U$ and tangent directions $a, b, c \in T_p\mathcal{M}$. We evaluate the two networks on concentrated inputs

$$A_\varepsilon = \mathrm{Exp}_p(\varepsilon a), \qquad B_\varepsilon = \mathrm{Exp}_p(\varepsilon b), \qquad C_\varepsilon = \mathrm{Exp}_p(\varepsilon c), \tag{4}$$

where ε controls feature spread. Normal-coordinate output differences give vector-valued errors and geodesic distance gives the deployment-level error. The two agree to leading order for small ε. We use the curvature convention $R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z$. All asymptotic remainders below are uniform for tangent directions and node parameters in fixed compact sets. Complete well-posedness statements and remainder proofs are in Appendix A.

3 Curvature creates a depth-dependent representation gap

The central calculation compares the depth-two block with its only admissible fixed surrogate. Write $x = a - b$ and $y = c - b$ for the two relative tangent directions. The common displacement b affects both outputs but cancels from their leading difference.

Theorem 3.1 (Forced fused weights and cubic gap)**.** *Let U be a local mean domain, $p \in U$, and $(\alpha, \beta, \gamma) \in (0,1)^3$. If $F_{\alpha,\beta,\gamma} = \mathcal{B}_w$ on a neighborhood of (p,p,p), then $w = \omega^0$ from Equation (3). For the probes in Equation (4),*

$$\begin{aligned} &\mathrm{Log}_p F_{\alpha,\beta,\gamma}(A_\varepsilon, B_\varepsilon, C_\varepsilon) - \mathrm{Log}_p \mathcal{B}_{\omega^0}(A_\varepsilon, B_\varepsilon, C_\varepsilon) \\ &\quad = \frac{\varepsilon^3}{3}\,\gamma(1-\gamma)(1-\alpha)\beta\, R_p(x,y)\big(\alpha x - (1-\beta)y\big) + \mathcal{O}(\varepsilon^4). \end{aligned} \tag{5}$$

The proof expands a multivariable Fréchet mean in normal coordinates and then propagates the expansion through the computation graph. It is important that Equation (5) is a vector identity for arbitrary (a, b, c), rather than a scalar calculation for one hand-picked triangle. The full derivation, including a derivative-controlled remainder, appears in Appendices A and B.

Interpretation. The leading possible error is the product of feature spread ε^3, architecture factor $g_{\alpha,\beta,\gamma}/3$, and directional curvature $R_p(x,y)(\alpha x - (1-\beta)y)$. A nonzero directional factor makes the cubic term the first nonzero discrepancy; curvature at p alone does not make every probe nondegenerate. The architecture factor vanishes as $\alpha \to 1$, $\beta \to 0$, or $\gamma \to 0, 1$, where the limiting computation loses a relevant branch or reduces to a single geodesic. Other boundary limits need not reduce the gap: $\alpha \to 0$ and $\beta \to 1$ can retain a genuine three-leaf tree, as quantified in corollary B.10 and theorem D.4.

For orthonormal $X, Y \in T_p\mathcal{M}$ with sectional curvature $K_p(X,Y)$,

$$\|R_p(X,Y)(\alpha X - (1-\beta)Y)\| \geq |K_p(X,Y)|\sqrt{\alpha^2 + (1-\beta)^2}. \tag{6}$$

Thus every nonzero sectional-curvature plane yields a concrete input witness. This gives an exact classification, not only an asymptotic lower bound.

Theorem 3.2 (Exact static-fusion classification)**.** *For a fixed interior parameter triple, there exists one $w \in \Delta_2$ such that $F_{\alpha,\beta,\gamma}(A,B,C) = \mathcal{B}_w(A,B,C)$ for all $(A,B,C) \in U^3$ if and only if $R \equiv 0$ on U. In the flat case, the unique fused weights are ω^0.*

Necessity follows from Equations (5) and (6) at each point. For sufficiency, a flat, strongly convex neighborhood admits Euclidean coordinates in which every Fréchet mean is an affine average, so ordinary path-product fusion is exact. Completeness is unnecessary for this local result. When global means are desired, sufficiently concentrated inputs have the same unique local and global minimizers; lemma B.7 transfers the witness to the global ManifoldNet setting.

Equal-weight block. For the commonly used symmetric choice $\alpha = \beta = \gamma = 1/2$, the forced surrogate has weights $(1/4, 1/2, 1/4)$. On the probe $(a, b, c) = (X, 0, Y)$,

$$\operatorname{Log}_p F_{\mathrm{eq}} - \operatorname{Log}_p \mathcal{B}_{(1/4,1/2,1/4)} = -\frac{\varepsilon^3}{96} R_p(X, Y)(Y - X) + \mathcal{O}(\varepsilon^4). \tag{7}$$

This display is only the equal-weight specialization. For arbitrary interior (α, β, γ), take orthonormal $X, Y \in T_p\mathcal{M}$ and use the probe $(\operatorname{Exp}_p(\varepsilon X), p, \operatorname{Exp}_p(\varepsilon Y))$. Then Equation (5) applies with $(x, y) = (X, Y)$ and the path-product weights ω^0 from Equation (3); no equal-weight assumption is used in the classification or the following space-form formula. For a space of constant curvature κ, the corresponding distance obeys

$$\lim_{\varepsilon \downarrow 0} \varepsilon^{-3} d(F_{\alpha,\beta,\gamma}, \mathcal{B}_{\omega^0}) = \frac{|\kappa|}{3} \gamma(1 - \gamma)(1 - \alpha)\beta \sqrt{\alpha^2 + (1 - \beta)^2}. \tag{8}$$

The vector formula (5) predicts the direction and sign reversal with curvature. The nonnegative distance formula (8) predicts magnitude, linear dependence on $|\kappa|$, and radius scaling, but cannot recover the sign of κ.

Consequences for learning a surrogate. For a fixed $w \neq \omega^0$, a suitable axial probe exposes a $\Theta(\varepsilon)$ error. For $w = \omega^0$ and $R_p \neq 0$, the local *worst-case* distance over the unit tangent input set is $\Theta(\varepsilon^3)$ (proposition C.12). A particular probe has this order only when its displayed cubic coefficient is nonzero. Likewise, expected squared loss is $\Theta(\varepsilon^6)$ only if the expected squared cubic coefficient is positive; a distribution supported on one geodesic can have zero loss. If $R_p = 0$ but curvature is nonzero nearby, the expansion at p has no cubic term even though exact fusion on a full neighborhood remains impossible.

Optimizing fixed weights at a finite radius can improve distributional loss, but if $R|_U \not\equiv 0$, it cannot make the two maps identical on all of U^3. In particular, $R_p \neq 0$ excludes equality on every neighborhood of (p, p, p). This does not exclude equality on a flat open subregion of a nonflat smooth manifold; example F.1 gives an explicit example. Allowing $w = w(A, B, C)$ is a different problem and succeeds only when the pointwise Karcher convex-hull condition holds. These distinctions determine what the teacher–student comparison in Section 5 can test.

Why adaptive weights can still fail. Let $W = \operatorname{span}\{X, Y\}$ for independent X, Y. An exact adaptive match on $(\operatorname{Exp}_p(\varepsilon X), p, \operatorname{Exp}_p(\varepsilon Y))$ forces $w - \omega^0 = \mathcal{O}(\varepsilon^2)$ by linear identifiability. Weight adjustment changes the linear output only within W and changes its cubic term only at fifth order. Thus a nonzero $W^\perp$ component of $D = -R_p(X, Y)(Y - X)/96$ cannot be removed. In affine-invariant SPD geometry, $X = \operatorname{diag}(1, 0)$ and $Y = (E_{12} + E_{21})/\sqrt{2}$ at I give $\langle D, \operatorname{diag}(0, 1)\rangle_F = -1/384$, where E_{ij} are matrix units. No convex weights then represent the nested output at small positive radii. The uniform proof in proposition F.5 permits arbitrary weight selections, without a smoothness assumption. This limits a universal interpretation of adaptive fusion, not the recorded oracle results on tested space forms.

4 From one block to general aggregation networks

The source block gives a complete classification, but a practical model may contain heterogeneous arities, repeated inputs, and many levels. Unfold such a computation graph into a rooted tree $\mathcal{T}$. Each internal node N has child weights θ^N, and each leaf carries one of the external input labels. The effective linear weight of input i is the sum, over leaves labeled i, of the products of node weights along the corresponding root-to-leaf paths. Radial probes show that these effective weights $\omega^{\mathcal{T}}$ are again the unique possible fixed fused weights.

4.1 A bottom-up cubic recursion

For tangent inputs $x_i(\rho) = \operatorname{Exp}_p(\rho v_i)$, write the normal-coordinate output of a subtree rooted at N as $\operatorname{Log}_p F_N(\rho) = \rho\mu_N + \rho^3 Q_N + \mathcal{O}(\rho^4)$. At a leaf labeled i, set $(\mu_N, Q_N) = (v_i, 0)$. An internal node with children N_j obeys

$$\mu_N = \sum_j \theta_j \mu_{N_j}, \qquad Q_N = \sum_j \theta_j Q_{N_j} + \frac{1}{3} \sum_j \theta_j R_p(\mu_N, \mu_{N_j}) \mu_{N_j}. \tag{9}$$

The fused one-layer mean has the same $\mu_{\mathcal{T}}$ and cubic coefficient $Q_\star^{\mathcal{T}} = \sum_i \omega_i^{\mathcal{T}} R_p(\mu_{\mathcal{T}}, v_i) v_i / 3$. Their difference $\mathfrak{D}_{\mathcal{T}}^p := Q_{\mathcal{T}} - Q_\star^{\mathcal{T}}$ is the *cubic architecture defect*. The complete result in theorem C.2 and corollary C.4 gives $\mathrm{Log}_p F_{\mathcal{T}}(x(\rho)) - \mathrm{Log}_p \mathcal{B}_{\omega\mathcal{T}}(x(\rho)) = \rho^3 \mathfrak{D}_{\mathcal{T}}^p(v) + \mathcal{O}(\rho^4)$.

Equation (9) is evaluated in one post-order pass and mirrors ordinary forward propagation. It supplies three ML-facing tools:

- a *fusion certificate*: one nonzero value of $\mathfrak{D}_{\mathcal{T}}^p(v)$ proves that no fixed one-layer Fréchet mean represents the tree near p;
- a *local error predictor*: $\rho^3 \left\| \mathfrak{D}_{\mathcal{T}}^p(v) \right\|$ estimates the output distortion before actually replacing the network; and
- a *curvature-aware regularizer*: under a compactly supported tangent distribution, the expected squared fusion loss is $\rho^6 \mathbb{E} \left\| \mathfrak{D}_{\mathcal{T}}^p(v) \right\|^2 + \mathcal{O}(\rho^7)$.

The recursion also yields a depth-dependent upper bound. If V bounds all leaf tangent norms and $L_{\mathcal{T}} = \sum_\ell \pi_\ell d_\ell$ is the expected number of internal nodes along a root-to-leaf path under the leaf-mass distribution, then $\left\| \mathfrak{D}_{\mathcal{T}}^p(v) \right\| \le (L_{\mathcal{T}} + 1) \left\| R_p \right\|_{\mathrm{op}} V^3 / 3$. Here π_ℓ is the product of the node weights along the path from the root to leaf ℓ, and d_ℓ is the number of internal nodes on that path. Equivalently, $L_{\mathrm{leaf}} = 0$ and $L_N = 1 + \sum_j \theta_j L_{N_j}$. This is a weighted average depth, not a maximum of pathwise node masses, and the upper bound does not assert monotone growth of actual defects with depth. The finite-radius version requires a neighborhood bound on the next geometric jet as well as pointwise curvature; theorem C.9 states the controlled remainder. We therefore do not promote the cubic coefficient to a uniform finite-radius guarantee without its assumptions.

4.2 What the certificate does not say

Curvature does not make every syntactically deep tree non-fusible. For example, the following identity holds on every strongly convex neighborhood, regardless of curvature.

Proposition 4.1 (A depth-two block that always fuses). *For $s, t \in [0, 1]$, $A \#_s (A \#_t B) = A \#_{st} B$.*

Both operations follow the same minimizing geodesic, so the identity is just a reparameterization. Conversely, a nonzero cubic defect is sufficient but not necessary for failure of exact fusion. Proposition C.5 constructs a symmetric three-input tree whose entire cubic defect cancels but whose first nonzero error on the sphere is fifth order. A zero cubic score should therefore be read as "undecided," not "safe to fuse."

The correct architecture-independent statement is a universal one.

Theorem 4.2 (Universal static-fusion property). *On a sufficiently small local mean domain U, the following are equivalent: (i) every finite fixed-weight Fréchet-mean network fuses to one fixed Fréchet mean of its external inputs; (ii) the equal-weight block in Equation (7) fuses; and (iii) $R|_U \equiv 0$.*

Thus flatness is exactly the condition under which Euclidean layer-fusion rules are valid for the whole model class. On a curved manifold, fusibility must be checked at the level of the particular architecture and input regime.

5 Experiments

The experiments test the mechanism, distinguish compression objectives, and then evaluate frozen replacements in a downstream SPD task. We ask: (Q1) does the source-block error follow the predicted radius, curvature, and parameter law; (Q2) can training one fixed surrogate remove the gap, and how does this differ from input-conditioned fitting; and (Q3) does the cubic recursion scale to diverse trees, including cases where the cubic term cancels; and (Q4) how does representation fidelity relate to cross-session task accuracy? Full protocols, tabulated high-precision values, solver checks, and artifact boundaries are in Appendix E.

Geometries and numerical controls. We evaluate Euclidean and spherical/hyperbolic constant-curvature models with closed-form exponential and logarithm maps. Multi-input Fréchet means are

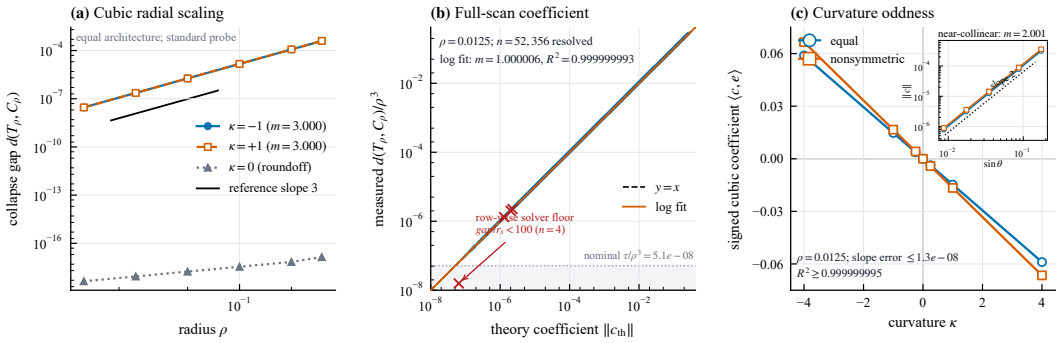


Figure 1: Source-block mechanism check. Left: cubic radius scaling for the equal block. Center: measured versus predicted coefficient over all eligible finest-radius rows; four solver-limited points are marked separately. Right: signed linear dependence on curvature and the predicted near-collinear degeneration.

computed by a Karcher solver; curvature is used only in the analytic prediction, not supplied to the solver. Every row records the stationarity residual. Nonzero-curvature configurations are paired across signs, $\kappa = 0$ is an exact-fusion control, and selected signals are recomputed at 80 and 120 decimal digits to separate geometric orders from float64 solver floors.

5.1 Q1: Does the curvature law predict the measured gap?

E1 contains 366,996 full-scan evaluations and 186 targeted evaluations: seven curvatures $\kappa \in \{-4, -1, -1/4, 0, 1/4, 1, 4\}$, six feature radii, the equal and nonsymmetric source blocks, 32 additional interior parameter triples, and one standard plus 256 random probes per architecture. On the standard witnesses, fitted distance orders lie in $[2.997181, 3.002752]$, matching the predicted cubic law. At the finest radius, the median and 95th-percentile relative errors of the predicted vector coefficient are 2.398×10^{-4} and 4.269×10^{-3}. The signed coefficient reverses with curvature; targeted coefficient-versus-curvature fits have relative slope error at most 2.424×10^{-6} and $R^2 \geq 0.999998758$. Euclidean gaps are at roundoff, and near-collinear probes exhibit the predicted directional degeneration rather than a spurious curvature-only lower bound.

Of the 52,360 eligible finest-radius curved rows, four have a gap-to-residual ratio below the predeclared safety factor of 100. They remain in the artifact and are marked in Figure 1. A tolerance sweep resolves the sole associated coefficient outlier, recovering order 3.001128 at tighter tolerance. The resolved coefficient regression has slope 1.0000056 and $R^2 = 0.9999999928$.

5.2 Q2: Fixed distillation versus input-conditioned matching

E2a first checks the error orders without learning. Across the four curved architecture–curvature strata, the forced path-product surrogate has distance order in $[2.999373, 3.000624]$ and random-input squared-loss order in $[5.998133, 6.001872]$. Deliberately wrong fixed weights have order one, as predicted by first-order identifiability.

E2b then uses 90 configurations: two source architectures, three curvatures, three radii, and five seeds, with 5,000/1,000/5,000 train/validation/test triples per run. On all 60 curved runs, optimizing one global fixed vector reduces held-out mean squared error to 0.676–0.802 of the path-product baseline, while the learned weights remain within 2.154×10^{-4} in ℓ_∞ of ω^0. The nonzero reported residual is consistent with the theoretical point: distributional optimization can improve the constant without changing the represented-function impossibility.

A weight vector fitted to one triangle attains fitted-example error at most 1.333×10^{-15}. Across the 60 curved runs, the run-level median of its test-error ratio is 1.865 relative to the path-product baseline. In Figure 2, each plotted point first averages the five seeds within one architecture–curvature–radius stratum, and the median of those 12 ratios is 2.917. A per-input constrained oracle returns simplex-feasible weights on every tested triple; the maximum across runs of its held-out mean squared error

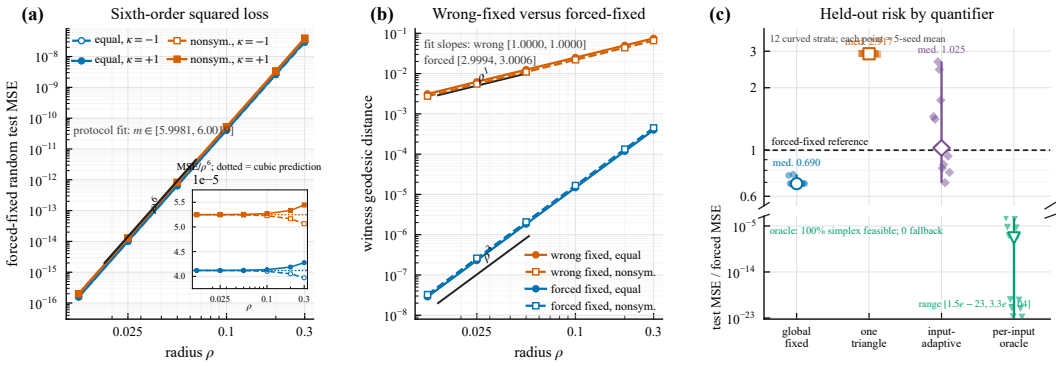


Figure 2: Compression objectives behave differently. (a) Forced fixed fusion has the predicted ρ^6 squared-loss scaling. (b) Wrong fixed weights expose a first-order error, versus the cubic path-product error. (c) Global fixed training improves average risk but remains inexact; one-example fitting does not generalize, while a per-input oracle occupies a different, near-zero scale. The upper axis shows the three finite-risk ratios and the broken lower axis shows the oracle ratios. Points are five-seed stratum means; Appendix E reports the weight and solver diagnostics.

is 3.331×10^{-16}, not a reported maximum over individual samples. The oracle does not contradict theorem 3.2: it selects a different weight vector after each input is known. Nor does this space-form experiment imply pointwise feasibility on SPD manifolds, which proposition F.5 can rule out even locally. The adaptive neural control is included only to visualize this quantifier change, not as a capacity-matched model ranking.

5.3 Q3: Does the recursion scale beyond the source block?

E3 samples 256 ordered full-binary trees at each depth 2 through 8, for 1,792 trees and 59,136 tree–probe pairs. Node weights lie in $[0.1, 0.9]$; random probes and tree parameters are paired across $\kappa \in \{-1, 1\}$, with a separate Euclidean control. Across all 38 curvature–label–depth strata, regressing measured vector coefficients on the recursion prediction gives slopes in $[0.999950, 1.000050]$ and $R^2 \geq 0.9999999988$. The largest observed ratio to the theorem's depth-dependent coefficient bound is 0.115784, with no violation. Of 118,272 finest-radius curved rows, 290 near-cancellation rows fall below the v1.1 resolution gate and are excluded only from coefficient-accuracy aggregates, not deleted.

Finally, E4 tests the zero-cubic-defect construction from proposition C.5. Its high-precision fitted order is 4.999965, and its measured fifth-order coefficient has relative error 1.196×10^{-6}. The same solver yields order 3.000123 on an ordinary cubic tree and numerical zero on the Euclidean version. This control confirms both sides of the certificate: a nonzero cubic defect detects non-fusibility, while a zero cubic defect can hide a higher-order gap.

5.4 Q4: Does approximate fusion preserve downstream accuracy?

We use all nine subjects of BCI Competition IV-2a (Tangermann et al., 2012), treating subjects as the independent units and using only the official training-to-evaluation-session direction. The two stochastic replacements are averaged over six seeds within each subject; deterministic methods are run once. Three Oracle Approximating Shrinkage (OAS)-regularized symmetric positive-definite (SPD) covariance views in $\mathrm{SPD}(22)$ (mu, low beta, and high beta) feed the nested teacher $(C_1 \#_{0.35} C_2) \#_{0.5} (C_2 \#_{0.65} C_3)$ and its frozen replacements. Evaluation-session data are not used in training or model selection and are scored once after freezing. We compare with tangent-space and filter bank common spatial pattern (FBCSP) (Ang et al., 2012) task baselines on the same retained trials.

The three subject-mean accuracies stay within 0.18 percentage points of the teacher; this is not a bound on each subject. The recomputed two-sided exact subject-wise sign-flip values are $p = .125$, 1.000, and $.969$ for path-product, learned-fixed, and input-adaptive fusion, respectively.

Table 2: BCI IV-2a official cross-session evaluation over nine subjects. Balanced accuracy is subject mean $\pm$ standard deviation in percent; Cohen's κ is the agreement score and Δ_{T} is the paired frozen-head change from the nested teacher. The learned-fixed and input-adaptive rows first average six seeds within each subject. The normalized AIRM gap divides the replacement–teacher distance by input dispersion, where AIRM denotes the affine-invariant Riemannian metric. Differences are computed from unrounded scores. Dashes denote task baselines without a teacher-fidelity target.

Method	Bal. acc.	κ	Norm. gap	Δ_{T} (pp)
Nested Fréchet teacher	60.28 ± 15.50	.471	$< 10^{-12}$	0.00
Path-product fixed	60.45 ± 15.45	.473	.0129	$+0.18$
Learned fixed (6 seeds)	60.28 ± 15.53	.471	.0105	0.00
Input-adaptive (6 seeds)	60.27 ± 15.57	.470	.0104	-0.01
Tangent-space concatenation	69.44 ± 15.32	.592	–	–
Broad-band tangent-space LR	62.19 ± 16.45	.496	–	–
FBCSP	$\mathbf{70.01 \pm 14.24}$	**.600**	–	–

The finite-sample testing interpretation requires sign-flip invariance under the null, not merely independent zero-mean differences (Appendix F.3); these tests use the absolute mean difference over all 2^9 possible sign configurations, after averaging stochastic seeds within subject. They do not establish statistical equivalence. For the two stochastic models, the standard deviation across the six nine-subject seed means is only .054 and .050 percentage points. Learning reduces the normalized AIRM gap from .0129 to about .0105, but neither the teacher nor its replacements beat FBCSP. Thus this benchmark supports downstream fidelity of approximate fusion, not a claim of improved BCI accuracy. Similar aggregate accuracy does not by itself establish agreement of the trial-wise predictions; a decision-disagreement rate would be needed for that separate claim.

6 Discussion, Limitations, and Conclusion

Practical reading of the theory. For static deployment, path-product weights are the only fused weights that preserve the first-order response. The cubic defect then estimates the error created by removing depth. A nonzero defect rules out lossless fusion; a small defect supports approximate fusion only at the evaluated feature scale and directions. Optimizing one fixed surrogate can further reduce distributional risk, but its validation objective should not be described as exact layer fusion. An input-conditioned surrogate may be exact on a restricted family only when its pointwise Karcher feasibility condition holds; SPD counterexamples show that retaining a weight-prediction mechanism does not guarantee this.

Limitations. Our exact classification concerns squared-distance Fréchet means, fixed interior node weights, smooth Riemannian geometry, and inputs in a region where all relevant means and logarithm maps are unique. Extrinsic averages, robust nonsquared objectives, attention-conditioned node weights, branch choices among nonunique global means, and truncated solvers define different operators. The leading coefficient is local: a finite-radius guarantee also needs the fourth-order remainder controlled in theorem C.9. The BCI experiment adds one nine-subject SPD application, but it is not evidence of an accuracy improvement: FBCSP is stronger, and the replacement comparison has limited subject-level power. Neither nonsignificant difference tests nor nearly equal subject-mean scores establish equivalence or trial-wise agreement. Its recorded batch-one timings are hardware-specific and do not establish an end-to-end speedup. Testing accuracy–compute tradeoffs in other SPD tasks, hyperbolic models, and pose-estimation networks remains important.

Conclusion. Euclidean averaging networks can hide their depth inside one set of effective weights; curved intrinsic networks generally cannot. For the canonical ManifoldNet block, a point with nonzero curvature admits an unavoidable cubic worst-case representation gap, and exact static fusion throughout a local mean domain is equivalent to flatness of that domain. For larger networks, a one-pass recursion exposes the same mechanism and separates certifiable non-fusion from higher-order undecided cases. The experiments separately show that surrogate objective and representation fidelity must be distinguished from downstream accuracy.

AI USE STATEMENT

We used generative AI tools to assist with reviewing mathematical claims, drafting and reviewing experimental code, developing counterexamples, revising proofs, checking symbolic identities, organizing supplementary material, and editing the manuscript for clarity and LaTeX formatting. AI tools also provided feedback on the description of experimental methods and the interpretation and scope of the results. The authors reviewed the AI-assisted material and take responsibility for the final manuscript, proofs, code, and reported results.

REPRODUCIBILITY STATEMENT

Appendices A to D and F state the assumptions and give complete mathematical arguments, including counterexamples delimiting the claims. Appendix E gives the experimental protocols and summary values. The supplementary package contains the canonical frozen bundles `e1-full-v1`, `e2-full-v1`, `e3-full-v1.1`, `e4-full-v1`, and `bci-iv2a-v1.1`. These bundles include machine-readable raw or run-level results, configurations, analysis and validation code, reports, manifests, and SHA-256 checksums; the BCI bundle also contains all nine subject-level records and the six-seed subject-first aggregate. The source contains the full proofs and the minimal sphere check described in Appendix E. No separate computer-algebra or formal-proof certificate is claimed.

ETHICS STATEMENT

The downstream results concern secondary analysis of the existing BCI Competition IV-2a dataset, as described in Appendix E; this revision introduces no new participant recruitment or data collection. The geometric results do not establish clinical utility, individual-subject equivalence, or a medical-device performance guarantee. Experimental records must be handled under the source dataset's applicable access and use terms; the supplementary package contains subject-level summary performance metrics, but no raw EEG recordings or direct personal identifiers.

REFERENCES

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. ISBN 978-0-691-13298-3. URL https://sites.uclouvain.be/absil/amsbook/.

Bijan Afsari. Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011. doi: 10.1090/S0002-9939-2010-10541-5.

Kai Keng Ang, Zheng Yang Chin, Chuanchu Wang, Cuntai Guan, and Haihong Zhang. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in Neuroscience*, 6:39, 2012. doi: 10.3389/fnins.2012.00039.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, 2007. ISBN 978-0-691-12918-1.

Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 15489–15500, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6e69ebbfad976d4637bb4b39de261bf7-Abstract.html.

Rudrasis Chakraborty, Jose Bouza, Jonathan H. Manton, and Baba C. Vemuri. ManifoldNet: A deep neural network for manifold-valued data with applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):799–810, 2022. doi: 10.1109/TPAMI.2020.3003846.

Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4868–4879, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0415740eaa4d9decbc8da001d3fd805f-Abstract.html.

Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré*, 10(4):215–310, 1948. URL https://www.numdam.org/item/AIHP_1948__10_4_215_0/.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5345–5355, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/dbab2adc8f9d078009ee3fa810bea142-Abstract.html.

Alexey V. Gavrilov. The double exponential map and covariant derivation. *Siberian Mathematical Journal*, 48(1):56–61, 2007. doi: 10.1007/s11202-007-0006-4.

Cedric E. Ginestet, Andrew Simmons, and Eric D. Kolaczyk. Weighted Fréchet means as convex combinations in metric spaces: Properties and generalized median inequalities. *Statistics & Probability Letters*, 82(10):1859–1863, 2012. doi: 10.1016/j.spl.2012.06.001.

David Groisser. Newton's method, zeroes of vector fields, and the Riemannian center of mass. *Advances in Applied Mathematics*, 33(1):95–135, 2004. doi: 10.1016/j.aam.2003.08.003.

Zhiwu Huang and Luc Van Gool. A Riemannian network for SPD matrix learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2036–2042, 2017. doi: 10.1609/aaai.v31i1.10866.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977. doi: 10.1002/cpa.3160300502.

Wilfrid S. Kendall. Probability, convexity, and harmonic maps with small image I: Uniqueness and fine existence. *Proceedings of the London Mathematical Society*, s3-61(2):371–406, 1990. doi: 10.1112/plms/s3-61.2.371.

Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge Belongie, Ser-Nam Lim, and Christopher De Sa. Differentiating through the Fréchet mean. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6393–6403. PMLR, 2020. URL https://proceedings.mlr.press/v119/lou20a.html.

John Milnor. Curvatures of left invariant metrics on Lie groups. *Advances in Mathematics*, 21(3):293–329, 1976. doi: 10.1016/S0001-8708(76)80002-3.

Maher Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747, 2005. doi: 10.1137/S0895479803436937.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6338–6347, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html.

Xavier Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006. doi: 10.1007/s10851-006-6228-4.

Xavier Pennec. Curvature effects on the empirical mean in Riemannian and affine manifolds: A non-asymptotic high concentration expansion in the small-sample regime. arXiv preprint arXiv:1906.07418, 2019. URL https://arxiv.org/abs/1906.07418.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. doi: 10.1016/j.neuroimage.2014.01.060.

A Local geometry and cubic Karcher calculus

A.1 Convexity, conventions, and radial exactness

Let $(\mathcal{M}, g)$ be a smooth connected Riemannian manifold of dimension $d \geq 2$. We use the curvature convention

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z, \qquad R(X,Y,Z,W) = \langle R(X,Y)Z, W \rangle . \quad (10)$$

For an orthonormal pair, $K_p(X,Y) = R_p(X,Y,Y,X)$. The $(0,4)$ tensor is skew in the first and last pairs, is symmetric under exchange of pairs, and satisfies the first Bianchi identity.

Fix $p \in \mathcal{M}$ and a local mean domain U containing p in the sense of definition 2.1. Unless the superscript glob is present, every Fréchet mean below is the unique local mean in U. This is the standing local well-posedness hypothesis.

Lemma A.1 (Radial exactness). *Let $v \in T_p\mathcal{M}$, and let $s_1, \dots, s_m$ be small enough that $x_i = \operatorname{Exp}_p(s_i v)$ lie in U. Then*

$$\mathcal{B}_w(x_1, \dots, x_m) = \operatorname{Exp}_p\left(\Big(\sum_{i=1}^{m} w_i s_i \Big) v \right). \quad (11)$$

In particular, $\operatorname{Exp}_p(sv) \#_\theta \operatorname{Exp}_p(tv) = \operatorname{Exp}_p(((1-\theta)s + \theta t)v)$.

Proof. Let $\gamma(q) = \operatorname{Exp}_p(qv)$ and put $\bar{s} = \sum_i w_i s_i$ and $z = \gamma(\bar{s})$. Because all parameters are taken in a sufficiently short interval, the unique minimizing geodesic from z to $x_i = \gamma(s_i)$ is the corresponding subsegment of γ. If $V(q) = \dot{\gamma}(q)$, then V is parallel along γ, and therefore $\operatorname{Log}_z x_i = (s_i - \bar{s})V(\bar{s})$. Consequently, $\sum_i w_i \operatorname{Log}_z x_i = (\sum_i w_i (s_i - \bar{s}))V(\bar{s}) = 0$. Thus z is a zero of the Karcher vector field. By the standing local well-posedness hypothesis, it is the unique minimizer of E_w in U. This proves (11). The binary identity is the case $m = 2$. □

A.2 A derivative-controlled world-function expansion

Write $\exp(u) = \operatorname{Exp}_p(u)$ and identify $T_p\mathcal{M}$ with $\mathbb{R}^d$ through an orthonormal frame. Let $H(u,v) = \frac{1}{2}\,\mathrm{d}(\exp(u), \exp(v))^2$. On a sufficiently small product ball, H is smooth because the two points are joined by a unique minimizing geodesic. The fourth-order Taylor polynomial of the world function in normal coordinates is standard (Gavrilov, 2007; Pennec, 2019).

Lemma A.2 (World function in C^1 form). *There are $r_0, C > 0$ such that, whenever $\|u\| + \|v\| \leq r_0$,*

$$H(u,v) = \frac{1}{2}\|u - v\|^2 - \frac{1}{6} R_p(u,v,v,u) + r_5(u,v), \quad (12)$$

where

$$|r_5(u,v)| \leq C(\|u\| + \|v\|)^5, \qquad \|D_u r_5(u,v)\| + \|D_v r_5(u,v)\| \leq C(\|u\| + \|v\|)^4. \quad (13)$$

Consequently, with the gradient taken with respect to the fixed inner product g_p,

$$\nabla_u^{(p)} H(u,v) = u - v - \frac{1}{3} R_p(u,v)v + \mathcal{O}((\|u\| + \|v\|)^4), \quad (14)$$

with a uniform remainder.

Proof. The smooth function H has no cut-locus singularity on the chosen product ball. Its Taylor polynomial through total degree four is $\frac{1}{2}\|u - v\|^2 - \frac{1}{6} R_p(u,v,v,u)$; this follows either from the normal-coordinate metric expansion or from the fourth-order expansion of the double exponential map. Taylor's theorem applied to H and to its first derivatives yields (13). This derivative estimate is essential: a scalar statement $r_5 = \mathcal{O}(r^5)$ alone would not justify differentiating the remainder.

For $Q_v(u) = R_p(u,v,v,u)$ and $h \in T_p\mathcal{M}$, multilinearity and pair symmetry give $DQ_v(u)[h] = R_p(h,v,v,u) + R_p(u,v,v,h) = 2R_p(u,v,v,h) = 2\langle R_p(u,v)v, h \rangle$. Differentiating (12) and invoking (13) gives $D_u H(u,v)[h] = \langle u - v - \frac{1}{3} R_p(u,v)v, h \rangle + \mathcal{O}((\|u\| + \|v\|)^4)\,\|h\|$, which is equivalent to (14). □

Lemma A.3 (Normal-coordinate comparison of distance)**.** *For every $A > 0$ there are constants $\rho_0, C > 0$ such that, whenever $0 < \rho < \rho_0$ and $\|u\|, \|v\| \le A\rho$,*

$$(1 - C\rho^2)\,\|u - v\| \le \mathrm{d}(\mathrm{Exp}_p u, \mathrm{Exp}_p v) \le (1 + C\rho^2)\,\|u - v\|\,. \tag{15}$$

The constants are uniform when u/ρ and v/ρ range over a fixed compact set.

Proof. Shrink the normal-coordinate ball, if necessary, so that it is strongly convex. The normal-coordinate metric satisfies

$$(1 - C_0\,\|z\|^2)\,\|\xi\|^2 \le g_z(\xi, \xi) \le (1 + C_0\,\|z\|^2)\,\|\xi\|^2 \tag{16}$$

for z in that ball. The coordinate segment from u to v remains in the ball of radius $A\rho$, so its Riemannian length gives the upper bound in (15).

For the lower bound, let σ be the minimizing geodesic from $\mathrm{Exp}_p u$ to $\mathrm{Exp}_p v$ and let $z(t) = \mathrm{Log}_p \sigma(t)$. The preceding upper bound gives $L_g(\sigma) = \mathcal{O}(\rho)$. On the fixed coordinate ball, the Euclidean and Riemannian norms are uniformly equivalent; hence the Euclidean length of z is also $\mathcal{O}(\rho)$. Since $\|u\| \le A\rho$, every point of z lies in a ball of radius $A'\rho$, where A' is independent of u, v, and ρ. Applying the lower estimate in (16) along z yields $L_g(\sigma) \ge (1 - C_1\rho^2)L_{\mathrm{E}}(z) \ge (1 - C_1\rho^2)\,\|u - v\|$. After enlarging the constant, this is the required lower bound. □

A.3 The multivariable cubic expansion

The form needed for network recursion allows the tangent inputs themselves to vary with the small parameter.

Theorem A.4 (Cubic expansion of a weighted Fréchet mean)**.** *Let w range over a compact subset of Δ_{m-1}, let $u_1, \dots, u_m \in T_p\mathcal{M}$, and put $r = \max_i \|u_i\|$ and $\bar{u} = \sum_i w_i u_i$. As $r \downarrow 0$,*

$$\mathrm{Log}_p \mathcal{B}_w(\mathrm{Exp}_p u_1, \dots, \mathrm{Exp}_p u_m) = \bar{u} + \frac{1}{3}\sum_{i=1}^{m} w_i R_p(\bar{u}, u_i)u_i + \mathcal{O}(r^4). \tag{17}$$

The remainder is uniform for the stated parameter ranges. In particular, for fixed v_i and $u_i = \varepsilon v_i$,

$$\mathrm{Log}_p \mathcal{B}_w(\mathrm{Exp}_p(\varepsilon v_1), \dots, \mathrm{Exp}_p(\varepsilon v_m)) = \varepsilon\bar{v} + \frac{\varepsilon^3}{3}\sum_i w_i R_p(\bar{v}, v_i)v_i + \mathcal{O}(\varepsilon^4), \tag{18}$$

where $\bar{v} = \sum_i w_i v_i$.

Proof. Let $z = \mathrm{Log}_p \mathcal{B}_w(\mathrm{Exp}_p u_1, \dots, \mathrm{Exp}_p u_m)$ and define $\Phi(z) = \sum_i w_i H(z, u_i)$. Since the differential of Exp_p is nonsingular in the chosen normal ball, the intrinsic Karcher equation is equivalent to $\nabla^{(p)}\Phi(z) = 0$. By lemma A.2,

$$0 = z - \bar{u} - \frac{1}{3}\sum_i w_i R_p(z, u_i)u_i + \mathcal{O}((\|z\| + r)^4). \tag{19}$$

At $(u_1, \dots, u_m, z) = (0, \dots, 0, 0)$, the derivative with respect to z is the identity for every w satisfying $\sum_i w_i = 1$. To apply the parameter-dependent implicit-function theorem at boundary weights, extend the smooth coordinate equation to an open neighborhood of the closed simplex within that affine hyperplane; negative weights in this auxiliary extension are used only for the theorem, not to define a layer. Compactness of the closed simplex gives a common small input ball and uniform derivative bounds. After shrinking this ball, the coordinate Hessian is uniformly positive definite, so the local branches agree on overlaps and give a unique smooth stationary solution with $z = \mathcal{O}(r)$. This solution lies in U and is therefore the stated local mean by uniqueness of its Karcher zero. Thus the uniform assertion includes zero weights and requires no unjustified inference from uniqueness to differentiability at remote input configurations.

Substituting $z = \mathcal{O}(r)$ into (19) first yields $z - \bar{u} = \mathcal{O}(r^3)$. Write $z = \bar{u} + \delta$ with $\delta = \mathcal{O}(r^3)$. Trilinearity of R_p gives $R_p(z, u_i)u_i = R_p(\bar{u}, u_i)u_i + R_p(\delta, u_i)u_i$, and the second term is $\mathcal{O}(r^5)$. Equation (19) now becomes $0 = \delta - \frac{1}{3}\sum_i w_i R_p(\bar{u}, u_i)u_i + \mathcal{O}(r^4)$, which proves (17). Setting $u_i = \varepsilon v_i$ gives (18). Uniformity follows from the uniform estimates in lemma A.2 and the parameter-uniform implicit-function theorem. □

Corollary A.5 (Binary expansion). *For $a, b \in T_p\mathcal{M}$ and $t \in [0,1]$, $\mathrm{Log}_p(\mathrm{Exp}_p(\varepsilon a)\#_t \mathrm{Exp}_p(\varepsilon b)) = \varepsilon((1-t)a + tb) + \frac{\varepsilon^3 t(1-t)}{3} R_p(a,b)(b-a) + \mathcal{O}(\varepsilon^4)$.*

Proof. For $t = 0$ or $t = 1$, the identity follows directly from the definition of $\#_t$. Assume $0 < t < 1$. Apply (18) with weights $(1-t, t)$ and $\mu = (1-t)a + tb$. The cubic coefficient is $\{(1-t)R(\mu, a)a + tR(\mu, b)b\}/3$. Because $R(\mu, a)a = -tR(a, b)a$ and $R(\mu, b)b = (1-t)R(a, b)b$, this coefficient is $t(1-t)R(a, b)(b-a)/3$. □

Remark A.6 (Parity on locally symmetric spaces). Suppose that the geodesic symmetry s_p is an isometry on the relevant neighborhood, as it is in a locally symmetric space, and recall that $s_p(\mathrm{Exp}_p v) = \mathrm{Exp}_p(-v)$. Shrink to an s_p-invariant normal ball inside the domain of this local isometry. For sufficiently small probes, the mean branches and all intermediate outputs lie there by the implicit-function theorem and continuity at the diagonal. The isometry transports the Karcher equation to itself, and uniqueness identifies the transported root. Weighted Fréchet means and every network assembled from them are consequently equivariant under s_p on these probes. Hence the normal-coordinate output of a probe $\mathrm{Exp}_p(\varepsilon v_i)$ is an odd function of ε. A collapse defect beginning at order ε^3 then has its next possible term at order ε^5, so the cubic defect divided by ε^3 has an even $\mathcal{O}(\varepsilon^2)$ bias. This exact parity statement, rather than an informal differentiation of an $\mathcal{O}$ term, explains the rate observed in the sphere experiment.

B Fixed-weight collapse of the ManifoldNet family

B.1 Exact identification of the candidate weights

We first prove the non-asymptotic part of theorem 3.1.

Proposition B.1 (Exact identifiability). *If $\omega \in \Delta_2$ and $F_{\alpha,\beta,\gamma} = \mathcal{B}_\omega$ on a neighborhood of (p, p, p), then $\omega = \omega^0$ from (3).*

Proof. Fix a nonzero $X \in T_p\mathcal{M}$ and take t sufficiently small. For the input $(A, B, C) = (\mathrm{Exp}_p(tX), p, p)$, radial exactness gives $A\#_\alpha B = \mathrm{Exp}_p((1-\alpha)tX)$ and $B\#_\beta C = p$. A second application of lemma A.1 yields $F_{\alpha,\beta,\gamma}(A, B, C) = \mathrm{Exp}_p((1-\gamma)(1-\alpha)tX)$. The one-layer side is $\mathrm{Exp}_p(\omega_A tX)$, so injectivity of Exp_p on the normal ball gives $\omega_A = (1-\gamma)(1-\alpha)$.

For $(A, B, C) = (p, \mathrm{Exp}_p(tX), p)$, the two inner outputs are $\mathrm{Exp}_p(\alpha tX)$ and $\mathrm{Exp}_p((1-\beta)tX)$. They lie on the same radial geodesic, whence $F_{\alpha,\beta,\gamma}(A, B, C) = \mathrm{Exp}_p(((1-\gamma)\alpha + \gamma(1-\beta))tX)$. The one-layer side is $\mathrm{Exp}_p(\omega_B tX)$, and therefore $\omega_B = (1-\gamma)\alpha + \gamma(1-\beta)$. Finally, the input $(p, p, \mathrm{Exp}_p(tX))$ gives $F_{\alpha,\beta,\gamma} = \mathrm{Exp}_p(\gamma\beta tX)$ and hence $\omega_C = \gamma\beta$. No curvature approximation is involved. □

B.2 The cubic defect for arbitrary tangent triples

We now prove the second assertion of theorem 3.1. The calculation is written in full because it simultaneously fixes the sign, shows that all terms depending on the common tangent displacement cancel, and supplies the formula needed later for random inputs.

Put $\ell = 1 - \alpha$, $\lambda = (1-\gamma)\ell$, $\eta = \gamma\beta$, and $g = g_{\alpha,\beta,\gamma} = \lambda\eta = \gamma(1-\gamma)(1-\alpha)\beta$. Thus $\omega^0 = (\lambda, 1-\lambda-\eta, \eta)$. Write $a = b + x$ and $c = b + y$. By corollary A.5, the logarithms of the two inner nodes have the form

$$\mathrm{Log}_p(A_\varepsilon\#_\alpha B_\varepsilon) = \varepsilon m_L + \varepsilon^3 q_L + \mathcal{O}(\varepsilon^4), \quad m_L = b + \ell x, \quad q_L = \frac{\alpha\ell}{3} R_p(b, x)x, \tag{20}$$

$$\mathrm{Log}_p(B_\varepsilon\#_\beta C_\varepsilon) = \varepsilon m_R + \varepsilon^3 q_R + \mathcal{O}(\varepsilon^4), \quad m_R = b + \beta y, \quad q_R = \frac{\beta(1-\beta)}{3} R_p(b, y)y. \tag{21}$$

For example, the left cubic term follows from $R_p(b + x, b)(b - (b + x)) = R_p(b, x)x$.

Apply the multivariable expansion (17) to the outer binary node. The cubic perturbations q_L, q_R contribute through their weighted average, while inserting them into the curvature term would change

that term only by $\mathcal{O}(\varepsilon^5)$. Therefore

$$\mathrm{Log}_p F_{\alpha,\beta,\gamma} = \varepsilon\mu + \varepsilon^3 Q_F + \mathcal{O}(\varepsilon^4), \qquad \mu = (1-\gamma)m_L + \gamma m_R = b + \lambda x + \eta y, \tag{22}$$

where

$$Q_F = (1-\gamma)q_L + \gamma q_R + \frac{\gamma(1-\gamma)}{3} R_p(m_L, m_R)(m_R - m_L). \tag{23}$$

The last curvature factor expands as

$$R_p(m_L, m_R) = \beta R_p(b, y) - \ell R_p(b, x) + \ell\beta R_p(x, y), \tag{24}$$

$$m_R - m_L = \beta y - \ell x. \tag{25}$$

Combining (20), (21), and (24)–(25), and using $\lambda(1-\lambda) = (1-\gamma)\ell(\alpha + \gamma\ell)$ and $\eta(1-\eta) = \gamma\beta(1-\beta+(1-\gamma)\beta)$, gives

$$Q_F = C_b + \frac{g}{3} R_p(x, y)(\beta y - \ell x), \tag{26}$$

where the common-displacement term is $C_b := \frac{1}{3}\Big[\lambda(1-\lambda)R_p(b,x)x - \lambda\eta R_p(b,x)y - \lambda\eta R_p(b,y)x + \eta(1-\eta)R_p(b,y)y\Big]$. For completeness, the coefficient of $R_p(b,x)x$ in (23) is $(1-\gamma)\alpha\ell + \gamma(1-\gamma)\ell^2 = \lambda(1-\lambda)$; the coefficient of $R_p(b,y)y$ is $\gamma\beta(1-\beta) + \gamma(1-\gamma)\beta^2 = \eta(1-\eta)$; and each mixed term has coefficient $-\gamma(1-\gamma)\ell\beta = -\lambda\eta$. This proves (26) without suppressing any b-dependent contribution.

We next expand the forced one-layer candidate. Its linear tangent mean is the same vector $\mu = b + \lambda x + \eta y$. By (18),

$$\mathrm{Log}_p \mathcal{B}_{\omega^0}(A_\varepsilon, B_\varepsilon, C_\varepsilon) = \varepsilon\mu + \varepsilon^3 Q_G + \mathcal{O}(\varepsilon^4), \tag{27}$$

with

$$3Q_G = \lambda R_p(\mu, b+x)(b+x) + (1-\lambda-\eta)R_p(\mu, b)b + \eta R_p(\mu, b+y)(b+y). \tag{28}$$

The three curvature operators in (28) are

$$R_p(\mu, b+x) = (1-\lambda)R_p(b,x) - \eta R_p(b,y) - \eta R_p(x,y), \tag{29}$$

$$R_p(\mu, b) = -\lambda R_p(b,x) - \eta R_p(b,y), \tag{30}$$

$$R_p(\mu, b+y) = (1-\eta)R_p(b,y) - \lambda R_p(b,x) + \lambda R_p(x,y). \tag{31}$$

After these operators are applied to $b+x$, b, and $b+y$, respectively, every term acting on b cancels. Indeed, the coefficient of $R_p(b,x)b$ is $\lambda(1-\lambda) - \lambda(1-\lambda-\eta) - \lambda\eta = 0$, and the coefficient of $R_p(b,y)b$ vanishes analogously; the two $R_p(x,y)b$ terms cancel pairwise. The remaining terms give

$$Q_G = C_b + \frac{g}{3} R_p(x, y)(y - x). \tag{32}$$

Subtracting (32) from (26) and using $1-\ell = \alpha$ gives $Q_F - Q_G = \frac{g}{3}R_p(x,y)\{\beta y - \ell x - (y-x)\} = \frac{g}{3}R_p(x,y)(\alpha x - (1-\beta)y)$. Together with (22) and (27), this proves (5). Uniformity follows from the uniform remainder in theorem A.4. The proof of theorem 3.1 is complete.

B.3 Detection of nonzero curvature

Lemma B.2 (Sectional lower bound)**.** *Let $X, Y \in T_p\mathcal{M}$ be orthonormal and put $K = K_p(X,Y)$. Then*

$$\langle R_p(X,Y)(\alpha X - (1-\beta)Y), X\rangle = -(1-\beta)K, \tag{33}$$

$$\langle R_p(X,Y)(\alpha X - (1-\beta)Y), Y\rangle = -\alpha K. \tag{34}$$

Consequently, (6) *holds.*

Proof. The endomorphism $R_p(X,Y)$ is skew-adjoint. Hence its diagonal components in the orthonormal pair vanish, while $\langle R_p(X,Y)X, Y\rangle = -\langle R_p(X,Y)Y, X\rangle = -K$. The two displayed projections follow immediately. The squared norm of a vector is at least the sum of the squares of its components in the orthonormal directions X and Y, giving $\|R_p(X,Y)(\alpha X - (1-\beta)Y)\|^2 \geq K^2((1-\beta)^2 + \alpha^2)$. □

For the equal architecture, a useful scalar projection is slightly more symmetric.

Lemma B.3 (Equal-architecture sectional identity)**.** *For arbitrary $X, Y \in T_p\mathcal{M}$, one has $\langle R_p(X,Y)(Y-X), X+Y\rangle = 2R_p(X,Y,Y,X)$. If X, Y are orthonormal, then $\|R_p(X,Y)(Y-X)\| \geq \sqrt{2}\,|K_p(X,Y)|$.*

Proof. Expanding the left-hand side and then applying Cauchy–Schwarz gives

$$\begin{aligned}\langle R(X,Y)(Y-X), X+Y\rangle &= R(X,Y,Y,X) + R(X,Y,Y,Y)\\ &\quad - R(X,Y,X,X) - R(X,Y,X,Y),\\ 2|K| \leq \|R(X,Y)(Y-X)\|\,\|X+Y\| &= \sqrt{2}\,\|R(X,Y)(Y-X)\|\,.\end{aligned}$$

The middle two curvature terms in the first identity vanish by skew-symmetry in the last pair, while $R(X,Y,X,Y) = -R(X,Y,Y,X)$. This proves both assertions. □

B.4 Flatness: necessity and exact sufficiency

The necessity in theorem 3.2 follows immediately from the cubic defect. Exact sufficiency requires a local affine development.

Lemma B.4 (Euclidean development of a flat convex neighborhood)**.** *Let U be a strongly convex normal neighborhood with $R|_U \equiv 0$. Then there are a convex Euclidean domain $V \subset \mathbb{R}^d$ and a Riemannian isometry $\Psi : U \to V$.*

Proof. Fix $p \in U$. Strong convexity makes $H(t,x) = \operatorname{Exp}_p(t \operatorname{Log}_p x)$ a contraction of U to p; in particular, U is simply connected. Because the Levi–Civita connection is flat on U, parallel transport is path-independent on this simply connected domain. Parallel-transport an orthonormal basis of $T_p\mathcal{M}$ to obtain a global parallel orthonormal frame $E_1, \dots, E_d$ on U, and let $\theta^1, \dots, \theta^d$ be its dual coframe.

Each θ^i is parallel. Since the Levi–Civita connection is torsion-free, $\mathrm{d}\theta^i(X,Y) = (\nabla_X\theta^i)(Y) - (\nabla_Y\theta^i)(X) + \theta^i(\nabla_X Y - \nabla_Y X - [X,Y]) = 0$; hence every θ^i is closed. Simple connectedness gives smooth functions $f^i : U \to \mathbb{R}$, normalized by $f^i(p) = 0$, with $\mathrm{d}f^i = \theta^i$. Define $\Psi = (f^1, \dots, f^d)$. For every $X \in T_x\mathcal{M}$, one has $\|D\Psi_x X\|^2_{\mathbb{R}^d} = \sum_i \theta^i(X)^2 = \|X\|^2_g$, so Ψ is a local isometry and a local diffeomorphism.

Let $x, y \in U$ and let γ be their unique geodesic in U. A local isometry preserves the Levi–Civita connection, hence $\Psi \circ \gamma$ is the Euclidean segment from $\Psi(x)$ to $\Psi(y)$. It follows that $\Psi(U)$ is convex. If $\Psi(x) = \Psi(y)$, that segment is constant; since $D\Psi$ is injective, $\dot\gamma = 0$ and $x = y$. Therefore Ψ is injective and is a Riemannian isometry of U onto the convex domain $V := \Psi(U)$. □

Lemma B.5 (Weighted means are affine in a flat development)**.** *Under the isometry of lemma B.4, one has $\Psi\big(\mathcal{B}_w(x_1, \dots, x_m)\big) = \sum_{i=1}^m w_i \Psi(x_i)$.*

Proof. The isometry carries the Fréchet energy to $z \mapsto \frac{1}{2}\sum_i w_i \|z - \Psi(x_i)\|^2$ on the convex Euclidean domain V. Its unique minimizer is the weighted average, which belongs to V by convexity. Applying Ψ^{-1} proves the assertion. □

Proof of theorem 3.2. If $R|_U = 0$, lemma B.5 turns both inner nodes and the outer node in (1) into affine combinations. Their composition has coefficients $(1-\gamma)(1-\alpha)$, $(1-\gamma)\alpha + \gamma(1-\beta)$, and $\gamma\beta$, so $F_{\alpha,\beta,\gamma} = \mathcal{B}_{\omega^0}$ exactly on U^3.

Conversely, suppose $F_{\alpha,\beta,\gamma} = \mathcal{B}_\omega$ on U^3. By proposition B.1, $\omega = \omega^0$. Fix $p \in U$ and choose a normal ball V_p whose closure is contained in U. The asserted equality restricts to V_p^3, so the expansion in theorem 3.1, carried out in normal coordinates at this arbitrary base point p, yields $R_p(X,Y)(\alpha X - (1-\beta)Y) = 0$ for every $X, Y \in T_p\mathcal{M}$. Taking X, Y orthonormal and applying either projection in lemma B.2 gives $K_p(X,Y) = 0$. An algebraic Riemann tensor is determined by its sectional numerators; the explicit polarization is given in theorem D.3. Hence $R_p = 0$. Since p was arbitrary, $R|_U \equiv 0$. Uniqueness of the weights is already contained in proposition B.1. □

Proof of theorem 4.2. The implication (i)⇒(ii) is immediate, and (ii)⇒(iii) is the equal-parameter case of theorem 3.2. If (iii) holds, use the isometry in lemma B.4. Every internal weighted-mean node becomes an affine map. Recursively distributing the node weights expresses the output as one Euclidean convex combination of the external inputs; the coefficient of an external input is the sum of the products of node weights along all root-to-leaf paths carrying that label. Applying Ψ^{-1} and lemma B.5 gives one fixed weighted Fréchet mean. This proves (iii)⇒(i). □

B.5 Resolution of the source conjecture and space forms

Corollary B.6 (Equal-weight ManifoldNet block)**.** *The only possible fixed one-layer weights for F_{eq} are $(1/4, 1/2, 1/4)$. For $A_\varepsilon = \mathrm{Exp}_p(\varepsilon X)$, $B_\varepsilon = p$, and $C_\varepsilon = \mathrm{Exp}_p(\varepsilon Y)$,*

$$\mathrm{Log}_p F_{\mathrm{eq}}(A_\varepsilon, p, C_\varepsilon) - \mathrm{Log}_p \mathcal{B}_{(1/4,1/2,1/4)}(A_\varepsilon, p, C_\varepsilon) = -\frac{\varepsilon^3}{96} R_p(X, Y)(Y - X) + \mathcal{O}(\varepsilon^4). \quad (35)$$

At a nonflat point, there are arbitrarily small triples for which the two outputs differ. More precisely, for every $w \in \Delta_2$ there are arbitrarily small inputs with $F_{\mathrm{eq}} \neq \mathcal{B}_w$: a radial input suffices when $w \neq (1/4, 1/2, 1/4)$, and the displayed curvature probe suffices for the remaining weight vector.

Proof. The weights follow from proposition B.1; the expansion is (5) with $\alpha = \beta = \gamma = 1/2$ and $(a, b, c) = (X, 0, Y)$. If $R_p \neq 0$, some orthonormal plane has $K_p(X, Y) \neq 0$. By lemma B.3, the cubic coefficient is nonzero, so the right-hand side of (35) is nonzero for all sufficiently small nonzero ε. Finally, if $w \neq (1/4, 1/2, 1/4)$, at least one coordinate of w differs from the path-product weight. Moving only the corresponding external input along a radial geodesic and fixing the other two at p gives an exact discrepancy already at first order, by the proof of proposition B.1. □

Lemma B.7 (Localization of concentrated global means)**.** *Assume that $\mathcal{M}$ is complete, fix $p \in \mathcal{M}$, and let U be a local mean domain containing p. Choose radii $0 < r < R/2$ such that the closed ball $\overline{B(p, R)}$ is contained in U and $B(p, r)$ is strongly convex. If $x_1, \ldots, x_m \in B(p, r)$ and $w \in \Delta_{m-1}$, then the global Fréchet energy $E_w(z) = \frac{1}{2} \sum_i w_i \, \mathrm{d}(z, x_i)^2$ has a unique minimizer on $\mathcal{M}$, and*

$$\mathcal{B}_w^{\mathrm{glob}}(x_1, \ldots, x_m) = \mathcal{B}_w(x_1, \ldots, x_m). \quad (36)$$

In particular, every binary weighted mean of two points of $B(p, r)$ is the corresponding geodesic interpolation and remains in $B(p, r)$; hence a finite binary network with inputs in $B(p, r)$ has the same value whether its nodes are interpreted as local or global Fréchet means.

Proof. Completeness and the Hopf–Rinow theorem imply that the continuous energy E_w attains its minimum. Indeed, for $d(p, z) \geq r$, the triangle inequality gives $d(z, x_i) \geq d(p, z) - r$ for every i, and therefore $E_w(z) \geq \frac{1}{2}(d(p, z) - r)^2$; thus every sublevel set is bounded, and closed bounded subsets of a complete finite-dimensional Riemannian manifold are compact.

If $z \notin B(p, R)$, then $d(z, x_i) \geq R - r$ for every i, so $E_w(z) \geq \frac{1}{2}(R - r)^2$. On the other hand, $E_w(p) < \frac{1}{2} r^2$. Since $R > 2r$, one has $(R - r)^2 > r^2$, and no global minimizer can lie outside $B(p, R)$. Every global minimizer therefore belongs to U, where the standing local well-posedness hypothesis gives the unique minimizer $\mathcal{B}_w(x_1, \ldots, x_m)$. This proves (36) and global uniqueness.

For two inputs, the local mean is the geodesic interpolation. Strong convexity of $B(p, r)$ keeps the entire interpolating segment inside that ball. Induction over the nodes proves the final assertion for finite binary networks. □

Corollary B.8 (Resolution of the fixed-weight ManifoldNet conjecture)**.** *Let $\mathcal{M}$ be a complete Riemannian manifold with nonconstant sectional curvature, as in the standing setting of Chakraborty et al. (2022). Fix the equal two-layer parameter vector u_0 corresponding to F_{eq}. Then, for every $w \in \Delta_2$, there exist points $A, B, C \in \mathcal{M}$, which may be chosen arbitrarily close to one another, such that*

$$F_{\mathrm{eq}}(A, B, C) \neq \mathcal{B}_w^{\mathrm{glob}}(A, B, C), \quad (37)$$

where each binary node in F_{eq} is also its unique global binary Fréchet mean. Thus the quantifier statement $\exists u \, \forall w \, \exists (A, B, C)$ in Appendix C of Chakraborty et al. (2022) holds.

Proof. Nonconstant sectional curvature excludes identically zero curvature, so there exists $p \in \mathcal{M}$ with $R_p \neq 0$. Choose a local mean domain U about p and radii r, R as in lemma B.7. The witnesses in corollary B.6 can be chosen inside $B(p, r)$ because they are obtained from arbitrarily short radial or curvature probes. For $w \neq (1/4, 1/2, 1/4)$, a radial probe gives an exact first-order discrepancy; for $w = (1/4, 1/2, 1/4)$, an orthonormal plane of nonzero sectional curvature gives a nonzero cubic discrepancy. By lemma B.7, all local means appearing in these comparisons agree with their unique global counterparts. Hence (37) holds for every w. □

The conclusion is stronger than the original nonconstant-curvature assertion: the same equal architecture fails to collapse on every nonflat local mean domain, including nonzero space forms.

Corollary B.9 (Space forms)**.** *Suppose $\mathcal{M}$ has constant sectional curvature κ. For orthonormal X, Y and the forced weights,*

$$\lim_{\varepsilon \downarrow 0} \varepsilon^{-3}\, \mathrm{d}\Big(F_{\alpha,\beta,\gamma}(\mathrm{Exp}_p(\varepsilon X), p, \mathrm{Exp}_p(\varepsilon Y)),\ \ \mathcal{B}_{\omega^0}(\mathrm{Exp}_p(\varepsilon X), p, \mathrm{Exp}_p(\varepsilon Y))\Big) = \frac{|\kappa|}{3} g_{\alpha,\beta,\gamma} \sqrt{\alpha^2 + (1-\beta)^2}. \tag{38}$$

Thus, on every sufficiently small local mean domain in a space form, fixed-weight collapse of the source family occurs exactly when $\kappa = 0$.

Proof. With our convention, $R(X, Y)Z = \kappa(\langle Y, Z\rangle X - \langle X, Z\rangle Y)$. Hence $R(X, Y)(\alpha X - (1 - \beta)Y) = -\kappa((1-\beta)X + \alpha Y)$, whose norm is $|\kappa|\sqrt{\alpha^2 + (1-\beta)^2}$. The normal-coordinate difference is given by (5). If u_ε and v_ε denote the two output logarithms, then $u_\varepsilon, v_\varepsilon = \mathcal{O}(\varepsilon)$, and lemma A.3 gives $\mathrm{d}(\mathrm{Exp}_p u_\varepsilon, \mathrm{Exp}_p v_\varepsilon) = \|u_\varepsilon - v_\varepsilon\|\,(1 + \mathcal{O}(\varepsilon^2))$. Combining this estimate with (5) proves the limit, including the case $\kappa = 0$. The classification follows from theorem 3.2. □

Appendix C of Chakraborty et al. (2022) also gives a constant-curvature construction in a fixed geodesic triangle. There the triangle is fixed first, the putative output of the equal-weight one-layer mean is identified with a triangle center, and only then is the outer two-layer weight selected so that this point lies on the geodesic joining the first-layer output to the third vertex. Even granting the geometric identification made in that argument, the resulting quantifier order is only $\forall(A, B, C)\, \exists u(A, B, C)$ with the one-layer equal weights fixed. It does not establish pointwise representability of any fixed two-layer map, which would require $\exists u_0\, \forall(A, B, C)\, \exists w(A, B, C)$, and a fortiori it does not establish fixed-weight functional collapse. Thus the construction does not prove the fixed-parameter constant-curvature claim, while corollary B.9 supplies an explicit local obstruction to that claim.

Corollary B.10 (Genuine three-leaf binary trees)**.** *Near a nonflat point, neither $(A\#_\alpha B)\#_\gamma C$ nor $A\#_\gamma(B\#_\beta C)$ is a fixed one-layer weighted Fréchet mean when all displayed parameters lie in $(0, 1)$. Their uniquely forced one-layer weights are, respectively, $((1-\gamma)(1-\alpha), (1-\gamma)\alpha, \gamma)$ and $(1-\gamma, \gamma(1-\beta), \gamma\beta)$. For the probe $(\mathrm{Exp}_p(\varepsilon X), p, \mathrm{Exp}_p(\varepsilon Y))$, the corresponding leading defects are*

$$\frac{\varepsilon^3}{3}\gamma(1-\gamma)(1-\alpha)\alpha R_p(X, Y)X, \qquad -\frac{\varepsilon^3}{3}\gamma(1-\gamma)\beta(1-\beta)R_p(X, Y)Y. \tag{39}$$

Proof. The asserted weights are forced without approximation. Set two external inputs equal to p and move the remaining input along $\mathrm{Exp}_p(tX)$. Repeated use of radial exactness shows that the root moves by the corresponding path-product coefficient; comparison with a one-layer mean therefore identifies each of the three weights. This argument applies separately to the two displayed trees.

For the left-associated tree, the inner output is $\mathrm{Exp}_p(\varepsilon(1-\alpha)X)$ by radial exactness. Applying corollary A.5 to the outer node gives the cubic coefficient $Q_L = \gamma(1-\gamma)(1-\alpha)R_p(X, Y)(Y - (1-\alpha)X)/3$. Its effective one-layer weights are $((1-\gamma)(1-\alpha), (1-\gamma)\alpha, \gamma)$, whose cubic coefficient on $(X, 0, Y)$ is $\gamma(1-\gamma)(1-\alpha)R_p(X, Y)(Y - X)/3$. Multiplying their difference by ε^3 gives the first vector in (39), with an $\mathcal{O}(\varepsilon^4)$ remainder in the output difference.

For the right-associated tree, the inner output is $\mathrm{Exp}_p(\varepsilon\beta Y)$. The outer binary expansion has cubic coefficient $Q_R = \gamma(1-\gamma)\beta R_p(X, Y)(\beta Y - X)/3$, whereas the effective one-layer weights $(1-\gamma, \gamma(1-\beta), \gamma\beta)$ give $\gamma(1-\gamma)\beta R_p(X, Y)(Y - X)/3$. Multiplying their difference by ε^3 gives the second vector in (39), again with an $\mathcal{O}(\varepsilon^4)$ remainder in the output difference.

If X, Y are orthonormal and $K_p(X,Y) \neq 0$, then $\langle R_p(X,Y)X, Y\rangle = -K_p(X,Y)$ and $\langle R_p(X,Y)Y, X\rangle = K_p(X,Y)$, so the respective coefficients are nonzero. □

C Arbitrary weighted-mean trees and quantitative consequences

C.1 Trees, effective weights, and the cubic recursion

Let $\mathcal{T}$ be a finite rooted tree. Each leaf ℓ carries a label $i(\ell) \in \{1, \dots, m\}$, and every label is assumed to occur at least once. An internal node N with children $N_1, \dots, N_k$ carries a fixed weight vector $\theta^N = (\theta_1^N, \dots, \theta_k^N)$ in the open simplex and returns the weighted Fréchet mean of its child outputs. Repeated labels permit a directed acyclic network with shared external inputs to be unfolded into such a tree without changing its function.

For a leaf ℓ, let π_ℓ be the product of the node weights along the root-to-leaf path. The effective weight of external label i is

$$\omega_i^{\mathcal{T}} := \sum_{\ell:\, i(\ell)=i} \pi_\ell. \tag{40}$$

These weights are positive and sum to one.

Proposition C.1 (Effective weights are forced)**.** *Let $F_{\mathcal{T}}$ be the function represented by the tree. If $w \in \Delta_{m-1}$ and $F_{\mathcal{T}}$ and a one-layer mean $\mathcal{B}_w$ have the same first jet at the diagonal $(p, \dots, p)$, then $w = \omega^{\mathcal{T}}$. In particular, every exact fixed-weight representation of $F_{\mathcal{T}}$ uses the effective path-product weights.*

Proof. Fix a label i and a nonzero $X \in T_p\mathcal{M}$. Put the ith external input equal to $\mathrm{Exp}_p(tX)$ and every other external input equal to p. Every leaf value then lies on the same radial geodesic. Induction from the leaves to the root, using lemma A.1 at each internal node, gives $F_{\mathcal{T}}(p, \dots, \mathrm{Exp}_p(tX), \dots, p) = \mathrm{Exp}_p(\omega_i^{\mathcal{T}} tX)$. The corresponding one-layer mean is $\mathrm{Exp}_p(w_i tX)$. Equality of first jets implies $\omega_i^{\mathcal{T}} X = w_i X$, hence $w_i = \omega_i^{\mathcal{T}}$. Since i was arbitrary, $w = \omega^{\mathcal{T}}$. Exact equality near the diagonal implies equality of first jets and is therefore covered by the same argument. □

Take tangent probes $x_i(\rho) = \mathrm{Exp}_p(\rho v_i)$ and let $F_N(\rho)$ denote the output of the subtree rooted at N.

Theorem C.2 (Cubic tree recursion)**.** *There are vectors $\mu_N, Q_N \in T_p\mathcal{M}$ such that $\mathrm{Log}_p F_N(\rho) = \rho\mu_N + \rho^3 Q_N + \mathcal{O}(\rho^4)$. At a leaf ℓ labeled by i, one has $\mu_\ell = v_i$ and $Q_\ell = 0$. If N has children $N_1, \dots, N_k$ and weights $\theta_1, \dots, \theta_k$, then*

$$\mu_N = \sum_{j=1}^{k} \theta_j \mu_{N_j}, \tag{41}$$

$$Q_N = \sum_{j=1}^{k} \theta_j Q_{N_j} + \frac{1}{3} \sum_{j=1}^{k} \theta_j R_p(\mu_N, \mu_{N_j})\mu_{N_j}. \tag{42}$$

At the root, $\mu_{\mathcal{T}} = \sum_i \omega_i^{\mathcal{T}} v_i$. For a fixed finite tree, all remainders are uniform when the tangent tuple ranges over a compact set and the node weights range over compact subsets of their open simplices.

Proof. The assertion is immediate at a leaf. Suppose it holds for all children of an internal node N, and write $u_j(\rho) = \mathrm{Log}_p F_{N_j}(\rho) = \rho\mu_{N_j} + \rho^3 Q_{N_j} + \mathcal{O}(\rho^4)$. Their weighted coordinate average is

$$\bar{u}_N(\rho) = \sum_j \theta_j u_j(\rho) = \rho \sum_j \theta_j \mu_{N_j} + \rho^3 \sum_j \theta_j Q_{N_j} + \mathcal{O}(\rho^4). \tag{43}$$

Apply the multivariable expansion (17) directly to the ρ-dependent inputs $u_j(\rho)$. This is the point at which a formula stated only for fixed directions would be insufficient. Since the curvature term is trilinear,

$$\frac{1}{3} \sum_j \theta_j R_p(\bar{u}_N(\rho), u_j(\rho))u_j(\rho) = \frac{\rho^3}{3} \sum_j \theta_j R_p\left(\sum_s \theta_s \mu_{N_s}, \mu_{N_j}\right)\mu_{N_j} + \mathcal{O}(\rho^5). \tag{44}$$

The fourth-order remainder in (17) dominates the displayed $\mathcal{O}(\rho^5)$ correction. Combining (43) and (44) gives (41)–(42).

The formula for $\mu_{\mathcal{T}}$ follows by induction: each passage through an internal node multiplies a leaf contribution by the corresponding node weight, and repeated labels are summed. The same induction, together with the uniformity in theorem A.4, proves the stated uniform remainder. □

The one-layer mean with the effective weights has cubic coefficient $Q^{\mathcal{T}}_\star(v_1,\dots,v_m) := \frac{1}{3}\sum_{i=1}^m \omega_i^{\mathcal{T}} R_p(\mu_{\mathcal{T}}, v_i)v_i$.

Definition C.3 (Architecture defect)**.** The cubic architecture defect of $\mathcal{T}$ at p is $\mathfrak{D}^p_{\mathcal{T}}(v_1,\dots,v_m) := Q_{\mathcal{T}}(v_1,\dots,v_m) - Q^{\mathcal{T}}_\star(v_1,\dots,v_m)$. It is homogeneous of degree three in the tangent tuple and linear in R_p.

Corollary C.4 (General local expansion and cubic certificate)**.** *Let $F_{\mathcal{T}}$ be the tree function and let $G_{\mathcal{T}} = \mathcal{B}_{\omega^{\mathcal{T}}}$ be its effective one-layer mean. Then*

$$\operatorname{Log}_p F_{\mathcal{T}}(x_1(\rho),\dots,x_m(\rho)) - \operatorname{Log}_p G_{\mathcal{T}}(x_1(\rho),\dots,x_m(\rho)) = \rho^3 \mathfrak{D}^p_{\mathcal{T}}(v_1,\dots,v_m) + \mathcal{O}(\rho^4). \quad (45)$$

If $\mathfrak{D}^p_{\mathcal{T}}(v) \neq 0$ for some tangent tuple v, then $F_{\mathcal{T}}$ has no exact fixed-weight one-layer representation on any neighborhood of the diagonal at p. If $\mathfrak{D}^p_{\mathcal{T}} \equiv 0$, the tree and its uniquely forced candidate have the same third jet, but no conclusion about exact collapse follows without higher-order analysis.

Proof. Apply theorem C.2 to the network and (18) to the one-layer mean. Their linear terms agree by (40); subtracting the cubic coefficients gives (45). If an exact fixed-weight representation existed, proposition C.1 would force its weights to be $\omega^{\mathcal{T}}$, and the left-hand side of (45) would vanish identically. A nonzero cubic coefficient rules this out. When the coefficient vanishes identically, the expansion yields only an $\mathcal{O}(\rho^4)$ remainder, which is precisely equality of third jets and does not control higher-order terms. □

C.2 Vanishing cubic defect does not imply exact collapse

The preceding corollary is deliberately one-sided. The next proposition gives an explicit tree for which the entire cubic defect cancels although exact collapse fails.

Proposition C.5 (A fifth-order obstruction with zero cubic defect)**.** *Define the three cyclic trees $T_{AB|C} := (A\#_{1/2}B)\#_{1/3}C$, $T_{BC|A} := (B\#_{1/2}C)\#_{1/3}A$, and $T_{CA|B} := (C\#_{1/2}A)\#_{1/3}B$, and set $H(A,B,C) := \mathcal{B}_{(1/3,1/3,1/3)}(T_{AB|C}, T_{BC|A}, T_{CA|B})$. Then the effective external weights of H are $(1/3,1/3,1/3)$ and $\mathfrak{D}^p_H \equiv 0$ for every base point p and every algebraic curvature tensor R_p. Nevertheless, on the unit sphere $\mathbb{S}^2$, if $p=(0,0,1)$ and $X=(1,0,0)$, $Y=(0,1,0)$, then*

$$\begin{aligned}&\operatorname{Log}_p H(\operatorname{Exp}_p(\varepsilon X), p, \operatorname{Exp}_p(\varepsilon Y))\\ &\quad - \operatorname{Log}_p \mathcal{B}_{(1/3,1/3,1/3)}(\operatorname{Exp}_p(\varepsilon X), p, \operatorname{Exp}_p(\varepsilon Y)) = \frac{7\varepsilon^5}{29160}(X+Y) + \mathcal{O}(\varepsilon^7).\end{aligned} \quad (46)$$

Consequently, H has no exact fixed-weight one-layer representation on any neighborhood of the diagonal at p in $\mathbb{S}^2$, despite having identically vanishing cubic architecture defect.

Proof. Each cyclic subtree has effective weights $(1/3,1/3,1/3)$, and hence so does H. For $a,b,c \in T_p\mathcal{M}$, set $\mathcal{S}(a,b,c) := R(a,b)(b-a) + R(b,c)(c-b) + R(c,a)(a-c)$. For $T_{AB|C}$, the binary formula and the tree recursion give

$$Q_{AB|C} = \frac{1}{18}R(a,b)(b-a) + \frac{1}{27}R(a+b,c)\left(c - \frac{a+b}{2}\right). \quad (47)$$

The two other coefficients are obtained by cyclic permutation. Using only $R(u,v) = -R(v,u)$, direct expansion gives $\sum_{\text{cyc}} R(a+b,c)(c-(a+b)/2) = \frac{3}{2}\mathcal{S}(a,b,c)$. Indeed, the terms containing $R(a,b)$ combine to $\frac{3}{2}R(a,b)(b-a)$, and the $R(b,c)$ and $R(c,a)$ terms combine in the same way. Summing (47) cyclically therefore yields $\sum_{\text{cyc}} Q_{AB|C} = \mathcal{S}/9$.

All three cyclic subtrees have the same linear tangent mean $\mu = (a+b+c)/3$. The local curvature term at the root of H is consequently zero, because every child linear mean equals the parent linear

mean. Thus $Q_H = \frac{1}{3}\sum_{\mathrm{cyc}} Q_{AB|C} = \mathcal{S}/27$. On the other hand, $Q_\star = \frac{1}{9}\sum_{v\in\{a,b,c\}} R(\mu,v)v = \mathcal{S}(a,b,c)/27$, so $\mathfrak{D}_H^p = Q_H - Q_\star = 0$ identically.

It remains to show that the cancellation does not persist to all orders. Work on $\mathbb{S}^2$ and put $A_\varepsilon = (\sin\varepsilon, 0, \cos\varepsilon)$, $B_\varepsilon = p$, and $C_\varepsilon = (0, \sin\varepsilon, \cos\varepsilon)$. For unit vectors u, v with $\theta = \arccos\langle u, v\rangle$, spherical interpolation, with its continuous value at $u = v$, is $u\#_t v = \frac{\sin((1-t)\theta)}{\sin\theta}u + \frac{\sin(t\theta)}{\sin\theta}v$. For $\varepsilon > 0$, set $q_\varepsilon = A_\varepsilon \#_{1/2} B_\varepsilon = (\sin(\varepsilon/2), 0, \cos(\varepsilon/2))$. Also set $\theta_\varepsilon = \arccos\langle q_\varepsilon, C_\varepsilon\rangle = \arccos(\cos(\varepsilon/2)\cos\varepsilon)$. Then

$$z_1 := T_{AB|C}(A_\varepsilon, B_\varepsilon, C_\varepsilon) = \frac{\sin(2\theta_\varepsilon/3)}{\sin\theta_\varepsilon} q_\varepsilon + \frac{\sin(\theta_\varepsilon/3)}{\sin\theta_\varepsilon} C_\varepsilon. \tag{48}$$

The second cyclic child z_2 is obtained from z_1 by interchanging its first two coordinates. For the third child, put $r_\varepsilon = (A_\varepsilon + C_\varepsilon)/\left\|A_\varepsilon + C_\varepsilon\right\|_{\mathbb{R}^3}$ and $\phi_\varepsilon = \arccos\langle p, r_\varepsilon\rangle$. Since r_ε lies on the radial geodesic generated by $(X+Y)/\sqrt{2}$, radial exactness gives

$$z_3 := T_{CA|B}(A_\varepsilon, B_\varepsilon, C_\varepsilon) = \mathrm{Exp}_p\left(\frac{2\phi_\varepsilon}{3\sqrt{2}}(X+Y)\right). \tag{49}$$

The angles satisfy $\cos\theta_\varepsilon = \cos(\varepsilon/2)\cos\varepsilon$ and, for $\varepsilon > 0$, $\tan\phi_\varepsilon = \tan\varepsilon/\sqrt{2}$. Expanding these two exact identities at the origin gives $\theta_\varepsilon = \sqrt{5}\,\varepsilon/2 - \sqrt{5}\,\varepsilon^3/60 - \sqrt{5}\,\varepsilon^5/600 + \mathcal{O}(\varepsilon^7)$ and $\phi_\varepsilon = \varepsilon/\sqrt{2} + \varepsilon^3/(6\sqrt{2}) + \varepsilon^5/(60\sqrt{2}) + \mathcal{O}(\varepsilon^7)$. Write $z_1 = (u_1, u_2, w_1)$ and $z_3 = (u_3, u_3, w_3)$. Equations (48) and (49) give the exact coordinate identities

$$u_1 = \frac{\sin(2\theta_\varepsilon/3)}{\sin\theta_\varepsilon}\sin(\varepsilon/2), \qquad u_2 = \frac{\sin(\theta_\varepsilon/3)}{\sin\theta_\varepsilon}\sin\varepsilon,$$
$$w_1 = \frac{\sin(2\theta_\varepsilon/3)}{\sin\theta_\varepsilon}\cos(\varepsilon/2) + \frac{\sin(\theta_\varepsilon/3)}{\sin\theta_\varepsilon}\cos\varepsilon, \quad u_3 = \frac{\sin(2\phi_\varepsilon/3)}{\sqrt{2}}, \qquad w_3 = \cos(2\phi_\varepsilon/3).$$

Taylor expansion of these identities gives the following one-sided coefficients; smoothness of the interpolation maps and the parity argument below yield the corresponding two-sided Taylor expansions. The second child is $z_2 = (u_2, u_1, w_1)$, and

$$u_1 = \frac{\varepsilon}{3} + \frac{2\varepsilon^3}{81} + \frac{2\varepsilon^5}{3645} + \mathcal{O}(\varepsilon^7), \qquad u_2 = \frac{\varepsilon}{3} + \frac{\varepsilon^3}{162} - \frac{89\varepsilon^5}{29160} + \mathcal{O}(\varepsilon^7),$$
$$u_3 = \frac{\varepsilon}{3} + \frac{7\varepsilon^3}{162} - \frac{7\varepsilon^5}{14580} + \mathcal{O}(\varepsilon^7), \qquad w_1 = 1 - \frac{\varepsilon^2}{9} - \frac{4\varepsilon^4}{243} - \frac{173\varepsilon^6}{131220} + \mathcal{O}(\varepsilon^8),$$
$$w_3 = 1 - \frac{\varepsilon^2}{9} - \frac{17\varepsilon^4}{486} - \frac{713\varepsilon^6}{131220} + \mathcal{O}(\varepsilon^8).$$

The coordinate interchange $\sigma(x_1, x_2, x_3) = (x_2, x_1, x_3)$ is an isometry fixing the two input sets $\{A_\varepsilon, B_\varepsilon, C_\varepsilon\}$ and $\{z_1, z_2, z_3\}$. Uniqueness of the relevant means implies that both means lie on the fixed geodesic of σ. Hence each has the form $m(k) = \mathrm{Exp}_p(k(X+Y))$, with $m(k) = \left(\frac{\sin(\sqrt{2}k)}{\sqrt{2}}, \frac{\sin(\sqrt{2}k)}{\sqrt{2}}, \cos(\sqrt{2}k)\right)$. For a unit vector z, put $q = \langle m(k), z\rangle$ and $\psi(q) = \arccos(q)/\sqrt{1-q^2}$, with the analytic continuation $\psi(1) = 1$. Since $\langle m'(k), m(k)\rangle = 0$, the component of the Karcher equation along the fixed tangent line is

$$\sum_j \psi(\langle m(k), z_j\rangle)\langle m'(k), z_j\rangle = 0. \tag{50}$$

The full Karcher vector is fixed by σ and tangent at $m(k)$, so it lies on that one-dimensional line; thus (50) is equivalent to the full equation. The analytic factor satisfies $\psi(1-s) = 1 + s/3 + 2s^2/15 + 2s^3/35 + \mathcal{O}(s^4)$. The geodesic symmetry at p maps the input triple at ε to the input triple at $-\varepsilon$, and both constructions are equivariant under this isometry. Their normal-coordinate outputs are therefore odd in ε. We may consequently write $k = a_1\varepsilon + a_3\varepsilon^3 + a_5\varepsilon^5 + \mathcal{O}(\varepsilon^7)$. Substitution into (50) for the direct input set $(A_\varepsilon, B_\varepsilon, C_\varepsilon)$ gives

$$0 = (2 - 6a_1)\varepsilon + \left(\frac{2a_1}{3} - 6a_3\right)\varepsilon^3$$

$$+\left(\frac{8a_1^3}{45}-\frac{4a_1^2}{15}+\frac{2a_1}{45}+\frac{2a_3}{3}-6a_5\right)\varepsilon^5+\mathcal{O}(\varepsilon^7),$$

so $a_1 = 1/3$, $a_3 = 1/27$, and $a_5 = 2/729$. For the child set (z_1, z_2, z_3), the displayed coordinate expansions give

$$0=(2-6a_1)\varepsilon+\left(\frac{2}{9}-6a_3\right)\varepsilon^3+\left(\frac{29}{1620}-6a_5\right)\varepsilon^5+\mathcal{O}(\varepsilon^7),$$

so $a_1 = 1/3$, $a_3 = 1/27$, and $a_5 = 29/9720$. Therefore

$$\mathrm{Log}_p \mathcal{B}_{(1/3,1/3,1/3)}(A_\varepsilon, B_\varepsilon, C_\varepsilon)=\left(\frac{\varepsilon}{3}+\frac{\varepsilon^3}{27}+\frac{2\varepsilon^5}{729}\right)(X+Y)+\mathcal{O}(\varepsilon^7),$$
$$\mathrm{Log}_p H(A_\varepsilon, B_\varepsilon, C_\varepsilon)=\left(\frac{\varepsilon}{3}+\frac{\varepsilon^3}{27}+\frac{29\varepsilon^5}{9720}\right)(X+Y)+\mathcal{O}(\varepsilon^7).$$

Their difference is (46), because $29/9720 - 2/729 = 7/29160$. By proposition C.1, any exact one-layer representation of H would have to use the equal weights, so the nonzero fifth-order term excludes every fixed one-layer representation. □

C.3 Translation invariance of the cubic architecture defect

The full three-input formula in theorem 3.1 depends only on $a - b$ and $c - b$. The next theorem shows that this is a general property of the architecture defect.

Theorem C.6 (Common-shift covariance and defect invariance)**.** *For a subtree N, let ω_i^N be its effective external weights, with $\omega_i^N = 0$ for labels absent from N, and let $\mu_N = \sum_i \omega_i^N v_i$. For $z \in T_p\mathcal{M}$, write $v + z = (v_1 + z, \dots, v_m + z)$. Then*

$$Q_N(v+z)-Q_N(v)=\frac{1}{3}\left\{R_p(\mu_N,z)\mu_N+\sum_i \omega_i^N R_p(z,v_i)v_i\right\}. \tag{51}$$

The one-layer coefficient $Q_\star^N$ has the same transformation law. Consequently,

$$\mathfrak{D}_N^p(v_1+z,\dots,v_m+z)=\mathfrak{D}_N^p(v_1,\dots,v_m). \tag{52}$$

Proof. At a leaf labeled by i, both sides of (51) vanish: $Q_\ell = 0$, $\mu_\ell = v_i$, and $R(v_i, z)v_i + R(z, v_i)v_i = 0$. Assume the formula holds for all children of an internal node N. Under the common shift, every child mean and the parent mean change by z. The change in the local curvature term of (42) is

$$\frac{1}{3}\sum_j \theta_j\left[R_p(\mu_N+z,\mu_{N_j}+z)(\mu_{N_j}+z)-R_p(\mu_N,\mu_{N_j})\mu_{N_j}\right]$$
$$=\frac{1}{3}\left\{R_p(\mu_N,z)\mu_N+\sum_j \theta_j R_p(z,\mu_{N_j})\mu_{N_j}\right\}. \tag{53}$$

To verify (53), expand all factors by trilinearity. The terms $\sum_j \theta_j R_p(\mu_N, \mu_{N_j})z$ vanish because $\sum_j \theta_j \mu_{N_j} = \mu_N$; the terms $R_p(\mu_N, z)z$ and $\sum_j \theta_j R_p(z, \mu_{N_j})z = R_p(z, \mu_N)z$ cancel; and the remaining terms are precisely those displayed.

The weighted sum of the child transformations is

$$\frac{1}{3}\sum_j \theta_j R_p(\mu_{N_j},z)\mu_{N_j}+\frac{1}{3}\sum_i \omega_i^N R_p(z,v_i)v_i. \tag{54}$$

The first term in (54) cancels the second term in (53), because $R(\mu_{N_j}, z) = -R(z, \mu_{N_j})$. The surviving terms give (51).

For the one-layer coefficient, write $\mu = \sum_i \omega_i v_i$. Expanding $\sum_i \omega_i R(\mu + z, v_i + z)(v_i + z)$ shows that $\sum_i \omega_i R(\mu, v_i)z = R(\mu, \mu)z = 0$ and that $R(\mu, z)z$ cancels $\sum_i \omega_i R(z, v_i)z = R(z, \mu)z$. What remains is exactly the right-hand side of (51). Subtracting the two transformation laws proves (52). □

Remark C.7. The common shift is an algebraic invariance of the cubic tangent polynomial; it is not a claim that a curved manifold admits a global translation action. It is nevertheless the correct coordinate statement: a simultaneous first-order displacement of all leaves changes the tree and one-layer cubic terms in the same way, so their difference depends only on relative tangent directions.

C.4 Depth-dependent bounds with a controlled remainder

Define the curvature operator norm by $\|R_p\|_{\mathrm{op}} = \sup_{\|x\|=\|y\|=\|z\|=1} \|R_p(x,y)z\|$. For a subtree, define its weighted internal depth recursively by $L_\ell = 0$ at leaves and $L_N = 1 + \sum_j \theta_j L_{N_j}$ at an internal node. Equivalently, $L_{\mathcal{T}} = \sum_\ell \pi_\ell d_\ell$, where d_ℓ is the number of internal nodes on the root-to-leaf path.

Theorem C.8 (Cubic depth bound)**.** *If* $\max_i \|v_i\| \le V$*, then every subtree satisfies* $\|\mu_N\| \le V$ *and* $\|Q_N\| \le L_N \|R_p\|_{\mathrm{op}} V^3/3$*. At the root,*

$$\|\mathfrak{D}^p_{\mathcal{T}}(v_1,\dots,v_m)\| \le \frac{L_{\mathcal{T}}+1}{3}\|R_p\|_{\mathrm{op}} V^3. \tag{55}$$

Proof. The bound on μ_N follows inductively from (41) and convexity of the norm. At an internal node, (42) and the induction hypothesis give

$$\begin{aligned}\|Q_N\| &\le \sum_j \theta_j \|Q_{N_j}\| + \frac{1}{3}\sum_j \theta_j \|R_p\|_{\mathrm{op}} \|\mu_N\| \|\mu_{N_j}\|^2 \\ &\le \frac{\|R_p\|_{\mathrm{op}} V^3}{3}\left(\sum_j \theta_j L_{N_j} + 1\right) = \frac{L_N}{3}\|R_p\|_{\mathrm{op}} V^3.\end{aligned}$$

The one-layer coefficient obeys $\|Q^{\mathcal{T}}_\star\| \le \frac{1}{3}\sum_i \omega^{\mathcal{T}}_i \|R_p\|_{\mathrm{op}} \|\mu_{\mathcal{T}}\| \|v_i\|^2 \le \|R_p\|_{\mathrm{op}} V^3/3$. The triangle inequality gives (55). □

A cubic coefficient bound is not, by itself, a finite-radius error bound. The fourth-order remainder must be retained.

Theorem C.9 (Uniform finite-radius estimate)**.** *Fix* p*, a finite tree* $\mathcal{T}$*, compact sets of admissible node weights, and the normalized tangent set* $K = \{(v_1,\dots,v_m) : \max_i \|v_i\| \le 1\}$*. There exist* $\rho_0 > 0$ *and* $C_4 < \infty$ *such that, for* $0 < \rho < \rho_0$ *and* $v \in K$*,*

$$\mathrm{d}(F_{\mathcal{T}}(\mathrm{Exp}_p(\rho v_1),\dots,\mathrm{Exp}_p(\rho v_m)), G_{\mathcal{T}}(\mathrm{Exp}_p(\rho v_1),\dots,\mathrm{Exp}_p(\rho v_m))) \le \rho^3 \|\mathfrak{D}^p_{\mathcal{T}}(v)\| + C_4\rho^4. \tag{56}$$

Consequently, for the same normalized probes,

$$\begin{aligned}&\mathrm{d}\big(F_{\mathcal{T}}(\mathrm{Exp}_p(\rho v_1),\dots,\mathrm{Exp}_p(\rho v_m)), G_{\mathcal{T}}(\mathrm{Exp}_p(\rho v_1),\dots,\mathrm{Exp}_p(\rho v_m))\big) \\ &\qquad \le \frac{L_{\mathcal{T}}+1}{3}\|R_p\|_{\mathrm{op}}\rho^3 + C_4\rho^4.\end{aligned} \tag{57}$$

Proof. The compact-uniform version of (45) gives normal-coordinate outputs $u_\rho, v_\rho = \mathcal{O}(\rho)$ such that $u_\rho - v_\rho = \rho^3\mathfrak{D}^p_{\mathcal{T}}(v) + r_\rho(v)$ and $\|r_\rho(v)\| \le C\rho^4$, uniformly for $v \in K$. After decreasing ρ_0, the coordinate segment $z_t = (1-t)u_\rho + tv_\rho$ remains in a ball of radius $C_1\rho$. The normal-coordinate metric satisfies $g_z(\xi,\xi) \le (1 + C_0\rho^2)\|\xi\|^2$ on that ball. The length of the coordinate segment therefore yields

$$\begin{aligned}\mathrm{d}(\mathrm{Exp}_p u_\rho, \mathrm{Exp}_p v_\rho) &\le (1 + C_2\rho^2)\|u_\rho - v_\rho\| \\ &\le \rho^3\|\mathfrak{D}^p_{\mathcal{T}}(v)\| + C_3\rho^4.\end{aligned}$$

Here the term $C_2\rho^5\|\mathfrak{D}^p_{\mathcal{T}}(v)\|$ is absorbed into $C_3\rho^4$ because $\mathfrak{D}^p_{\mathcal{T}}$ is bounded on the compact set K and $0 < \rho < \rho_0$. Enlarging the constant proves (56); theorem C.8 with $V = 1$ gives (57). □

Corollary C.10 (Certified local fusion criterion)**.** *Under the hypotheses of theorem C.9, and once a valid bound* C_4 *has been supplied, replacing* $F_{\mathcal{T}}$ *by* $G_{\mathcal{T}}$ *incurs error at most* τ *whenever* $(L_{\mathcal{T}} + 1)\|R_p\|_{\mathrm{op}}\rho^3/3 + C_4\rho^4 \le \tau$*. The leading cubic inequality obtained by deleting* $C_4\rho^4$ *is only an asymptotic heuristic, not a sufficient finite-radius condition.*

A valid constant C_4 may be bounded in terms of the uniform C^1 world-function remainder in lemma A.2, the normal-coordinate metric bounds, the finite architecture, the admissible parameter ranges, and the chosen neighborhood. It is therefore a geometric–analytic constant, not an implementation-dependent one. If Exp, Log, or the Fréchet means are evaluated approximately, the resulting solver error is a separate term and must be added to (56).

C.5 Directional scale, and why no universal lower law exists

For the source family and unit vectors X, Y, write

$$d_p^{\alpha,\beta,\gamma}(\rho; X, Y) := \mathrm{d}\Big(F_{\alpha,\beta,\gamma}(\mathrm{Exp}_p(\rho X), p, \mathrm{Exp}_p(\rho Y)), \\ \mathcal{B}_{\omega^0}(\mathrm{Exp}_p(\rho X), p, \mathrm{Exp}_p(\rho Y))\Big)$$

Set $\Xi_p(X, Y; \alpha, \beta) := \|R_p(X, Y)(\alpha X - (1-\beta)Y)\|$. Equation (5), together with lemma A.3, gives

$$d_p^{\alpha,\beta,\gamma}(\rho; X, Y) = \frac{g_{\alpha,\beta,\gamma}}{3}\Xi_p(X, Y; \alpha, \beta)\rho^3 + \mathcal{O}(\rho^4). \tag{58}$$

Since the maximal input displacement from p is ρ, division of (58) by ρ gives the normalized expansion $d_p^{\alpha,\beta,\gamma}(\rho; X, Y)/\rho = (g_{\alpha,\beta,\gamma}/3)\Xi_p(X, Y; \alpha, \beta)\rho^2 + \mathcal{O}(\rho^3)$. In particular, for fixed unit directions there is a constant C such that $d_p^{\alpha,\beta,\gamma}(\rho; X, Y)/\rho \le (g_{\alpha,\beta,\gamma}/3)\,\|R_p\|_{\mathrm{op}}\,\|\alpha X - (1-\beta)Y\|\,\rho^2 + C\rho^3$. The constant implicit in the final remainder is attached to the fixed metric and the chosen neighborhood; it is not determined by the pointwise tensor R_p alone. More precisely, no estimate uniform over a class of metrics can, in general, bound the fourth-order remainder solely through R_p: the homogeneous fourth-order term of a Fréchet-mean expansion generally contains ∇R_p, equivalently the next metric jet. A class-uniform finite-radius bound must therefore control that jet, or an equivalent neighborhood C^5 bound, in addition to $\|R_p\|_{\mathrm{op}}$ (Pennec, 2019).

There is no two-sided estimate depending only on a curvature scale and ρ^3. On a nonflat space form, take $X = Y$. All three inputs then lie on one geodesic, and lemma A.1 makes the hierarchical and one-layer means exactly equal, although $|\kappa| > 0$ and $\rho > 0$. More generally, fix an orthonormal pair X, Z and put $Y_\vartheta = \cos\vartheta\, X + \sin\vartheta\, Z$. Whenever $\sin\vartheta \neq 0$—in particular, for every sufficiently small nonzero ϑ—the plane spanned by X and Y_ϑ is the fixed plane $\mathrm{span}\{X, Z\}$, but $\Xi_p(X, Y_\vartheta; \alpha, \beta) \to 0$ as $\vartheta \to 0$. A $\Theta(|\kappa|\rho^3)$ statement is therefore valid only after a nondegeneracy assumption supplies a positive lower bound for the directional factor Ξ_p.

C.6 Random-input regularization and parameter identifiability

Let a, b, c be random tangent vectors supported in a fixed compact subset of $T_p\mathcal{M}$, and define $A_\rho = \mathrm{Exp}_p(\rho a)$, $B_\rho = \mathrm{Exp}_p(\rho b)$, $C_\rho = \mathrm{Exp}_p(\rho c)$. Put $x = a - b$ and $y = c - b$.

Proposition C.11 (Small-spread collapse regularizer)**.** *Let* $F_\rho = F_{\alpha,\beta,\gamma}(A_\rho, B_\rho, C_\rho)$, $G_\rho = \mathcal{B}_{\omega^0}(A_\rho, B_\rho, C_\rho)$, *and* $\mathcal{L}_{\mathrm{coll}}(\rho) := \mathbb{E}[d(F_\rho, G_\rho)^2]$. *Then*

$$\mathcal{L}_{\mathrm{coll}}(\rho) = \frac{\rho^6 g_{\alpha,\beta,\gamma}^2}{9}\mathbb{E}\Big[\|R_p(x, y)(\alpha x - (1-\beta)y)\|^2\Big] + \mathcal{O}(\rho^7). \tag{59}$$

Proof. By theorem 3.1, the logarithmic difference equals $\rho^3 V(a, b, c) + \mathcal{O}(\rho^4)$ uniformly on the support, where $V = (g_{\alpha,\beta,\gamma}/3)R_p(x, y)(\alpha x - (1-\beta)y)$. Both outputs have normal coordinates of order ρ, while their coordinate separation is of order ρ^3. By lemma A.3, $\mathrm{d}^2 = \|u_\rho - v_\rho\|^2\,(1 + \mathcal{O}(\rho^2))$ uniformly on the support; hence the metric correction is $\mathcal{O}(\rho^8)$. Squaring the uniform vector expansion gives $\rho^6\,\|V\|^2 + \mathcal{O}(\rho^7)$. Compact support permits the uniform remainder to be integrated, proving (59). More general distributions require corresponding moment and uniform-integrability hypotheses; pointwise $\mathcal{O}$ notation alone cannot be passed through an expectation. □

The expansion in proposition C.11 is a $\Theta(\rho^6)$ law only when its expected squared cubic coefficient is strictly positive. The identical observation applies to the tree regularizer with $\mathfrak{D}_{\mathcal{T}}^p$ in place of the source coefficient. Compactly supported collinear data can make both losses identically zero; no nondegeneracy follows from nonzero ambient curvature alone.

Finally, the error order associated with a wrong surrogate weight vector is a worst-case statement about maps, not a claim about every individual sample. For a tangent triple v, write $x_i^\rho(v) = \mathrm{Exp}_p(\rho v_i)$ and define $\mathcal{E}_{\widehat{\omega}}(\rho) := \sup_{\max_i \|v_i\| \le 1} \mathrm{d}\Big(F_{\alpha,\beta,\gamma}(x_A^\rho(v), x_B^\rho(v), x_C^\rho(v)),\ \mathcal{B}_{\widehat{\omega}}(x_A^\rho(v), x_B^\rho(v), x_C^\rho(v))\Big)$.

Proposition C.12 (Worst-case parameter identifiability)**.** *Let $\widehat{\omega} \in \Delta_2$ and $\delta_i = \omega_i^0 - \widehat{\omega}_i$. If $\widehat{\omega} \neq \omega^0$, then*

$$\lim_{\rho \downarrow 0} \frac{\mathcal{E}_{\widehat{\omega}}(\rho)}{\rho} = \sum_{i \in \{A,B,C\}} |\delta_i| > 0. \tag{60}$$

If $\widehat{\omega} = \omega^0$, then $\mathcal{E}_{\omega^0}(\rho) = \mathcal{O}(\rho^3)$; at a nonflat point, $\mathcal{E}_{\omega^0}(\rho) = \Theta(\rho^3)$.

Proof. The differentials at the diagonal are the Euclidean path-weight maps. Uniformly for $\max_i \|v_i\| \le 1$, one has $\mathrm{Log}_p F_{\alpha,\beta,\gamma} - \mathrm{Log}_p \mathcal{B}_{\widehat{\omega}} = \rho \sum_i \delta_i v_i + \mathcal{O}(\rho^3)$ because both maps have no quadratic term. The operator norm of $(v_A, v_B, v_C) \mapsto \sum_i \delta_i v_i$ from the product max norm to $T_p\mathcal{M}$ is $\sum_i |\delta_i|$: the triangle inequality gives the upper bound, and equality is attained by taking $v_i = \mathrm{sgn}(\delta_i) e$ for a fixed unit vector e, with the convention $\mathrm{sgn}(0) = 0$. The comparison in lemma A.3 is uniform on the compact unit tangent set; taking suprema and then the limit proves (60).

For $\widehat{\omega} = \omega^0$, (5) gives the uniform upper bound $\mathcal{O}(\rho^3)$. If $R_p \neq 0$, choose an orthonormal plane with nonzero sectional curvature and take $(v_A, v_B, v_C) = (X, 0, Y)$. By lemma B.2, the cubic coefficient is nonzero, furnishing a matching lower bound. Particular samples may still have zero error, for example when all tangent directions are collinear; the proposition concerns the local worst-case map norm. □

D Curvature reconstruction, sensitivity, and model geometries

D.1 Vector and sectional curvature from an equal-layer defect

For $X, Y \in T_p\mathcal{M}$, define

$$F_\rho(X,Y) := F_{\mathrm{eq}}(\mathrm{Exp}_p(\rho X), p, \mathrm{Exp}_p(\rho Y)),$$
$$G_\rho(X,Y) := \mathcal{B}_{(1/4,1/2,1/4)}(\mathrm{Exp}_p(\rho X), p, \mathrm{Exp}_p(\rho Y)).$$

For $\rho \neq 0$, set

$$\widehat{\Delta}_{p,\rho}(X,Y) := \frac{96}{\rho^3}\big(\mathrm{Log}_p G_\rho(X,Y) - \mathrm{Log}_p F_\rho(X,Y)\big). \tag{61}$$

Theorem D.1 (Vector-valued curvature recovery)**.** *For every $X, Y \in T_p\mathcal{M}$, $\Delta_p(X,Y) := \lim_{\rho \to 0} \widehat{\Delta}_{p,\rho}(X,Y) = R_p(X,Y)(Y-X)$. In general, $\widehat{\Delta}_{p,\rho} = \Delta_p + \mathcal{O}(\rho)$ uniformly on compact tangent sets. If the local geodesic symmetry at p is an isometry, then $\widehat{\Delta}_{p,\rho}$ is an even function of ρ and $\widehat{\Delta}_{p,\rho} = \Delta_p + \mathcal{O}(\rho^2)$.*

Proof. Equation (35) with the sign reversed gives $\mathrm{Log}_p G_\rho - \mathrm{Log}_p F_\rho = \rho^3 R_p(X,Y)(Y-X)/96 + \mathcal{O}(\rho^4)$, proving the limit and the general rate. Under the local symmetry hypothesis, remark A.6 shows that each output logarithm is odd in ρ. Their difference is odd, and division by ρ^3 produces an even smooth function. Its first correction to the constant term is therefore of order ρ^2. □

Let $S_p(X,Y) = R_p(X,Y,Y,X)$ be the sectional numerator. No normalization is required. By lemma B.3,

$$S_p(X,Y) = \frac{1}{2}\langle \Delta_p(X,Y), X+Y\rangle. \tag{62}$$

For linearly independent X, Y, the sectional curvature is consequently $K_p(X,Y) = \frac{\langle \Delta_p(X,Y), X+Y\rangle}{2(\|X\|^2\|Y\|^2 - \langle X,Y\rangle^2)}$. Thus one two-direction experiment recovers the sectional curvature of its plane to leading order.

D.2 Reconstruction of the full curvature tensor

The biquadratic map S_p determines the complete algebraic curvature tensor. For $a, b, c, d \in T_p\mathcal{M}$, define

$$P_p(a, c; b, d) := \frac{1}{16}\big[S_p(a + c, b + d) - S_p(a + c, b - d) - S_p(a - c, b + d) + S_p(a - c, b - d)\big]. \tag{63}$$

Lemma D.2 (Double polarization)**.** *The quantity in* (63) *satisfies*

$$P_p(a, c; b, d) = \frac{1}{2}\big(R_p(a, b, d, c) + R_p(a, d, b, c)\big). \tag{64}$$

Proof. First polarize the first variable of $S(x, y) = R(x, y, y, x)$. The curvature symmetries give $S(a + c, y) - S(a - c, y) = 4R(a, y, y, c)$. Indeed, the two cross terms are $R(a, y, y, c)$ and $R(c, y, y, a)$, and pair symmetry followed by the two skew symmetries identifies the latter with the former. Next polarize the quadratic map $y \mapsto R(a, y, y, c)$: $R(a, b+d, b+d, c)-R(a, b-d, b-d, c) = 2R(a, b, d, c) + 2R(a, d, b, c)$. Combining the two polarizations produces a factor 8 multiplying the sum on the right. Division by the normalizing factor 16 in (63) proves (64). □

Theorem D.3 (Full Riemann tensor from collapse defects)**.** *For all $a, b, c, d \in T_p\mathcal{M}$,*

$$R_p(a, b, c, d) = \frac{2}{3}\big(P_p(a, d; b, c) - P_p(a, c; b, d)\big). \tag{65}$$

Since P_p is obtained from S_p and S_p from (62)*, the map $(X, Y) \mapsto \Delta_p(X, Y)$ determines R_p completely.*

Proof. Set $A = R(a, b, c, d)$, $B = R(a, c, b, d)$, and $C = R(a, d, b, c)$. The first Bianchi identity, pair symmetry, and skew-symmetry give $A + C - B = 0$, hence $B = A + C$. By lemma D.2,

$$P(a, d; b, c) = \frac{1}{2}\{R(a, b, c, d) + R(a, c, b, d)\} = \frac{1}{2}(A + B) = A + \frac{C}{2},$$
$$P(a, c; b, d) = \frac{1}{2}\{R(a, b, d, c) + R(a, d, b, c)\} = \frac{1}{2}(-A + C) = \frac{C}{2} - \frac{A}{2}.$$

Their difference is $3A/2$, which is equivalent to (65). □

If $e_1, \ldots, e_d$ is an orthonormal basis, evaluating $S_p(e_i \pm e_k, e_j \pm e_\ell)$ supplies all values required by (63); (65) then recovers every component $R_p(e_i, e_j, e_k, e_\ell)$. The standard curvature symmetries reduce the number of distinct probes.

D.3 Finite-radius bias correction

For a general smooth geometry, the numerator in (61) vanishes to third order at $\rho = 0$; smooth divisibility by ρ^3 therefore extends $\widehat{\Delta}_{p,\rho}$ smoothly through $\rho = 0$. Write its two-sided expansion as $\widehat{\Delta}_{p,\rho} = \Delta_p + c_1\rho + c_2\rho^2 + \mathcal{O}(\rho^3)$. Richardson extrapolation removes the linear term: $\widehat{\Delta}^{\mathrm{R}}_{p,\rho} := 2\widehat{\Delta}_{p,\rho/2} - \widehat{\Delta}_{p,\rho} = \Delta_p + \mathcal{O}(\rho^2)$. Alternatively, the symmetric average

$$\widehat{\Delta}^{\mathrm{sym}}_{p,\rho} := \frac{1}{2}\big(\widehat{\Delta}_{p,\rho} + \widehat{\Delta}_{p,-\rho}\big) = \Delta_p + \mathcal{O}(\rho^2) \tag{66}$$

removes all odd powers whenever both signs can be sampled. On a locally symmetric space, theorem D.1 shows that the deterministic estimator is already even; in that case (66) does not improve the analytic bias, although it may reduce antisymmetric numerical error.

D.4 Optimization of the guaranteed sectional-curvature gain

For an orthonormal plane with $K_p(X,Y) \neq 0$, lemma B.2 separates a guaranteed sectional-curvature factor from the architecture. Define $\mathscr{G}(\alpha,\beta,\gamma) := \gamma(1-\gamma)(1-\alpha)\beta\sqrt{\alpha^2+(1-\beta)^2}$. Then, for the forced one-layer weights, $\lim_{\rho\downarrow 0} \frac{d_p^{\alpha,\beta,\gamma}(\rho;X,Y)}{|K_p(X,Y)|\rho^3} = \frac{g_{\alpha,\beta,\gamma}}{3|K_p(X,Y)|}\Xi_p(X,Y;\alpha,\beta) \geq \frac{1}{3}\mathscr{G}(\alpha,\beta,\gamma)$. On a space form, the final inequality is an equality. The function $\mathscr{G}$ extends continuously to $[0,1]^3$; boundary points represent degenerate binary branches.

Theorem D.4 (Architecture-gain optimization)**.** *On the closed cube* $[0,1]^3$*,*

$$\max \mathscr{G} = \frac{1}{16}. \tag{67}$$

The maximum is attained precisely at $\gamma = 1/2$ *and either* $(\alpha,\beta) = (0,1/2)$ *or* $(\alpha,\beta) = (1/2,1)$*. On the open cube,* $1/16$ *is the supremum and is not attained. The equal architecture has* $\mathscr{G}(1/2,1/2,1/2) = 1/(16\sqrt{2})$*, namely* $1/\sqrt{2}$ *of the global supremum.*

Proof. The factor $\gamma(1-\gamma)$ is at most $1/4$, with equality only at $\gamma = 1/2$. Put $x = \alpha$ and $y = 1-\beta$. It remains to maximize $h(x,y) = (1-x)(1-y)\sqrt{x^2+y^2}$ on $[0,1]^2$.

On the boundary, $h(0,y) = y(1-y) \leq 1/4$, $h(x,0) = x(1-x) \leq 1/4$, and $h(1,y) = h(x,1) = 0$. At an interior critical point, logarithmic differentiation gives $-1/(1-x) + x/(x^2+y^2) = 0$ and $-1/(1-y) + y/(x^2+y^2) = 0$. Equivalently, $x - 2x^2 - y^2 = 0$ and $y - 2y^2 - x^2 = 0$. Subtracting yields $(x-y)(1-x-y) = 0$. The case $x+y = 1$ would imply $3x(1-x) = 1$, which has no real solution. Thus $x = y$, and the critical equations give $x = y = 1/3$. At this point $h = 4\sqrt{2}/27 < 1/4$. Therefore the boundary values $h(0,1/2) = h(1/2,0) = 1/4$ are the global maxima. Multiplying by the maximal γ factor gives (67) and the stated maximizers. Direct substitution gives the equal-architecture value. □

The maximizing points lie on the boundary because one nominal branch becomes degenerate. For the binary source architecture, if every node weight is constrained to be at least η_0, then necessarily $0 < \eta_0 \leq 1/2$ and $(\alpha,\beta,\gamma) \in [\eta_0, 1-\eta_0]^3$. Continuity gives an attained constrained optimum on this nonempty compact cube; the same two-variable function determines it.

D.5 Canonical geometries

Spherical and hyperbolic space forms. On a space form of curvature κ, the leading source-family vector on an orthonormal plane is $-\kappa((1-\beta)X + \alpha Y)$. Positive and negative curvature reverse its direction but have the same magnitude at equal $|\kappa|$. Equation (38) therefore applies equally to spherical and hyperbolic models.

The rotation group. Equip $\mathrm{SO}(3)$ with an Ad-invariant metric. For left-invariant vectors,

$$\nabla_X Y = \frac{1}{2}[X,Y], \qquad R(X,Y)Z = -\frac{1}{4}[[X,Y],Z], \qquad K(X,Y) = \frac{1}{4}\,\|[X,Y]\|^2$$

for orthonormal X, Y (Milnor, 1976). In $\mathfrak{so}(3)$, the bracket is a metric-dependent constant multiple of the cross product, so $\|[X,Y]\|$ is independent of the orthonormal plane. Thus the standard bi-invariant metric is locally a positive-curvature space form, up to normalization, and corollary B.9 explains the fixed-weight depth gap analytically.

Symmetric positive-definite matrices. Let $\mathcal{P}_n$ be the manifold of real symmetric positive-definite matrices with the affine-invariant metric $\langle X,Y\rangle_P = \mathrm{tr}(P^{-1}XP^{-1}Y)$. At the identity, tangent vectors are symmetric matrices and $R_I(X,Y)Z = -\frac{1}{4}[[X,Y],Z]$, $K_I(X,Y) = -\frac{1}{4}\frac{\|[X,Y]\|_F^2}{\|X\|_F^2\|Y\|_F^2 - \langle X,Y\rangle_F^2}$ for linearly independent X, Y (Moakher, 2005; Bhatia, 2007; Absil et al., 2008). Linearly independent commuting tangent matrices span a flat plane, whereas a noncommuting pair has negative sectional curvature. The equal-architecture defect at I is $\mathrm{Log}_I F_{\mathrm{eq}} - \mathrm{Log}_I \mathcal{B}_{(1/4,1/2,1/4)} = \frac{\rho^3}{384}[[X,Y],Y-X] + \mathcal{O}(\rho^4)$. Thus a commuting data family may exhibit exact or higher-order restricted collapse without implying collapse on an open subset of $\mathcal{P}_n$.

Products and flat data restrictions. For a product manifold $\mathcal{M}_1 \times \cdots \times \mathcal{M}_s$ with product metric, weighted Fréchet means, curvature, and the architecture defect act componentwise. If one factor has nonzero curvature at a point in the relevant projection, then the source family cannot collapse on any ambient neighborhood of the corresponding product point. By contrast, let $S \subset \mathcal{M}$ be a totally geodesic flat submanifold and choose a sufficiently small geodesically convex patch $W \subset S$ contained in the ambient local mean domain. Shrinking W if necessary, the unique ambient minimizing geodesic between any two points of W is the corresponding intrinsic geodesic in S. Consequently $\operatorname{Log}_z^{\mathcal{M}} x = \operatorname{Log}_z^{S} x \in T_z S$ for $x, z \in W$. If m_S is the intrinsic weighted mean of points $x_i \in W$, then $\sum_i w_i \operatorname{Log}_{m_S}^{\mathcal{M}} x_i = \sum_i w_i \operatorname{Log}_{m_S}^{S} x_i = 0$. The uniqueness of the ambient Karcher zero therefore gives $m_S = \mathcal{B}_w^{\mathcal{M}}(x_1, \ldots, x_m)$. Since the intrinsic mean is affine in a local flat development of S, induction over the nodes shows that every network restricted to W is affine and collapses to its effective one-layer mean. This data-restricted statement is strictly weaker than functional collapse on an open ambient neighborhood.

E Numerical verification and concluding discussion

E.1 Artifact boundary and numerical protocol

The analytic results above do not depend on computation. The supplement includes the canonical append-only bundles `e1-full-v1`, `e2-full-v1`, `e3-full-v1.1`, `e4-full-v1`, and `bci-iv2a-v1.1` for the source-family, quantifier-comparison, random-tree, high-precision-control, and downstream experiments. Each bundle contains its protocol snapshot, machine-readable results, reports, manifest, and SHA-256 checksum list. We verified every listed checksum before preparing this version. No training or data-generation experiment was rerun; the reported summaries, figures, and BCI paired statistics were recomputed or cross-checked from those frozen records.

E1 contains the full raw table, metadata, targeted controls, and solver-floor audit. E2 contains the deterministic scaling table and all 90 run-level JSON files. E3 contains its raw tree/probe tables, and E4 contains the frozen cyclic core together with the preassigned cubic and flat controls. BCI v1.1 contains the complete v1 summary, five additional stochastic-seed files, the validated subject-first aggregate, the nine subject-level records, and the implementation/test snapshot. The immutable ancestor bundles and `mechanism-regression-v1` remain available for provenance but are not substituted for these canonical paper bundles. This material supports the stated numerical checks; reproducing BCI training from scratch additionally requires lawful access to the source dataset and a compatible software and hardware environment.

For a lightweight check independent of the large bundles, run the following command from the manuscript source directory:

```
python supplementary/minimal_sphere_probe.py
```

The script requires Python 3.10 or newer and NumPy 1.25 or newer. It evaluates the equal-weight sphere probe directly and prints two cubic components converging to $-1/96$; the expected final line begins `PASS`. This check supports the transparent sphere witness only and does not replace the frozen experiments.

The E1 and E3 scans use exact exponential and logarithm maps on two-dimensional space forms and an intrinsic Karcher iteration with backtracking accepted when either the energy or the residual decreases. Every reported solve stores the norm of the Karcher residual at the one-layer mean and the largest residual over the tree. The solver tolerances are 10^{-13} for E1 and 10^{-14} for E3; E4 instead uses the symmetry-reduced high-precision equation described below at both 80 and 120 decimal digits. E2 uses the same exact space-form maps in a separate teacher–student protocol and records optimization, simplex-feasibility, and Karcher-residual diagnostics. Flat-space evaluations are preassigned controls. The E3 probes and E1 standard/targeted controls are curvature-paired; the E1 full-scan random probes are independently drawn within each curvature block. None of these assignments is a post hoc filter.

E.2 Sphere formulas and numerical protocol

For the representative high-precision calculations, we use the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$, whose sectional curvature is one. For $v \in T_x\mathbb{S}^2$, let $r = \|v\|$ and define $\operatorname{sinc}(r) = \sin(r)/r$ for $r \neq 0$ and $\operatorname{sinc}(0) = 1$. Then $\operatorname{Exp}_x(v) = \cos(r)x + \operatorname{sinc}(r)v$. For $y \notin \{x, -x\}$, with $\theta = \arccos\langle x, y\rangle$, $\operatorname{Log}_x(y) = \frac{\theta}{\sin\theta}(y - \cos\theta\, x)$, $\operatorname{Log}_x x = 0$. The separate definition at $y = x$ removes the apparent $0/0$ singularity; the antipode is excluded because the logarithm is not single-valued there.

We take $p = (0, 0, 1)$, $X = (1, 0, 0)$, and $Y = (0, 1, 0)$. All test inputs lie in $\overline{B(p, 0.2)}$. Set $U = B(p, \pi/4)$ and choose radii $0.2 < r < R/2$ with $R < \pi/4$. To verify definition 2.1, fix an arbitrary finite configuration in U and choose $\bar{r} < \pi/4$ containing that configuration in $B(p, \bar{r})$. For every data point $x \in B(p, \bar{r})$ and every $z \in U$, one has $r_x := \mathrm{d}(x, z) < \bar{r} + \pi/4 < \pi/2$. If e_x is the unit radial direction from x at z, the spherical Hessian formula is $\operatorname{Hess}_z\big(\frac{1}{2}\,\mathrm{d}(\,\cdot\,, x)^2\big)\,(\xi, \xi) = \langle \xi, e_x\rangle^2 + r_x \cot(r_x)\big(\|\xi\|^2 - \langle \xi, e_x\rangle^2\big)$, with its continuous value $\|\xi\|^2$ at $z = x$. It is positive for every nonzero ξ because $0 \le r_x < \pi/2$. Hence every weighted energy is strictly geodesically convex on U. The local center-of-mass theorem supplies a minimizer in $B(p, \bar{r}) \subset U$, and strict convexity makes it the unique Karcher zero in U. Thus U is a local mean domain. Moreover, lemma B.7 identifies all means used in the experiment with their unique global counterparts, while strong convexity keeps every intermediate binary mean in $B(p, r)$ (Afsari, 2011).

Binary means are evaluated by the exact formula $x \#_t y = \operatorname{Exp}_x(t \operatorname{Log}_x y)$. For a three-point mean, write $m(u) = \operatorname{Exp}_p(u_1 X + u_2 Y)$ and define the intrinsic and pulled-back residuals by $\mathcal{R}(u) := \sum_i w_i \operatorname{Log}_{m(u)} x_i \in T_{m(u)}\mathbb{S}^2$, $\widetilde{\mathcal{R}}(u) := \big(D \operatorname{Exp}_p|_u\big)^* \mathcal{R}(u) \in T_p\mathbb{S}^2$. Because $D \operatorname{Exp}_p|_u$ is invertible in the working ball, $\mathcal{R}(u) = 0$ if and only if the X- and Y-components of $\widetilde{\mathcal{R}}(u)$ vanish. Newton's method was applied to this two-dimensional pulled-back equation, initialized at $u^{(0)} = \sum_i w_i \operatorname{Log}_p x_i$, using 80 decimal digits, and terminated when $\|\mathcal{R}(u)\| < 10^{-60}$. For the equal architecture, the coordinate-swap isometry interchanges A_ε and C_ε, while both the equal two-layer block and the equal one-layer mean are invariant under this simultaneous leaf exchange. Uniqueness therefore forces each output to lie on the fixed geodesic $u_1 = u_2$, and the equation was solved on that one-dimensional set; the two reported tangent components are consequently equal exactly. Repeating the calculations with 120 digits changes none of the displayed figures. The residual is a stationarity diagnostic, not an interval-arithmetic enclosure of the solution error; analytic uniqueness, rather than the residual alone, identifies the relevant root.

Under the curvature convention (10), $R_p(X, Y)Z = \langle Y, Z\rangle\, X - \langle X, Z\rangle\, Y$. Hence $R_p(X, Y)(Y - X) = X + Y$.

E.3 Representative high-precision source-family probes

E.3.1 Equal architecture

Define $D_\varepsilon^{\mathrm{eq}} := \varepsilon^{-3}(\operatorname{Log}_p F_{\mathrm{eq}} - \operatorname{Log}_p \mathcal{B}_{(1/4,1/2,1/4)})$, with both maps evaluated at $(\operatorname{Exp}_p(\varepsilon X), p, \operatorname{Exp}_p(\varepsilon Y))$. The predicted limit is $D_0^{\mathrm{eq}} = -(X + Y)/96$. Its X- and Y-components are both -0.0104166666667.

Table 3: Equal-architecture cubic coefficient on $\mathbb{S}^2$. Coordinate-interchange symmetry gives $\langle D_\varepsilon^{\mathrm{eq}}, X\rangle = \langle D_\varepsilon^{\mathrm{eq}}, Y\rangle$ exactly. The error is $\|D_\varepsilon^{\mathrm{eq}} - D_0^{\mathrm{eq}}\|$.

ε	common tangent component	error
0.20000	$-1.046167135771 \times 10^{-2}$	$6.364624444 \times 10^{-5}$
0.10000	$-1.042794313057 \times 10^{-2}$	$1.594732819 \times 10^{-5}$
0.05000	$-1.041948733308 \times 10^{-2}$	$3.989024702 \times 10^{-6}$
0.02500	$-1.041737192970 \times 10^{-2}$	$9.973925513 \times 10^{-7}$
0.01250	$-1.041684298845 \times 10^{-2}$	$2.493566509 \times 10^{-7}$

Halving ε reduces the error by a factor tending to four. This is the $\mathcal{O}(\varepsilon^2)$ scaled bias predicted by the exact parity argument in remark A.6 and theorem D.1.

E.3.2 A NONSYMMETRIC INTERIOR ARCHITECTURE

Take $(\alpha, \beta, \gamma) = (0.3, 0.7, 0.4)$. The effective weights are $\omega^0 = (0.42, 0.30, 0.28)$: the three entries are $0.6 \cdot 0.7$, $0.6 \cdot 0.3 + 0.4 \cdot 0.3$, and $0.4 \cdot 0.7$. Define $D_\varepsilon^{\mathrm{gen}} := \varepsilon^{-3}(\mathrm{Log}_p F_{0.3,0.7,0.4} - \mathrm{Log}_p \mathcal{B}_{(0.42,0.30,0.28)})$, again evaluating both maps at $(\mathrm{Exp}_p(\varepsilon X), p, \mathrm{Exp}_p(\varepsilon Y))$. Here $g_{\alpha,\beta,\gamma}/3 = 0.0392$, while $R_p(X, Y)(0.3X - 0.3Y) = -0.3(X + Y)$. The predicted limit is therefore $D_0^{\mathrm{gen}} = -0.01176(X + Y)$.

Table 4: General-weight cubic coefficient on $\mathbb{S}^2$ for $(\alpha, \beta, \gamma) = (0.3, 0.7, 0.4)$ and effective weights $(0.42, 0.30, 0.28)$. The error is $\|D_\varepsilon^{\mathrm{gen}} - D_0^{\mathrm{gen}}\|$.

ε	$\langle D_\varepsilon^{\mathrm{gen}}, X\rangle$	$\langle D_\varepsilon^{\mathrm{gen}}, Y\rangle$	error
0.20000	$-1.183054931337 \times 10^{-2}$	$-1.181984850152 \times 10^{-2}$	$9.251512714 \times 10^{-5}$
0.10000	$-1.177771503621 \times 10^{-2}$	$-1.177507058812 \times 10^{-2}$	$2.325822724 \times 10^{-5}$
0.05000	$-1.176443351260 \times 10^{-2}$	$-1.176377431511 \times 10^{-2}$	$5.822498482 \times 10^{-6}$
0.02500	$-1.176110867364 \times 10^{-2}$	$-1.176094399381 \times 10^{-2}$	$1.456118663 \times 10^{-6}$
0.01250	$-1.176027718685 \times 10^{-2}$	$-1.176023602436 \times 10^{-2}$	$3.640605075 \times 10^{-7}$

Both tangent components converge to -0.01176, and the errors again decrease quadratically after cubic scaling. No curvature tensor is supplied to the numerical mean solver; curvature enters only in the analytic prediction against which the two independently computed intrinsic outputs are compared.

E.4 FROZEN SOURCE-FAMILY MECHANISM CHECKS

Tables 3 and 4 isolate two transparent sphere witnesses, while the canonical E1 full scan and the independently frozen numerical-core regression corroborate their asymptotic coefficients at float64 accuracy. The complete E1 table broadens this check to space-form curvatures $\kappa \in \{-4, -1, -1/4, 0, 1/4, 1, 4\}$, six radii from 0.0125 to 0.3, the equal and nonsymmetric architectures, 32 additional random interior architectures, and one standard plus 256 random probes per architecture. The resulting 366,996 full-scan rows are accompanied by 186 separately targeted curvature-linearity and angle-degeneration rows. There are no solver failures. The preassigned standard-witness log–log gap slopes lie in the enclosing interval $[2.997181, 3.002752]$. Across the predeclared nondegenerate small-radius rows, the coefficient relative-error median, 95th percentile, and 99th percentile are 2.398×10^{-4}, 4.269×10^{-3}, and 7.528×10^{-3}, respectively.

The signed projected coefficient reverses with κ. Across the equal and nonsymmetric targeted fits, its measured slope as a function of curvature has relative error at most 2.424×10^{-6} and $R^2 \geq 0.999998758$. There are no standard-probe sign mismatches, the Euclidean raw gap is at most 1.145×10^{-16}, and the maximum full-scan solver residual is 9.9993×10^{-14}. In the near-collinear scan, the zero-angle gap is at most 1.112×10^{-16}, while the measured small-angle orders agree with the architecture-dependent directional predictions. This last control is important: it corroborates the degeneration discussed after (58), rather than suggesting a nonexistent curvature-only lower bound.

Among the 52,360 eligible nondegenerate, nonzero-curvature rows at the finest radius, four have gap-to-residual ratio below the predeclared factor of 100 and are retained as solver-limited observations. The audit identifies only one corresponding coefficient/slope outlier. A frozen addendum reruns all four groups at tolerances 10^{-13}, 10^{-14}, and 10^{-15}. For the outlying group, tightening the tolerance from 10^{-13} to 10^{-14} changes the fitted order from 3.610417 to 3.001128 and the finest-radius coefficient error from 0.756718 to 2.988×10^{-5}; the 10^{-14} and 10^{-15} results agree at the stored precision. We therefore retain and mark the original row but treat its anomaly as a solver-floor diagnostic, not as geometric evidence.

At $\rho = 0.0125$, all 52,360 nonzero-curvature rows with theoretical coefficient norm at least 10^{-8} enter the measured-versus-predicted panel in Figure 3. Of these, 52,356 pass the factor-100 solver-safety gate. Their log–log coefficient regression has slope 1.0000056 and $R^2 = 0.9999999928$; the four solver-limited rows are shown separately rather than silently removed.

Table 5 collects the accepted scope of all four experiment lines before the E2–E4 details below.

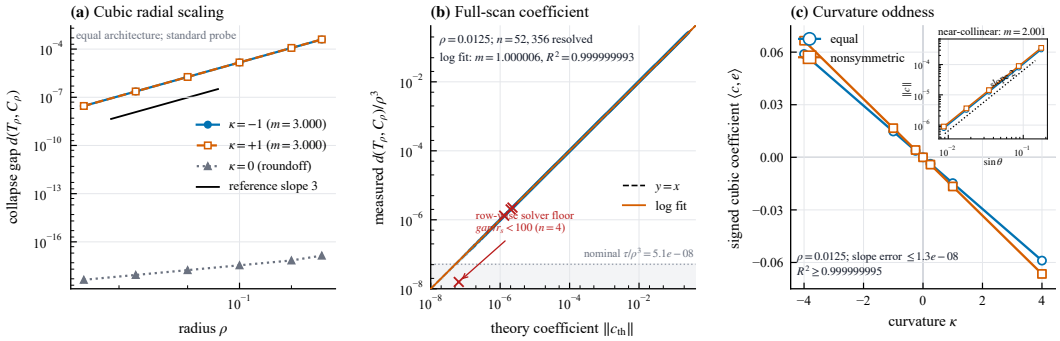


Figure 3: Canonical E1 verification of the source-family cubic mechanism from `e1-full-v1`. Left: the standard equal-architecture witness across $\kappa = -1, 0, 1$, with a reference slope of three and the flat signal marked as roundoff. Center: all eligible rows from the byte-preserved full table. The 52,356 solver-resolved rows determine the fit and four solver-floor rows are marked by crosses. Right: the targeted coefficient is linear in curvature for the equal and nonsymmetric architectures; the inset shows the predicted near-collinear quadratic degeneration.

Table 5: Canonical theorem-facing checks. Summary values and evaluated scopes are recomputed from the checksum-pinned frozen bundles included with the supplement.

Check	Evaluated scope	Primary outcome
E1	366,996 + 186 rows; seven curvatures; 34 architectures	Standard-witness gap orders in $[2.997181, 3.002752]$; coefficient-error median/p95 $2.398 \times 10^{-4}/4.269 \times 10^{-3}$; curvature-fit relative error at most 2.424×10^{-6}
E2	Two structures; three curvatures; six radii; 90 configurations, five seeds per stratum	Curved forced-distance orders in $[2.999373, 3.000624]$, squared-loss orders in $[5.998133, 6.001872]$, and wrong-weight orders equal to 1; all E2b acceptance gates pass
E3	1,792 trees; 59,136 tree–probe pairs; 532,224 measurements	All 38 strata pass; vector-regression slope in $[0.999950, 1.000050]$ with $R^2 \geq 0.9999999988$; coefficient-bound ratio at most 0.115784
E4	Cyclic core plus ordinary cubic and Euclidean flat controls; six radii at 80 and 120 digits	Cyclic order 4.999965 with coefficient error 1.196×10^{-6}; ordinary-control order 3.000123; flat gap/floor ≤ 0.0884

E.5 FROZEN FIXED-VERSUS-INPUT-DEPENDENT QUANTIFIER CHECK

E2 separates deterministic order checks from learning and pointwise fitting. E2a evaluates the equal and nonsymmetric source architectures at six radii for $\kappa \in \{-1, 0, 1\}$. On the four curved architecture–curvature strata, the forced path-product candidate has witness-distance slopes in $[2.999373, 3.000624]$ and random squared-loss slopes in $[5.998133, 6.001872]$; the deliberately wrong fixed weights have axial-distance slope one in every stratum. In flat space the forced signal is at numerical zero, so no spurious log–log order is assigned. These are the distance-level and risk-level counterparts of the cubic obstruction and the first-order identifiability statement.

E2b contains 90 unique configurations: two architectures, three curvatures, three radii, and five seeds, with $5{,}000/1{,}000/5{,}000$ train/validation/test triples per run. All 90 learned-fixed runs pass the predeclared check against the forced-fixed feasible point, all adaptive runs finish without a nonfinite event, and every one-triangle and per-input stationarity solve is simplex-feasible without optimization fallback. Across the 60 curved runs, the globally learned fixed weights remain within 2.154×10^{-4} in ℓ_∞ distance of the path-product weights and attain 0.676–0.802 times the forced-fixed held-out mean squared error. This changes the distributional constant but not the exact functional statement: the curved held-out risk remains nonzero.

The quantifier controls behave differently. One-triangle fitting reaches fitted-input mean squared error at most 1.333×10^{-15}, yet on new curved triangles its run-level test-error ratio is 0.993–5.719 relative to forced fixed (median 1.865 over 60 runs). Panel (c) instead plots 12 architecture–curvature–radius ratios after averaging five seeds within each stratum; those ratios range from 2.894 to 2.939, with median 2.917. The per-input oracle is simplex-feasible on every test triple, uses no fallback, and has maximum across runs of the held-out mean squared error 3.331×10^{-16}; its maximum stationarity and reconstructed Karcher residuals are 5.218×10^{-16} and 4.417×10^{-15}. Thus accurate fitting after observing each input is compatible with the impossibility of one globally fixed exact surrogate. The adaptive network, with 4,803 or 4,995 parameters depending on the source architecture, is reported only as a finite-capacity input-dependent control, not as a capacity-matched fixed-layer baseline.

Figure 4 is generated from the canonical E2a table and all 90 hash-verified E2b result files. It keeps the deterministic orders and risk comparison in separate visual roles; the oracle is deliberately shown on a split scale because it has a different input-wise quantifier. Across the 60 curved runs, the learned fixed-weight deviation has median 4.940×10^{-5} and maximum 2.153×10^{-4}, while the maximum-coordinate input-adaptive weight standard deviation has median 5.646×10^{-5} and maximum 3.609×10^{-4}.

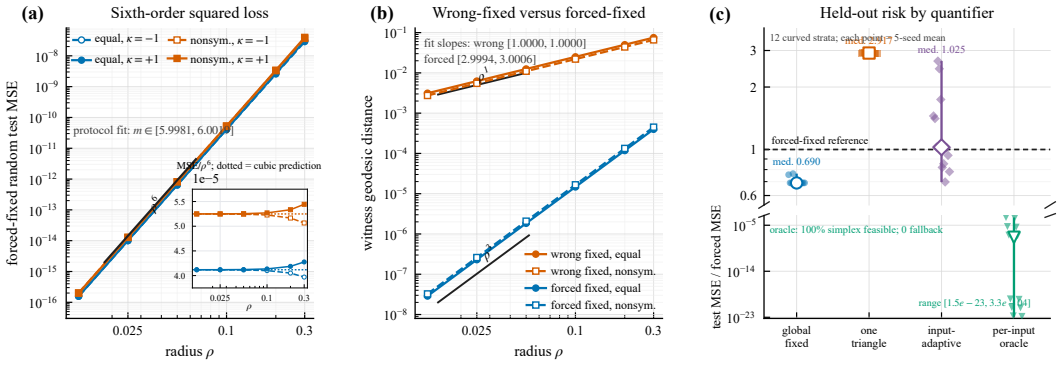


Figure 4: Canonical E2 fixed-versus-input-dependent quantifier check from `e2-full-v1`. (a) Forced-fixed random held-out squared distance follows the predicted ρ^6 law on the four curved strata; the inset shows the MSE/ρ^6 plateau against the parameter-free cubic prediction. (b) The protocol-defined axial wrong-fixed control has slope 1, while the forced path-product witness has slope 3. (c) Each point is a five-seed mean within one of 12 curved structure–curvature–radius strata, normalized by forced-fixed risk; global fixed, one-triangle, and input-adaptive methods occupy the upper panel, while the per-input oracle is shown on the lower split scale. Weight diagnostics for the 60 curved runs are reported in the text.

E.6 CANONICAL RANDOM-TREE VERIFICATION

The E3 protocol samples 256 unique ordered full-binary trees at every structural depth $2, \dots, 8$, for 1,792 trees in total. Depths at least three are stratified by $m \in \{3, 5, 8\}$ external labels; node weights are sampled in $[0.1, 0.9]$. Each tree receives one fixed $X0Y$ probe and 32 random unit tangent tuples. The probe set, tree, labels, and node weights are paired across $\kappa = -1$ and $\kappa = 1$, and each pair is evaluated at four radii. This gives 473,088 curved rows. A separate $\kappa = 0$ exact-collapse control at $\rho = 0.1$ adds 59,136 rows. The independent sampled unit is the tree; the 32 random probes are repeated measurements within that tree.

The generation protocol and original v1 acceptance threshold were fixed before the full scan. After the scan exposed solver-floor failures, v1.1 amended only the derived minimum-radius coefficient-resolution cutoff, using the already fixed tolerance, safety factor, radius, and cubic order rather than fitting the observed failure boundary. The amended cutoff is $\frac{100 \times 10^{-14}}{0.0125^3} = 5.12 \times 10^{-7}$. Of 118,272 finest-radius curved rows, 117,982 exceed this gate and have gap-to-residual ratio at least 359.427. The remaining 290 rows, arising from 145 near-cancellation tree–probe pairs, remain in the raw table and low-signal audit and are classified as numerically unresolved at the minimum radius. Under the amended acceptance protocol they are excluded only from the three finest-radius gates: solver safety, coefficient regression, and coefficient error.

Across the 38 curvature–label–depth strata, regression of measured vector coefficients on the recursion prediction has slope between 0.9999500803 and 1.0000499238, with minimum $R^2 = 0.9999999988$. The largest stratum median and 95th-percentile coefficient relative errors are 5.925×10^{-5} and 1.435×10^{-4}. The ratio $\frac{\|\mathfrak{D}^p_T(v)\|}{(L_T+1)\|R_p\|_{\mathrm{op}} V^3/3}$ is at most 0.1157836, with no violation. This is a check of the cubic *coefficient* bound in theorem C.8; it is not a claim that a finite-radius distance omits the remainder in theorem C.9. All Karcher solves converge, the maximum recorded residual is 9.9996×10^{-15}, paired curvature predictions flip sign exactly, and the largest Euclidean raw gap is 5.5511×10^{-17}.

Figure 5(a,b) shows both the coefficient agreement and the tree-level depth-bound ratios; its third panel supplies the E4 controls described below.

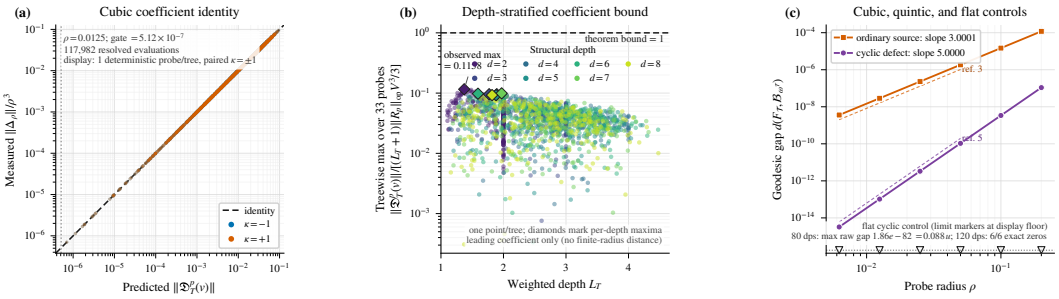


Figure 5: Canonical E3–E4 verification. (a) E3 measured versus recursively predicted coefficient norms at the finest radius for a deterministic tree-balanced display sample (one resolved probe per tree, paired across $\kappa = \pm 1$; 3,584 points); the diagonal is the parameter-free prediction. (b) E3 one point per tree (1,792 total), showing the maximum over its 33 probes of the cubic coefficient divided by the weighted-depth bound; diamonds mark structural-depth maxima and the dashed line is the theorem bound. (c) E4 on the common high-precision probe: the ordinary source tree gives the positive cubic control, the cyclic construction gives the fifth-order signal, and the Euclidean cyclic tree collapses at the declared precision floor. Panel (b) is a coefficient bound, not a finite-radius distance bound.

E.7 HIGH-PRECISION FIFTH-ORDER CHECK

For the cyclic construction H in proposition C.5, symmetry makes the two tangent components equal. Define the measured component coefficient

$$c_\rho := \frac{1}{\rho^5}\Big\langle \operatorname{Log}_p H(\operatorname{Exp}_p(\rho X), p, \operatorname{Exp}_p(\rho Y)) \\ - \operatorname{Log}_p \mathcal{B}_{(1/3,1/3,1/3)}(\operatorname{Exp}_p(\rho X), p, \operatorname{Exp}_p(\rho Y)), X\Big\rangle.$$

The predicted limit is $c_0 = 7/29160 = 2.400548696844993 \times 10^{-4}$. The complete frozen cyclic core grid is shown in Table 6; the displayed 80- and 120-digit evaluations agree exactly. Fitting the four smallest radii gives gap order 4.9999648.

Table 6: Fifth-order component coefficient for the cyclic zero-cubic-defect construction on $\mathbb{S}^2$. The two stored precision settings give identical displayed values.

ρ	c_ρ	$\lvert c_\rho - c_0\rvert/\lvert c_0\rvert$
0.20000	$2.396880252341 \times 10^{-4}$	1.528169×10^{-3}
0.10000	$2.399768611144 \times 10^{-4}$	3.249614×10^{-4}
0.05000	$2.400362160215 \times 10^{-4}$	7.770583×10^{-5}
0.02500	$2.400502591740 \times 10^{-4}$	1.920607×10^{-5}
0.01250	$2.400537203615 \times 10^{-4}$	4.787751×10^{-6}
0.00625	$2.400545825603 \times 10^{-4}$	1.196078×10^{-6}

The reported `e4-full-v1` bundle completes the original Figure 5 design with two preassigned controls evaluated by the same 80- and 120-digit framework. The ordinary equal source tree has a finest-four gap slope of 3.0001226757 and a finest-radius component relative error of 4.232×10^{-6} from the predicted $-1/96$. The Euclidean cyclic control has maximum gap-to-precision-floor ratio 0.08839; all six 120-digit rows are exact zeros, so no spurious log–log slope is fitted. The cyclic core and ordinary control therefore separate fifth- and third-order behavior, while the flat control checks that the plotted zero is numerical collapse rather than an omitted signal.

This verifies only the narrow claim made in proposition C.5: the specified cyclic sphere construction has vanishing cubic defect and a nonzero fifth-order coefficient. It is not promoted to a claim about every architecture whose cubic defect vanishes. The combined third panel in Figure 5 displays these controls together with the canonical E3 panels.

E.8 BCI COMPETITION IV-2A DOWNSTREAM PROTOCOL

The downstream experiment uses all nine official training-to-evaluation session splits of BNCI2014-001 / BCI Competition IV-2a (Tangermann et al., 2012); the evaluation session is scored once after freezing. Excluding the three EOG channels and source-flagged artifact trials leaves $2{,}328/2{,}368$ training/evaluation trials from $2{,}592/2{,}592$. Each continuous run is filtered before epoching (0.5–3.5 seconds after cue onset). Fourth-order zero-phase Butterworth filters and centered OAS covariance produce mu (8–13 Hz), low-beta (13–20 Hz), and high-beta (20–30 Hz) views in $\mathrm{SPD}(22)$, plus an 8–30 Hz broad-band control.

The nested teacher is $(C_1 \#_{0.35} C_2) \#_{0.5} (C_2 \#_{0.65} C_3)$, with path-product weights $(0.325, 0.35, 0.325)$. Learned-fixed and input-adaptive fusion minimize training-session AIRM error to the teacher for 20 and 30 fixed epochs; no evaluation metric is computed during optimization. The common frozen teacher head uses a training-only log-Euclidean reference, affine-log coordinates, standardization, and multinomial logistic regression, so Table 2 isolates replacement fidelity. Refit-head metrics were reported as part of that bundle. Tangent-space controls and FBCSP with shrinkage LDA use the same retained training trials.

The teacher and path-product are parameter-free, and the three task baselines use effectively deterministic solvers. Only learned-fixed and input-adaptive fusion are therefore replicated over six seeds. Seeds are averaged within subject before the nine independent subjects enter the mean/standard

deviation, subject-bootstrap 95% intervals, and paired sign-flip summaries; the 54 runs are not independent units. We recomputed the two-sided exact values using the absolute mean of the nine subject differences and all 2^9 sign assignments, counting ties in the tail. The resulting p values are 0.125, 1.000, and 0.96875 for path-product, learned-fixed, and input-adaptive fusion. A finite-sample valid interpretation additionally requires invariance of the subject-difference vector under independent coordinate sign changes; Appendix F.3 states this condition and its limits. Every Fréchet solve records convergence, residual, and iterations. Hardware-specific fusion, cached-prediction, and amortized preprocessing timings were reported for audit, not as evidence of causal online latency or speedup.

E.9 Scope, limitations, and conclusion

The main classification concerns the fixed-parameter map $F_{\alpha,\beta,\gamma}(A, B, C) = (A\#_\alpha B)\#_\gamma(B\#_\beta C)$ with interior node weights. Its only possible collapsed weights are the path-product weights, and its leading possible local discrepancy is the translation-invariant cubic vector in (5). For this particular family, that cubic vector detects every nonzero sectional plane; hence functional collapse on a local mean domain is equivalent to flatness. The equal case supplies the fixed two-layer parameter choice required by the ManifoldNet conjecture and also rules out fixed-weight collapse on nonzero space forms.

This conclusion is architecture-specific. Proposition 4.1 gives a curved-space network that collapses, whereas theorem 4.2 shows that flatness is equivalent to collapse of every finite fixed-weight Fréchet-mean network. For a specified tree, $\mathfrak{D}^p_{\mathcal{T}} \neq 0$ is a rigorous certificate that the bracketing is visible at cubic order. The converse is false: proposition C.5 has $\mathfrak{D}^p_{\mathcal{T}} \equiv 0$ but a nonzero fifth-order defect on $\mathbb{S}^2$. Thus the cubic architecture defect is the first possible obstruction, not a complete invariant of exact collapse.

The differential expansions and collapse classification are local because global Fréchet means may fail to exist or be unique. The sole global step is lemma B.7: under completeness, it identifies sufficiently concentrated local means with their unique global counterparts and thereby transfers the local witnesses to the original ManifoldNet conjecture. The remaining hypotheses are smooth Riemannian geometry, fixed node weights, exact squared-distance means, and exact geometric maps. Input-dependent attention, extrinsic means, nonsquared losses, branch selection among nonunique global minimizers, and truncated numerical solvers define different maps. Likewise, a finite-radius certificate requires a valid fourth-order constant in theorem C.9; the leading cubic coefficient alone is only asymptotic. E2 is a controlled teacher–student study of these quantifiers. The BCI experiment asks the separate downstream question for one SPD dataset: it supports fidelity of the tested frozen replacements, not improved task accuracy or universal generalization.

Within these hypotheses, the source-family classification is exact, the tree recursion records the complete cubic jet, and the equal-layer defect determines the Riemann tensor. The canonical computations corroborate the source-family ρ^3 law; the penalty for wrong fixed weights and the four surrogate quantifiers; the recursion over $1{,}792$ trees; and the cubic, fifth-order, and flat controls. Separately, BCI representation gaps coexist with nearly unchanged subject-mean teacher-head accuracy, without establishing trial-wise prediction agreement. These experiments clarify theoretical consequences and practical quantifiers but do not replace the proofs.

F Scope of local fusion and statistical interpretation

This appendix makes explicit the distinctions needed to interpret the local classification, the input-conditioned target, and the unchanged experimental summaries. The examples are analytic counterexamples, not additional experiments. They do not change the source-block formula, the tree recursion, or any of the recorded measurements.

F.1 The region of exact equality and the base point of an expansion

Example F.1 (A nonflat manifold with a flat open region)**.** There is a smooth, connected, complete nonflat Riemannian surface containing a nonempty open set on which every finite fixed-weight mean tree fuses exactly. Consequently, nonflatness of the ambient manifold does not exclude exact fusion on every open subset of its input space.

Proof. On $\mathbb{R}^2$ define $f(x) = 0$ for $x \le 0$ and $f(x) = e^{-1/x^2}$ for $x > 0$, and equip the plane with $g = e^{2f(x)}(dx^2 + dy^2)$. Every derivative of the right-hand branch tends to zero at the origin, so f and g are smooth. Since $0 \le f \le 1$, the metric is uniformly equivalent to the complete Euclidean metric and is complete. A direct calculation for this conformal metric gives $K = -e^{-2f} f''$. For $x > 0$, $f''(x) = e^{-1/x^2}(4x^{-6} - 6x^{-4})$, which is nonzero for sufficiently small positive x. Thus the surface is nonflat in every neighborhood of the origin, whereas the half-plane $x < 0$ is exactly Euclidean.

Choose a sufficiently small local mean domain U about the origin and a Euclidean convex disk V whose closure lies in $U \cap \{x < 0\}$. The inequality $g \ge g_{\mathrm{E}}$ implies $d_g(z, x) \ge \|z - x\|_{\mathrm{E}}$ for all points of the surface. For points in V, their Euclidean segment lies in the flat half-plane and attains this lower bound. Given inputs $x_i \in V$ and weights $w \in \Delta_{m-1}$, their Euclidean average $\bar{x} = \sum_i w_i x_i$ belongs to V. For any $z \in \mathbb{R}^2$, $\sum_i w_i d_g(z, x_i)^2 \ge \sum_i w_i \|z - x_i\|_{\mathrm{E}}^2 = \sum_i w_i \|\bar{x} - x_i\|_{\mathrm{E}}^2 + \|z - \bar{x}\|_{\mathrm{E}}^2$. At $z = \bar{x}$ the bound is attained. It is strict for $z \ne \bar{x}$, so $\bar{x}$ is the unique global, and hence local, mean. Induction over a finite tree identifies every intermediate output with its Euclidean average in V. Thus exact path-product fusion holds on V^m, even though $R|_U \not\equiv 0$. There is no contradiction with theorem 3.2, which asks for identity on all of U^3. □

Example F.2 (Nonflat neighborhoods with a zero-curvature base point)**.** On a small coordinate disk about $p = (0, 0)$, let $g = e^{2x^3}(dx^2 + dy^2)$. The same curvature calculation gives $K(x, y) = -6xe^{-2x^3}$. Therefore $R_p = 0$, although every neighborhood of p contains points of nonzero curvature. The cubic coefficient of every fixed tree at p is zero, by theorem C.2; its defect there is $\mathcal{O}(\rho^4)$. Nevertheless, theorem 3.2 excludes exact source-block fusion on any whole local mean domain about p.

The distinction also applies to input directions and distributions. Even when $R_p \ne 0$, collinear probes remain on one short geodesic and have exactly zero path-product defect by lemma A.1. A tangent law supported on these probes has zero expected loss. Thus a nonflat *domain*, a nonzero tensor at the chosen *point*, and a nonzero coefficient on the chosen *probe or distribution* are different hypotheses. The first governs exact regional classification, the second provides a cubic worst-case witness, and the third is needed for a particular $\Theta(\rho^3)$ distance or $\Theta(\rho^6)$ loss law.

F.2 Pointwise representability is a separate feasibility condition

Proposition F.3 (Pointwise Karcher feasibility)**.** *Let $x_1, \dots, x_m, y$ belong to a local mean domain U. There exists $w \in \Delta_{m-1}$ with $y = \mathcal{B}_w(x_1, \dots, x_m)$ if and only if $0 \in \mathrm{conv}\{\mathrm{Log}_y x_1, \dots, \mathrm{Log}_y x_m\} \subset T_y\mathcal{M}$.*

Proof. If y is the mean, its differential vanishes because y is interior to U, and the Karcher equation is $\sum_i w_i \mathrm{Log}_y x_i = 0$. Nonnegativity and unit sum of the weights make this exactly a convex-hull representation of zero. Conversely, such a convex combination supplies weights for which y is a Karcher zero. The defining uniqueness of that zero in U identifies it with $\mathcal{B}_w$. This argument does not infer sufficiency of stationarity on an arbitrary manifold or outside the specified local domain. □

Changing the weights gives additional freedom inside this convex-hull constraint, but does not ensure that a target produced by a nested mean satisfies it. The following obstruction identifies a component of the curvature defect that weight adjustment cannot remove.

Lemma F.4 (A normal component obstructs adaptive weights)**.** *Let $X, Y \in T_p\mathcal{M}$ be linearly independent and put $W = \mathrm{span}\{X, Y\}$. Let $P_{W^\perp}$ be orthogonal projection onto its complement. If $P_{W^\perp} R_p(X, Y)(Y - X) \ne 0$, then for all sufficiently small positive ρ there is no $w \in \Delta_2$ such that $F_{\mathrm{eq}}(\mathrm{Exp}_p(\rho X), p, \mathrm{Exp}_p(\rho Y)) = \mathcal{B}_w(\mathrm{Exp}_p(\rho X), p, \mathrm{Exp}_p(\rho Y))$. The conclusion allows weights chosen independently for each input and radius; no continuity or other regularity of their selection is assumed.*

Proof. Write F_ρ for the nested output, $G_w(\rho)$ for the single mean, and $\omega^0 = (1/4, 1/2, 1/4)$. By theorem A.4, for the whole closed simplex and a common sufficiently small radius,

$$\mathrm{Log}_p G_w(\rho) = \rho(w_A X + w_C Y) + \rho^3 Q_w + \mathcal{O}(\rho^4), \tag{68}$$

where $Q_w = \sum_i w_i R_p(\mu_w, v_i)v_i/3$, $(v_A, v_B, v_C) = (X, 0, Y)$, and $\mu_w = w_A X + w_C Y$. The coefficient Q_w is a polynomial in w, hence bounded and Lipschitz on the closed simplex. The nested output has linear term $\rho(X+Y)/4$ and a fixed bounded cubic term. If $F_\rho = G_w(\rho)$, subtraction and division by ρ give $(w_A - 1/4)X + (w_C - 1/4)Y = \mathcal{O}(\rho^2)$, with a bound independent of w. Linear independence makes the map $(s,t) \mapsto sX + tY$ bounded below on $\mathbb{R}^2$, so $|w_A - 1/4| + |w_C - 1/4| = \mathcal{O}(\rho^2)$. The unit-sum condition then implies $w - \omega^0 = \mathcal{O}(\rho^2)$ and $Q_w - Q_{\omega^0} = \mathcal{O}(\rho^2)$.

Set $D = -R_p(X, Y)(Y - X)/96$. Subtracting Equation (68) from the equal-source expansion gives $\mathrm{Log}_p F_\rho - \mathrm{Log}_p G_w(\rho) = \rho\big((1/4 - w_A)X + (1/4 - w_C)Y\big) + \rho^3 D + \mathcal{O}(\rho^4)$. The omitted weight change in the cubic term is $\mathcal{O}(\rho^5)$, absorbed in the uniform fourth-order remainder. Projecting onto $W^\perp$ eliminates the entire linear contribution, leaving $\rho^3 P_{W^\perp} D + \mathcal{O}(\rho^4)$. Its norm is at least $\rho^3 \|P_{W^\perp} D\|/2$ for all sufficiently small positive ρ, uniformly over any weights that could satisfy equality. This contradicts $F_\rho = G_w(\rho)$. Because the estimates are uniform in w, the argument excludes arbitrary selections, not just smooth weight functions. □

Proposition F.5 (An SPD counterexample to pointwise exact fusion)**.** *In $\mathcal{P}_2$ with the affine-invariant metric, take $p = I$ and*

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad Y = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad Z = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{69}$$

For all sufficiently small positive ρ, the equal two-layer block on $(e^{\rho X}, I, e^{\rho Y})$ has no exact one-layer representation with any convex weight vector, even one chosen after this triple is observed.

Proof. At I, the metric is the Frobenius inner product, and X, Y, Z are orthonormal. The curvature formula in Appendix D yields $D = [[X, Y], Y - X]/384$ for the equal-block defect. Direct multiplication gives

$$[[X, Y], Y - X] = \begin{pmatrix} 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & -1 \end{pmatrix}, \qquad \langle D, Z \rangle_F = -\frac{1}{384}.$$

Since $Z \perp \mathrm{span}\{X, Y\}$, the normal projection is nonzero. Lemma F.4 proves the claim. The matrix exponential maps these directions to SPD matrices, and all inputs and means lie in a common local mean domain when ρ is sufficiently small. □

The counterexample also embeds in $\mathcal{P}_n$ for $n \geq 2$ using block-diagonal matrices with an identity block in the remaining coordinates. The corresponding submanifold has the product affine-invariant metric, and the unique means preserve the block structure. Thus the pointwise exact target is not generally attainable even in a geometry relevant to the existing downstream experiment. This observation does not assert that its particular recorded inputs violate feasibility, or change their measured errors. Nor does it contradict the near-zero E2 oracle errors on its tested two-dimensional space forms, where the normal obstruction above is absent. Absence of that obstruction alone is not a sufficiency theorem.

F.3 Finite-sample interpretation of sign-flip summaries

Let $D = (D_1, \ldots, D_n)$ denote subject-level paired differences after any within-subject seed averaging. A sufficient null hypothesis for a sign-flip test is joint invariance $D \overset{d}{=} s \odot D$ for every $s \in \{-1, 1\}^n$. Independent distributions symmetric about zero satisfy this hypothesis; independence and zero means alone do not (Winkler et al., 2014). For a specified statistic T, the complete-enumeration quantity is

$$p_T(D) = 2^{-n} \sum_{s \in \{-1,1\}^n} \mathbf{1}\{T(s \odot D) \geq T(D)\}. \tag{70}$$

Here $s \odot D$ denotes componentwise multiplication. Under the stated invariance, this is a finite-sample valid p value: conditional on the sign orbit, the observed point is uniform over the distinct orbit elements, and its rank among the statistics is uniform in the absence of ties. Counting all tied values in the upper tail only enlarges p_T, so $\Pr(p_T \leq a) \leq a$ for $a \in [0, 1]$. Zero entries and repeated orbit values have constant multiplicity and do not invalidate this argument. Thus "exact" refers to a valid

finite-sample randomization calculation under its invariance null, not to validity for every mean-zero population.

For an explicit counterexample to the latter interpretation, let the D_i be independent, with $D_i = 0.01$ with probability 0.9 and $D_i = -0.09$ with probability 0.1. Then $\mathbb{E}D_i = 0$, and all differences are bounded in $[-1, 1]$. For $n = 9$, the event that every difference is positive has probability $0.9^9 = 0.387420489$. On this event, the two-sided statistic $T(D) = |\sum_i D_i|$ is maximal only for the all-positive and all-negative sign patterns, giving $p_T = 2/2^9 = 0.00390625$. A nominal 5% test of the unrestricted mean-zero null would therefore reject with probability at least 38.7%.

Equation (70) and this counterexample explain the conditions needed for an interpretation of the BCI summaries. The frozen bundle contains the implementation and all nine differences for each method. Using $T(D) = |\operatorname{mean}(D)|$, complete enumeration of all 2^9 sign assignments, and an inclusive upper tail reproduces $p = 0.125$, 1.000, and 0.96875 for path-product, learned-fixed, and input-adaptive fusion after within-subject seed averaging. The records certify this arithmetic, but they cannot by themselves establish the sign-invariance assumption. In particular, a nonsignificant difference test is not an equivalence test with a prespecified margin.

F.4 Aggregate error, accuracy, and prediction agreement

A maximum over run-level mean squared errors is not a maximum individual squared error. For nonnegative errors $e_1, \ldots, e_N$, $\mathrm{MSE} = N^{-1} \sum_i e_i$ implies only $\max_i e_i \le N\,\mathrm{MSE}$. With the existing E2 test size $N = 5000$, a bound 3.331×10^{-16} on MSE gives the weaker individual bound 1.6655×10^{-12}, not 3.331×10^{-16}. The former is an algebraic upper bound, not a newly measured maximum. The manuscript therefore reports the original value specifically as the maximum *run-level held-out MSE*.

Likewise, a small difference between subject-mean balanced accuracies gives neither a per-subject bound nor a trial-wise agreement rate. To see the latter without any probabilistic assumption, consider a balanced four-class test set. On half of the trials within each class let classifier h_1 predict the correct label and h_2 the next label cyclically; reverse their roles on the other half. Both balanced accuracies are $1/2$, but the classifiers disagree on every trial. Thus the retained BCI results establish nearly equal reported aggregate scores, not near-identical decisions. A claim about decision preservation requires the separate disagreement counts.