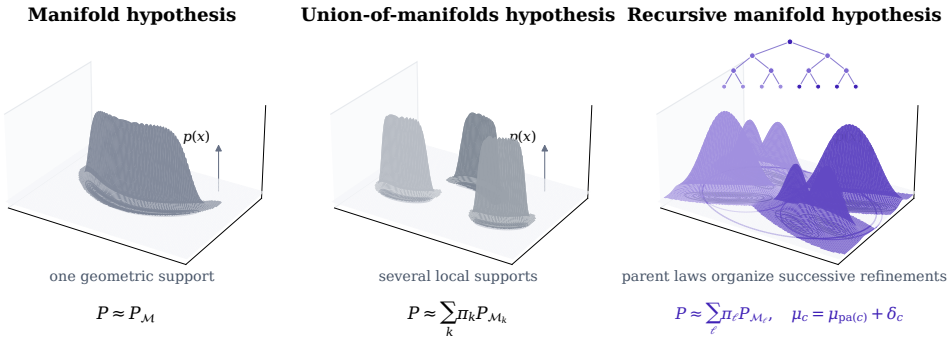


Figure 1: Three hypotheses about the organization of high-dimensional data. The manifold hypothesis posits one geometric support; the union-of-manifolds hypothesis allows several local supports. The recursive manifold hypothesis adds coarse-to-fine relations among their conditional laws. We formalize and test this relational claim; local manifold geometry remains an independent companion condition.

THE RECURSIVE MANIFOLD HYPOTHESIS FOR HIERARCHICAL DATA LAWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many real-world distributions are organized at more than one resolution. A coarse conditional law captures structure shared by its descendants; conditioning more finely reveals additional variation, while the resulting laws remain related through their common parent. We call this the *recursive manifold hypothesis*.

We formalize the hypothesis as multiscale conditional refinement of probability laws across successive coarse-to-fine relations. The energy distance metric between probability distributions compares the laws model-independently and allows us to decompose their variation exactly. Each parent–child relation has three coordinates: absolute law variation V, the fraction F revealed by the children, and parent organization A relative to the expected organization of a random parent map with the same group sizes. The product $R = FA$ yields a sufficient population criterion for one specified parent–child relation. We estimate this profile from pooled energy-distance U-statistics and quantify uncertainty in R by an observation jackknife.

Experiments center on ImageNet-1k and Amazon Reviews 2023 and include push-forward laws from DINOv2-L and Qwen3-Embedding-4B. Across image, text, and spiking-speech datasets, 13 of 14 relation–representation pairs reveal additional variation that the supplied parent organizes better than a random map with the same group sizes. ImageNet-1k, Amazon Reviews, FGVC-Aircraft, and ImageNet-R satisfy the criterion on two successive relations. Spiking Heidelberg Digits distinguishes structural strength across attributes: speakers are strongly organized within words, whereas the language attribute provides little organization of word laws. Excluding singleton branches preserves these empirical findings. These results suggest objectives and models that respect the hierarchical organization in data laws.

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.

1 INTRODUCTION

Real-world variation is resolved in stages. Aircraft from one manufacturer share properties that distinguish them from aircraft made by another; within a manufacturer, families retain common structure while their variants differ in finer ways. A spoken word constrains its acoustic distribution, within which individual speakers introduce further variation. Refinement in these examples does more than partition a population: it exposes variation hidden at a coarser level, and the refined laws remain related through structure shared within their parent.

We propose the following distributional hypothesis:

> Real-world data laws admit coarse-to-fine organizations in which parent laws capture structure shared by their descendants, while successive refinements introduce additional variation that remains structured within each parent. This organization occurs at multiple scales.

We call this statement the *recursive manifold hypothesis*. The name extends the lineage from a manifold to a union of manifolds and then to relations among conditional laws. In this distributional extension, multiscale effects need not change support: they may instead reorganize probability mass on the same support. The hypothesis therefore covers manifold-supported, full-dimensional, and discrete laws, alongside complementary local geometry.

The main technical task is to compare complete conditional laws in a model-independent way while distinguishing variation introduced by refinement from variation already resolved by the parent. Energy distance is an established metric between probability distributions (Székely & Rizzo, 2013; Lyons, 2013). Its Hilbert-space representation turns each conditional law into a vector and yields an exact variance decomposition. At parent–child relation j, we obtain

$$(V_j, F_j, A_j).$$

Here V_j is the absolute variation among child laws, F_j is the fraction revealed at that refinement, and A_j measures how much better the specified parent organizes that variation than the expectation under a size-matched random parent map.

For one specified relation, the formal condition requires newly revealed variation and better-than-random parent organization. Both hold at the population level exactly when

$$R_j = F_j A_j > 0.$$

The profile (V_j, F_j, A_j) describes the kind and scale of organization; R_j provides the relation's sufficient criterion and uncertainty interval.

Together, these quantities describe how much variation is present, where new variation enters, and how strongly the preceding scale organizes it. This view suggests learning systems that share structure where parent organization is strong and specialize where refinement introduces new variation.

2 A DISTRIBUTIONAL FRAMEWORK FOR RECURSIVE REFINEMENT

The hypothesis concerns relations among complete conditional laws. We therefore need a geometry of probability distributions that presupposes no density model and retains more than a selected set of moments. Energy distance provides such a geometry through an injective Hilbert representation. The resulting coarse-to-fine decomposition is a law of total variance for conditional laws. Proofs of all propositions are given in appendix A.

2.1 THE HIERARCHICAL DATA LAW

Let $X \sim P$ take values in a metric space $(\mathcal{X}, d)$ of strong negative type, and let

$$C_0 \prec C_1 \prec \cdots \prec C_L$$

be nested discrete variables. Here C_0 is constant and each C_{j-1} is a function of C_j. The observation law, metric, nested variables, and parent maps define a *conditional-law hierarchy*. Changing the representation or a parent map changes this mathematical object.

Write $P_{j,c} = P(X \in \cdot \mid C_j = c)$, $\pi_{j,c} = \Pr(C_j = c)$, and $v = \mathrm{pa}(c)$. The law of total probability fixes every parent law:

$$P_{j-1,v} = \sum_{c \in \mathrm{ch}(v)} \omega_{c|v} P_{j,c}, \qquad \omega_{c|v} = \frac{\pi_{j,c}}{\pi_{j-1,v}}. \tag{1}$$

Thus every parent is the probability-weighted mixture of its children, and passing to the next level resolves structure already present in the joint law without introducing another density model.

2.2 Variation revealed by conditioning

For laws P and Q with finite first moments, let $X, X' \sim P$ and $Y, Y' \sim Q$ be mutually independent. Their energy distance is

$$\mathcal{E}(P, Q) = 2\,\mathbb{E}d(X, Y) - \mathbb{E}d(X, X') - \mathbb{E}d(Y, Y'). \tag{2}$$

On a space of strong negative type, this established distributional distance admits an injective map $P \mapsto \mu_P$ into a Hilbert space $\mathcal{H}$, with $\mathcal{E}(P, Q) = 2\|\mu_P - \mu_Q\|^2_{\mathcal{H}}$ (Székely & Rizzo, 2013; Lyons, 2013; Sejdinovic et al., 2013). We call μ_P the *law embedding* of P. Mixture linearity maps eq. (1) to the Hilbert-space mean $\mu_{j-1,v} = \sum_c \omega_{c|v} \mu_{j,c}$.

The symbols now correspond directly to vectors: $\mu_{j,c}$ embeds the child law $P_{j,c}$, $\mu_{j-1,v}$ embeds its parent law, and $\mu_{0,\varnothing}$ embeds the root law. Define

$$\begin{aligned} \delta_{j,c} &= \mu_{j,c} - \mu_{j-1,v}, \\ I_j &= \sum_c \pi_{j,c} \|\delta_{j,c}\|^2_{\mathcal{H}}, \\ V_j &= \sum_c \pi_{j,c} \|\mu_{j,c} - \mu_{0,\varnothing}\|^2_{\mathcal{H}}. \end{aligned} \tag{3}$$

The innovation I_j is the law variation exposed when parents are resolved into children. The total V_j is all law variation visible at level j relative to the root; V_{j-1} uses the same probability measure after aggregation over parent laws.

Proposition 1 (Scale-wise law-variance decomposition)**.** *For every finite conditional-law hierarchy,*

$$V_j = V_{j-1} + I_j, \qquad V_L = \sum_{j=1}^{L} I_j.$$

This identity gives an exact coarse-to-fine account: child-level variation is the sum of variation already resolved by the parents and variation newly exposed by conditioning. It telescopes when one probability measure π is used throughout the hierarchy. Relation-specific balancing retains the exact one-step identity but need not sum to one terminal V_L.

Proposition 2 (Exact energy representation)**.** *At one relation, index the m child laws by a, write $\mathrm{pa}(a)$ for their parents, and abbreviate $\pi_a = \pi_{j,a}$ and $\pi_v = \sum_{a:\mathrm{pa}(a)=v} \pi_a$. Then*

$$\begin{aligned} V_j &= \frac{1}{2} \sum_{a<b} \pi_a \pi_b \mathcal{E}(P_a, P_b), \\ I_j &= \frac{1}{2} \sum_v \frac{1}{\pi_v} \sum_{\substack{a<b \\ \mathrm{pa}(a)=\mathrm{pa}(b)=v}} \pi_a \pi_b \mathcal{E}(P_a, P_b). \end{aligned} \tag{4}$$

Both quantities are therefore determined by pairwise energy distances among the child laws. For $V_j > 0$, define

$$F_j = \frac{I_j}{V_j}, \qquad \kappa_j = \frac{V_{j-1}}{V_j} = 1 - F_j. \tag{5}$$

The innovation fraction F_j is the share of child-law variation first revealed at this relation. Its complement κ_j is the share already resolved by the parent. Thus $F_j = 0$ means that conditioning introduces no population variation, whereas $\kappa_j = 0$ means that the proposed parent resolves none of the variation among its children.

2.3 THE CONDITIONAL-REFINEMENT PROFILE

Consider relation j, with m_j child laws and K_j parents. A size-matched reference assigns the children uniformly at random to nonempty parents while preserving the K_j parent sizes.

Proposition 3 (Random-hierarchy baseline)**.** *Give the m_j child laws equal probability. Conditional on their law embeddings, the random assignment G satisfies*

$$\mathbb{E}_G[I_j(G)] = \frac{m_j - K_j}{m_j - 1} V_j, \qquad \mathbb{E}_G[\kappa_j(G)] = \frac{K_j - 1}{m_j - 1}.$$

Let

$$b_j = \frac{K_j - 1}{m_j - 1}, \qquad t_j = 1 - b_j = \frac{m_j - K_j}{m_j - 1}.$$

The first quantity is the expected parent organization of a size-matched random map, and the second is its expected innovation. Therefore, the specified parent organizes the child laws better than this reference exactly when $\kappa_j > b_j$, or equivalently $F_j < t_j$.

Define the adjusted parent organization of relation j by

$$A_j = \frac{\kappa_j - b_j}{1 - b_j} = 1 - \frac{m_j - 1}{m_j - K_j}(1 - \kappa_j).$$

For equal child weights, A_j is algebraically the usual degrees-of-freedom-adjusted R^2 for a categorical parent factor with $K_j - 1$ fitted degrees of freedom (Karch, 2020). More precisely, A_j measures parent organization after subtracting the organization expected from the number and sizes of its groups. Together, (V_j, F_j, A_j) reports absolute law variation, the share first revealed by the children, and adjusted parent organization.

The single sufficient criterion is

$$R_j = F_j \frac{\kappa_j - b_j}{1 - b_j} = F_j \frac{t_j - F_j}{t_j}. \tag{6}$$

It is positive exactly when conditioning reveals variation $F_j > 0$ and the supplied parent organizes that variation better than the size-matched random expectation $A_j > 0$. It vanishes at either boundary and is negative when the parent is less informative than the reference.

On its positive interval, R_j is a concave quadratic with maximum $t_j/4$ at $F_j = t_j/2$. Since t_j depends on (m_j, K_j), the profile carries the substantive comparison across relations: V_j sets the absolute scale, F_j locates the new variation, and A_j quantifies the supplied parent's contribution. The score R_j determines whether their two requirements hold together.

Definition 1 (Multiscale conditional refinement)**.** *A relation exhibits* structured innovation *when $R_j > 0$. A specified conditional-law hierarchy exhibits* multiscale conditional refinement *when structured innovation occurs at two or more nontrivial relations.*

2.4 THE HYPOTHESIS AND ITS GEOMETRIC SPECIALIZATION

The *recursive manifold hypothesis* states that real-world data laws admit coarse-to-fine organizations in which parent laws capture shared structure and successive refinements reveal additional, structured variation at multiple scales. Multiscale conditional refinement is its formal, distributional counterpart. The random-parent expectation fixes the zero point for each relation, and repeated observations from the specified child laws determine sampling uncertainty.

A geometric specialization adds the premise that terminal conditional laws satisfy a finite-complexity manifold or union-of-manifolds account. The energy analysis evaluates relations among complete laws; controlled geometric tests can characterize their terminal supports.

3 FROM THE POPULATION PROFILE TO EMPIRICAL EVIDENCE

The preceding derivation reduces a hierarchical relation to pairwise distances among its conditional laws. We estimate those distances directly, pool the resulting matrices, and then form the profile. This order preserves the population decomposition without fitting a density to each child law.

3.1 A RELATION-BALANCED LAW

Our estimand is variation over the represented conditional laws, rather than over corpus observations. We therefore give every child law at relation j the weight $1/m_j$. This deliberate relation-balanced design prevents a frequent child from determining a statement intended to compare laws and gives each represented refinement equal influence. Because the balancing is performed anew at every relation, the profile rows describe separate relations rather than additive terms under one terminal measure.

Substituting the U-statistic distance matrix into proposition 2 yields $(\widehat{V}_j, \widehat{F}_j, \widehat{A}_j)$, and eq. (6) yields $\widehat{R}_j$. The estimators $\widehat{V}_j$ and $\widehat{I}_j$ are unbiased. Their ratio, and hence $\widehat{R}_j$, need not be exactly unbiased in finite samples.

3.2 ONE CRITERION AND ITS UNCERTAINTY

For computational feasibility and equal per-law precision, each dataset is reduced to a leaf-balanced subset and divided into mutually disjoint blocks. The selection rules and per-leaf sample counts are specified in section 4; every observation in the resulting subset contributes to the main profile. At each relation, we average the blockwise U-statistic distance matrices and compute $\widehat{I}_j/\widehat{V}_j$ only after this pooling. Averaging at the distance-matrix level removes the arbitrary block index from the estimand instead of treating computational blocks as replicate estimates of a nonlinear ratio.

Sampling uncertainty is estimated by a stratified delete-one jackknife. We remove one observation in turn from each block–child stratum and propagate its effect through both U-statistic terms. The resulting first-order variance estimate gives the approximate 95% normal intervals.

The inferential population consists of new observations from the fixed, named child laws under the relation-balanced design. A relation satisfies the sufficient condition when the lower interval endpoint for R_j exceeds zero; a system is multiscale when this occurs at two or more nontrivial relations. The profile describes each relation, while this single criterion determines support for the formal hypothesis.

4 EXPERIMENTAL DESIGN

Each experiment specifies an observation law, representation, metric, child weighting, and hierarchy, then applies the same estimator and sufficient criterion.

4.1 A COMMON MODEL-FREE PROTOCOL

We evaluate prespecified, interpretable conditional-law hierarchies. Their relations come from released dataset metadata or from deterministic paths in the WordNet hypernym graph. Prespecification prevents the measured distances from selecting the relations they later evaluate; interpretability makes those relations scientifically meaningful under finite data. Observations are then allocated to mutually disjoint, leaf-balanced subsets.

4.2 TAXONOMIES FOR IMAGES AND TEXT

CIFAR-100 provides an official superclass-to-class relation (Krizhevsky, 2009). We measure it using both 8-by-8 luminance block means and normalized DINOv2-L class tokens (Oquab et al., 2024). ImageNet-1k class synsets define coarse-to-middle and middle-to-class WordNet relations (Deng et al., 2009; Miller, 1995); both are measured in DINOv2-L. Amazon Reviews 2023 supplies an Electronics category hierarchy, represented by fixed hashed token counts and Qwen3-Embedding-4B text features (Hou et al., 2026; Zhang et al., 2025). CIFAR-100 and Amazon Reviews use five disjoint blocks of 48 observations per terminal law; ImageNet-1k uses two blocks of 24.

4.3 HIERARCHIES FOR AIRCRAFT, SHD, AND IMAGENET-R

The next three datasets test conditional-law hierarchies derived from interpretable relations: an engineering taxonomy, linguistic attributes, and semantic hypernyms.

FGVC-Aircraft. The released manufacturer–family–variant hierarchy has 30 manufacturers, 70 families, and 100 variants (Maji et al., 2013). The primary observation law is a deterministic 32-by-32 RGB bilinear resize in $[0, 1]$ with Euclidean distance; normalized, unprojected DINOv2-L features provide a representation sensitivity.

Spiking Heidelberg Digits. SHD contains English and German spoken digits with speaker metadata (Cramer et al., 2022). Its nested variables are language, digit word within language, and speaker identity within each word, giving 2, 20, and 240 conditional laws. For the primary law, we take the coordinate-wise square root of the normalized histogram, so that Euclidean distance matches Hellinger distance up to a constant.

ImageNet-R. The ImageNet-R dataset contains 200 rendition classes (Hendrycks et al., 2021). For each synset, we select the longest WordNet hypernym path with deterministic lexicographic tie-breaking and take ancestors four and two edges above the leaf. The primary law uses normalized, unprojected 1024-dimensional DINOv2-L features; 32-by-32 RGB is the representation sensitivity.

5 Empirical evidence

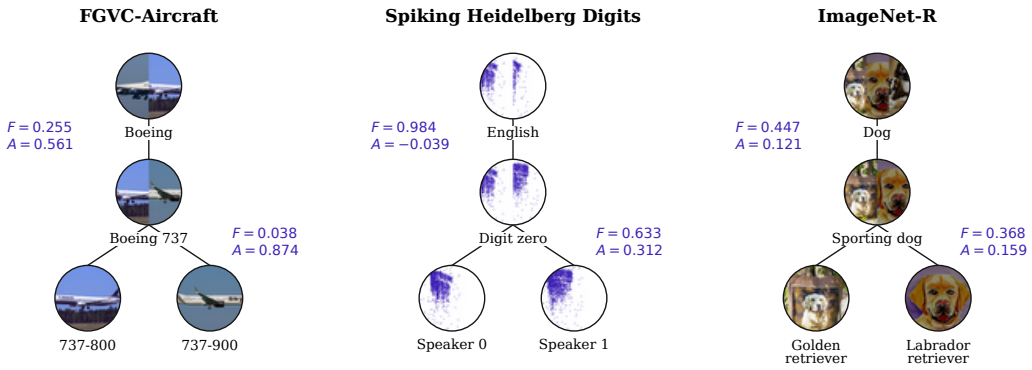


Figure 2: **Examples from three prespecified conditional-law hierarchies.** Each panel traces one parent, one child, and two grandchildren from the hierarchy used in the experiment. Thumbnails illustrate node contents; the adjacent F and A values are computed from all child laws at the corresponding relation. Lines denote nesting.

5.1 Refinement has a distinct profile at each scale

ImageNet-R exhibits structured refinement at both WordNet relations in the full DINOv2-L representation (fig. 3):

$$(V_2, F_2, A_2) = (0.1059, 0.447, 0.121), \qquad R_2 = 0.0540\ [0.0477, 0.0603],$$
$$(V_3, F_3, A_3) = (0.1346, 0.368, 0.159), \qquad R_3 = 0.0584\ [0.0541, 0.0626].$$

At both scales, the child laws reveal additional variation. The supplied parent organizes this variation better than a size-matched random map, which distinguishes the result from an arbitrary hierarchy with the same group sizes.

FGVC-Aircraft also satisfies the criterion at two scales, but allocates its variation differently:

$$(V_2, F_2, A_2) = (0.314, 0.255, 0.561), \qquad R_2 = 0.1428\ [0.1388, 0.1467],$$
$$(V_3, F_3, A_3) = (0.260, 0.038, 0.874), \qquad R_3 = 0.0334\ [0.0155, 0.0512].$$

Manufacturer strongly organizes its families, which introduce substantial variation. Families organize their variants even more strongly, although the variants add comparatively little. Structured refinement is present at both relations, with most innovation at the coarser one.

The Spiking Heidelberg Digits dataset shows why innovation and organization are separate concepts. From word to speaker, $(V_3, F_3, A_3) = (0.182, 0.633, 0.312)$ and $R_3 = 0.1976\ [0.1953, 0.1999]$:

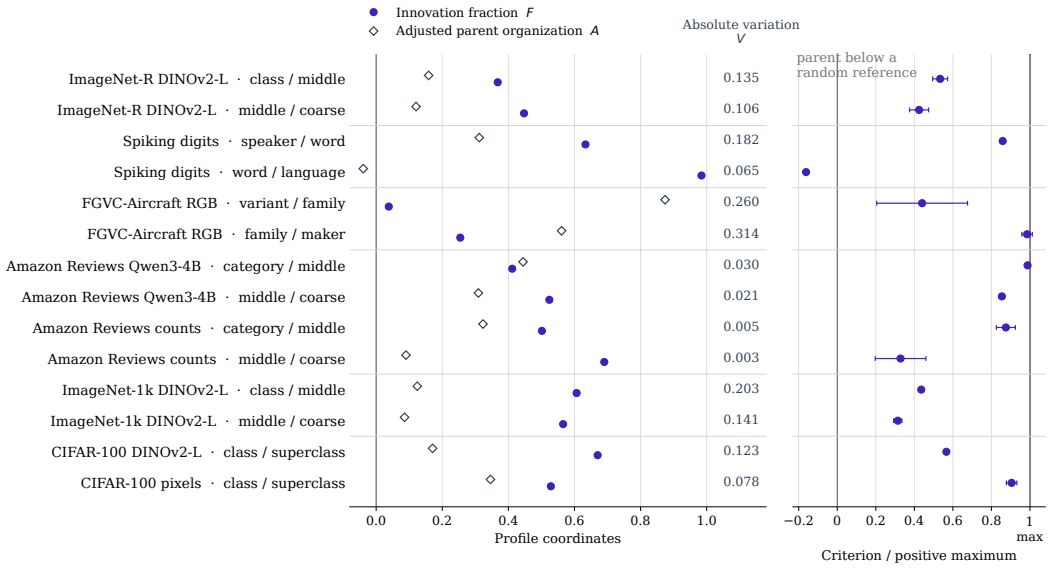


Figure 3: **Conditional-refinement profiles across data laws and relations.** Left: innovation fraction F_j, adjusted parent organization A_j, and absolute law variation V_j for every specified relation. Right: the sufficient criterion $R_j = F_j A_j$ with approximate 95% leaf-stratified observation-jackknife intervals, divided by its relation-specific positive maximum $t_j/4$. The heavy line at 1 marks this maximum; the negative range is relation-dependent. Exact values are listed in tables 3 and 4.

speakers introduce substantial variation that is organized within words. From language to word, words account for nearly all visible variation, $F_2 = 0.984$, while adjusted organization by the language attribute is $A_2 = -0.039$. The resulting score, $R_2 = -0.0383\ [-0.0410, -0.0357]$, shows that this attribute organizes the word distributions less effectively than the average size-matched grouping.

5.2 The pattern recurs across laws and representations

CIFAR-100 class laws satisfy the criterion in both 8-by-8 luminance block means and DINOv2-L. Their innovation fractions are 0.529 and 0.670, and their adjusted parent organizations are 0.346 and 0.171. Amazon Reviews also satisfies the criterion at two successive category relations in both hashed counts and Qwen3-Embedding-4B. In Qwen3, the coarse and fine relations have profiles $(V, F, A) = (0.0212, 0.524, 0.309)$ and $(0.0298, 0.412, 0.444)$, with score intervals $[0.1595, 0.1646]$ and $[0.1823, 0.1837]$.

ImageNet-1k provides another multiscale instance. In the full 1024-dimensional DINOv2-L feature, its coarse and fine WordNet relations have profiles $(V, F, A) = (0.1413, 0.566, 0.086)$ and $(0.2034, 0.606, 0.124)$, with score intervals $[0.0453, 0.0519]$ and $[0.0739, 0.0770]$. The supplied relations exceed their size-matched random-parent references at both scales.

The paired representations also show how a pushforward changes the law being measured. Aircraft DINOv2, ImageNet-R RGB, and untransformed SHD frequencies retain support for the same relations as their primary representations, but they change the magnitude and allocation of variation. For example, the ImageNet-R RGB interval is wider at the coarse relation, whereas DINOv2 resolves that relation more sharply. The evidence is model-independent and specific to each declared pushforward.

Across all 14 relation–law pairs in fig. 3, the non-singleton restriction preserves every conclusion under the $R_j > 0$ sufficiency condition (table 5).

6 RELATION TO PRIOR WORK

From manifolds to relations among laws. The manifold hypothesis explains why nonlinear coordinates can recover useful low-dimensional structure (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003) and becomes statistically testable once geometric complexity is controlled (Fefferman et al., 2016). The union-of-manifolds hypothesis allows disconnected components of different dimensions (Brown et al., 2023); the CW-complex hypothesis further permits smooth cells of different dimensions to meet along lower-dimensional boundaries (Wang & Wang, 2024). In particular, the pushforward argument of Brown et al. (2023) uses preservation of connectedness under continuous maps to show how disconnected observations can obstruct a single-manifold account. The recursive manifold hypothesis is a conceptual continuation of this progression. A manifold describes one support, a union describes several, and recursive refinement evaluates whether their conditional laws are organized by coarser laws at successive scales.

Distances between distributions. DISCO partitions distance-based variation into within- and between-group terms and supports consistent multi-sample distributional tests (Rizzo & Székely, 2010). Energy distance, distance covariance, and maximum mean discrepancy are linked through semimetrics of negative type and reproducing-kernel embeddings (Sejdinovic et al., 2013). Kernel mean embeddings give the broader interpretation of a probability distribution as a vector (Muandet et al., 2017), while energy and kernel two-sample tests are closely related (Gretton et al., 2012). Conditional MMD compares conditional distributions through operator, conditional-mean, and joint embeddings (Moskvichev et al., 2026). Our setting starts from a discrete conditional-law hierarchy and measures how successive parent maps resolve the variance of its law embeddings.

Distance decompositions also underlie hierarchical clustering. The joint between–within criterion of Székely & Rizzo (2005) extends Ward's method to powers of Euclidean distance and uses that decomposition to construct a hierarchy. Here an interpretable hierarchy is prespecified and evaluated against its exact size-matched expectation. The resulting (V, F, A) profile separates absolute variation, newly revealed variation, and adjusted parent organization. Its last coordinate is algebraically an adjusted R^2 (Karch, 2020); multiplication by innovation gives the sufficient relation-wise criterion.

Geometric multiresolution. Geometric multiresolution analysis (GMRA) constructs nested cells, fits local affine principal components, and represents changes across scales by geometric wavelets (Allard et al., 2012). Robust and adaptive variants give approximation guarantees for laws near low-dimensional sets (Maggioni et al., 2016; Liao & Maggioni, 2019), and related methods address multi-manifold structure (Medina et al., 2018). GMRA discovers a partition to approximate one observation law. A conditional-law hierarchy instead relates specified conditional distributions; a learner could combine the two ideas by inducing such relations under complexity control.

7 IMPLICATIONS AND OUTLOOK

7.1 REAL-WORLD VARIATION IS RESOLVED SCALE BY SCALE

The empirical profiles rule out a single uniform description of class variation: some refinements add little new variation but are strongly organized; others add substantial variation without better-than-random parent organization. The ImageNet-R and ImageNet-1k datasets instantiate this pattern on two successive WordNet relations. In the Aircraft dataset, we observe that most new variation occurs between manufacturers and families, while families strongly organize the smaller changes among variants. We observe an analogous contrast in SHD: word laws differ strongly, but the language attribute provides little organization, whereas speakers introduce variation that is organized within words. A named level in a taxonomy is therefore a statistical refinement only when conditioning changes the law and its parent organizes the resulting variation.

When the hypothesis is evaluated in a learned representation, it concerns the corresponding pushforward law. A feature map can sharpen one refinement, suppress another, or change which aspects of the complete conditional law are expressed. Comparing profiles across pushforwards measures which refinements a representation retains. The DINOv2 profiles may include semantic regularities

acquired during large-scale image pretraining; the RGB profiles show which of the tested relations remain in a direct image representation.

7.2 Distributional refinement and local geometry

The distributional formulation concerns organization among conditional laws. A geometric specialization can add controlled tests of dimension, reach, connectedness, or approximation error within their supports (Fefferman et al., 2016; Brown et al., 2023). This separation permits the same relational hypothesis to describe full-dimensional Gaussians, discrete laws, or terminal manifolds.

For experimental validity under finite data, we evaluate prespecified, interpretable relations from metadata and deterministic semantic paths. A learner fitted to the marginal law may instead induce or refine its own conditional-law hierarchy. Such a learner is responsible for controlling the complexity of that selection and evaluating it on independent data. The success of interpretable supplied relations suggests that hierarchy-inducing learners can use the same sufficient criterion to favor coarse-to-fine organizations that remain stable out of sample.

7.3 Consequences for learning and data design

At every scale, organization favors sharing information among descendants, whereas innovation requires specialization. The measured profile can therefore inform hierarchical experts, multiscale adapters, conditional transport maps, or priors whose sharing strength varies with scale. The proof-of-concept PPCA demonstration is a simple instance: tangent sharing helps when tangent structure is inherited along the relation, while the complete-law profile can also detect organization outside that parameterization (appendix B).

In particular, rate-reduction approaches to representation learning model multi-class data through class-conditional subspaces and mixtures of low-dimensional Gaussians (Chan et al., 2022). Within this Gaussian/subspace model, the natural multiscale analogue is a sequence of coding-rate gaps between each pooled conditional law and its refined components, revealing where a representation resolves structure beyond a single class partition.

The same principle applies to representations, curricula, and synthetic data. A representation may be selected for the relevant refinements it exposes. A curriculum may present coarse regularities before the innovations that depend on them, extending the easy-to-difficult ordering of curriculum learning (Bengio et al., 2009). A generator can be compared with real data through its refinement profile rather than only through marginal moments or aggregate complexity. Cellular-automata pretraining provides a precedent for connecting structural properties of a data source to transfer (Lee et al., 2026). Here the corresponding prediction is that matching relations relevant to a data-poor descendant should aid learning more than matching unrelated relations.

The profile also gives a statistical view of dataset design. A nominally fine distinction with little innovation is nearly redundant in the chosen law; a coarse category below the random-parent expectation poorly summarizes the relations among its children. These outcomes need not be annotation errors. They identify where semantic and statistical organization diverge, which is precisely where a learner must decide whether to share or specialize.

8 Conclusion

The recursive manifold hypothesis gives precise form to a familiar view of real-world data: coarse laws capture structure shared by their descendants, while successive refinements reveal additional variation that remains organized within each parent. Multiscale conditional refinement is its formal, distributional counterpart.

Across the specified image, text, and spiking-speech laws, ImageNet-1k, Amazon Reviews, FGVC-Aircraft, and ImageNet-R satisfy the criterion at multiple scales, hence at least two.

Combined with local geometric tests, conditional-refinement profiles can characterize relations among laws and structure within terminal laws; combined with learning mechanisms, they can guide sharing and specialization according to the measured refinement.

AI USE STATEMENT

Generative AI tools provided minor assistance across the workflow, with major contributions to scripting and writing programs. They assisted with literature search, conceptual and methodological critique, formalization and proofs, experimental design, implementation, figure development, and manuscript drafting. The author comprehensively validated the mathematical arguments, code, experiments, results, citations, and prose and takes responsibility for the final content.

REPRODUCIBILITY STATEMENT

The appendices contain complete proofs, estimator details, sensitivity and calibration analyses, and the PPCA demonstration. The anonymous supplement contains the implementation, experiment configurations, tests, result bundles with checksum manifests, and scripts used to verify the results and regenerate the figures. Data and model sources, revisions, preprocessing, and usage terms are recorded in `DATA_PROVENANCE.md`; raw third-party data and derived representations of individual examples are not redistributed.

REFERENCES

William K. Allard, Guangliang Chen, and Mauro Maggioni. Multi-scale geometric methods for data sets ii: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 32 (3):435–462, 2012. doi:10.1016/j.acha.2011.08.001.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi:10.1162/089976603321780317.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48. ACM, 2009. doi:10.1145/1553374.1553380.

Bradley C. A. Brown, Anthony L. Caterini, Brendan L. Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=aJp8UXRKvVm.

Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. ReduNet: A white-box deep network from the principle of maximizing rate reduction. *Journal of Machine Learning Research*, 23(114):1–103, 2022. URL https://jmlr.org/papers/v23/21-0631.html.

Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke. The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2744–2757, 2022. doi:10.1109/TNNLS.2020.3044364.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi:10.1109/CVPR.2009.5206848.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi:10.1090/jams/852.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi:10.1198/016214506000001437.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Christian Horvat and Jean-Pascal Pfister. Density estimation on low-dimensional manifolds: An inflation–deflation approach. *Journal of Machine Learning Research*, 24(61):1–37, 2023.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders. *arXiv preprint arXiv:2403.03952*, 2026. URL https://arxiv.org/abs/2403.03952.

Julian Karch. Improving on adjusted R-squared. *Collabra: Psychology*, 6(1):45, 2020. doi:10.1525/collabra.343.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Dan Lee, Seungwook Han, Akarsh Kumar, and Pulkit Agrawal. Training language models via neural cellular automata. *arXiv preprint arXiv:2603.10055*, 2026. URL https://arxiv.org/abs/2603.10055.

Wenjing Liao and Mauro Maggioni. Adaptive geometric multiscale approximations for intrinsically low-dimensional data. *Journal of Machine Learning Research*, 20(98):1–63, 2019. URL https://jmlr.org/papers/v20/17-252.html.

Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013. doi:10.1214/12-AOP803.

Mauro Maggioni, Stanislav Minsker, and Nate Strawn. Multiscale dictionary learning: Non-asymptotic bounds and robustness. *Journal of Machine Learning Research*, 17(2):1–51, 2016. URL https://jmlr.org/papers/v17/maggioni16a.html.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. URL https://arxiv.org/abs/1306.5151.

F. Patricia Medina, Linda Ness, Melanie Weber, and Karamatou Yacoubou Djima. Heuristic framework for multi-scale testing of the multi-manifold hypothesis, 2018. URL https://arxiv.org/abs/1807.00349.

George A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. doi:10.1145/219717.219748.

Peter Moskvichev, Siu Lun Chau, and Dino Sejdinovic. Measuring differences between conditional distributions using kernel embeddings, 2026. URL https://arxiv.org/abs/2605.02260.

Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141, 2017. doi:10.1561/2200000060.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Maria L. Rizzo and Gábor J. Székely. DISCO analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010. doi:10.1214/09-AOAS245.

Sam Roweis, Lawrence Saul, and Geoffrey Hinton. Global coordination of local linear models. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. doi:10.1126/science.290.5500.2323.

Dino Sejdinovic, Bharath K. Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi:10.1214/13-AOS1140.

Gábor J. Székely and Maria L. Rizzo. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, 22(2):151–183, 2005. doi:10.1007/s00357-005-0012-9.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi:10.1016/j.jspi.2013.03.018.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi:10.1126/science.290.5500.2319.

Michael E. Tipping and Christopher M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999a. doi:10.1162/089976699300016728.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, 61(3):611–622, 1999b. doi:10.1111/1467-9868.00196.

Ye Wang, Antonio Canale, and David Dunson. Scalable geometric density estimation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pp. 857–865. PMLR, 2016. URL https://proceedings.mlr.press/v51/wang16e.html.

Yi Wang and Zhiren Wang. CW complex hypothesis for image data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 51596–51613. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24bs.html.

Christopher K. I. Williams. A MCMC approach to hierarchical mixture modelling. In *Advances in Neural Information Processing Systems*, volume 12, 2000.

XianXing Zhang, David Dunson, and Lawrence Carin. Tree-structured infinite sparse factor model. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 785–792, 2011.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

A PROOFS

Proof of proposition 1. Fix level j, let $v = \mathrm{pa}(c)$, and decompose the displacement of child c from the root as

$$\mu_{j,c} - \mu_{0,\varnothing} = \big(\mu_{j,c} - \mu_{j-1,v}\big) + \big(\mu_{j-1,v} - \mu_{0,\varnothing}\big) = \delta_{j,c} + r_v.$$

After squaring and weighting, the only non-additive term is

$$2 \sum_{v} \sum_{c \in \mathrm{ch}(v)} \pi_{j,c} \langle \delta_{j,c}, r_v \rangle_{\mathcal{H}}.$$

For each parent v, mixture linearity gives

$$\sum_{c \in \mathrm{ch}(v)} \pi_{j,c} \delta_{j,c} = \pi_{j-1,v} \left(\sum_{c} \omega_{c|v} \mu_{j,c} - \mu_{j-1,v} \right) = 0.$$

Hence every parent-specific cross term vanishes. The squared child displacements sum to I_j, while $\sum_v \pi_{j-1,v} \|r_v\|_{\mathcal{H}}^2 = V_{j-1}$. Therefore $V_j = V_{j-1} + I_j$. Since $V_0 = 0$, applying this identity for $j = 1, \ldots, L$ yields $V_L = \sum_{j=1}^{L} I_j$. □

Proof of proposition 2. For vectors $z_1, \dots, z_m$ with weights $w_a \geq 0$, $\sum_a w_a = 1$, and mean $\bar{z} = \sum_a w_a z_a$, direct expansion gives

$$\sum_a w_a \|z_a - \bar{z}\|^2 = \sum_a w_a \|z_a\|^2 - \|\bar{z}\|^2 = \sum_{a<b} w_a w_b \|z_a - z_b\|^2.$$

Apply this identity globally to $z_a = \mu_{P_a}$ and $w_a = \pi_a$. Since $\mathcal{E}(P_a, P_b) = 2\|\mu_{P_a} - \mu_{P_b}\|^2$, the result is

$$V_j = \frac{1}{2} \sum_{a<b} \pi_a \pi_b \mathcal{E}(P_a, P_b).$$

Within parent v, apply the same identity with normalized weights $w_{a|v} = \pi_a / \pi_v$. Multiplication by the parent mass π_v gives

$$\sum_{a:\mathrm{pa}(a)=v} \pi_a \|\mu_{P_a} - \mu_{P_v}\|^2 = \frac{1}{2\pi_v} \sum_{\substack{a<b \\ \mathrm{pa}(a)=\mathrm{pa}(b)=v}} \pi_a \pi_b \mathcal{E}(P_a, P_b).$$

Summing over v yields the expression for I_j. □

Proof of proposition 3. Let the centered law embeddings be $z_a = \mu_{P_a} - \bar{\mu}$, so that $\sum_a z_a = 0$ under equal child weights. For a uniformly random subset S of size s, the standard finite-population identity is

$$\mathbb{E}\left[\sum_{a \in S} \|z_a - \bar{z}_S\|^2\right] = \frac{s-1}{m_j - 1} \sum_{a=1}^{m_j} \|z_a\|^2.$$

One way to verify it is to use the pairwise form $\sum_{a \in S} \|z_a - \bar{z}_S\|^2 = s^{-1} \sum_{a<b \in S} \|z_a - z_b\|^2$ and the inclusion probability $s(s-1)/[m_j(m_j-1)]$ for every pair. For fixed parent sizes $s_1, \dots, s_{K_j}$, weighting each within-parent sum by its parent probability s_k / m_j gives

$$\mathbb{E}_G[I_j(G)] = \frac{\sum_k (s_k - 1)}{m_j - 1} V_j = \frac{m_j - K_j}{m_j - 1} V_j.$$

Finally, proposition 1 implies $\kappa_j(G) = 1 - I_j(G)/V_j$. Taking expectations yields $\mathbb{E}_G[\kappa_j(G)] = (K_j - 1)/(m_j - 1)$. □

B Law-estimation consequences and a PPCA demonstration

B.1 A finite-sample consequence for law estimation

The decomposition also determines when estimating a child law can benefit from its parent. Let $\widehat{\mu}_a$ be an unbiased empirical estimate of child embedding μ_a, with positive average sampling error

$$T = \sum_a \pi_a \mathbb{E} \|\widehat{\mu}_a - \mu_a\|_{\mathcal{H}}^2.$$

From independent reference data, let $\widehat{\tau}_v$ and $\widehat{\tau}_0$ be unbiased estimates of parent embedding μ_v and root embedding μ_0. Their average estimation errors are

$$S_{\mathrm{par}} = \sum_a \pi_a \mathbb{E} \|\widehat{\tau}_{\mathrm{pa}(a)} - \mu_{\mathrm{pa}(a)}\|_{\mathcal{H}}^2, \qquad S_{\mathrm{root}} = \mathbb{E} \|\widehat{\tau}_0 - \mu_0\|_{\mathcal{H}}^2.$$

Proposition 4 (Finite-reference advantage of parent sharing)**.** *For $\lambda \in [0, 1]$, shrink each child estimate toward either its parent or the root:*

$$\widehat{\mu}_a^{\mathrm{par}}(\lambda) = (1 - \lambda)\widehat{\mu}_a + \lambda \widehat{\tau}_{\mathrm{pa}(a)}, \qquad \widehat{\mu}_a^{\mathrm{root}}(\lambda) = (1 - \lambda)\widehat{\mu}_a + \lambda \widehat{\tau}_0.$$

Their minimum weighted squared-error risks are

$$\frac{T(I_j + S_{\mathrm{par}})}{T + I_j + S_{\mathrm{par}}} \quad \textit{and} \quad \frac{T(V_j + S_{\mathrm{root}})}{T + V_j + S_{\mathrm{root}}},$$

respectively. Parent sharing has lower oracle-tuned risk exactly when $I_j + S_{\mathrm{par}} < V_j + S_{\mathrm{root}}$. With known targets, this reduces to $I_j < V_j$; either optimally weighted target also improves on the empirical child estimate whenever its target discrepancy is finite.

Thus I_j is the structural discrepancy incurred by shrinking a child toward its parent, whereas V_j is the corresponding discrepancy to the root. Finite reference samples add the explicit estimation terms S_{par} and S_{root}.

Proof of proposition 4. Consider first shrinkage toward the parent. Subtracting the true child embedding gives

$$\widehat{\mu}_a^{\mathrm{par}}(\lambda) - \mu_a = (1-\lambda)(\widehat{\mu}_a - \mu_a) + \lambda(\widehat{\tau}_{\mathrm{pa}(a)} - \mu_{\mathrm{pa}(a)}) + \lambda(\mu_{\mathrm{pa}(a)} - \mu_a).$$

Unbiasedness and independence make the three pairwise cross terms vanish in expectation. Summing with weights π_a therefore gives

$$R_{\mathrm{par}}(\lambda) = (1-\lambda)^2 T + \lambda^2 (I_j + S_{\mathrm{par}}).$$

The same expansion around the root yields

$$R_{\mathrm{root}}(\lambda) = (1-\lambda)^2 T + \lambda^2 (V_j + S_{\mathrm{root}}).$$

For either expression, first take the non-sampling term $D > 0$. Differentiating $(1-\lambda)^2 T + \lambda^2 D$ gives the unique minimizer $\lambda^* = T/(T+D)$, which lies in $(0,1)$, and substitution gives the minimum $TD/(T+D)$. This value is strictly increasing in D because its derivative is $T^2/(T+D)^2 > 0$. Hence parent shrinkage has lower minimum risk exactly when $I_j + S_{\mathrm{par}} < V_j + S_{\mathrm{root}}$. Setting the reference errors to zero gives $I_j < V_j$. Finally, $TD/(T+D) < T$ for finite $T, D > 0$, so either optimally weighted target improves on the empirical child estimate, whose risk is T. If $D = 0$, the same conclusion follows at $\lambda = 1$. □

B.2 Hierarchical probability models

Hierarchical models exploit relations among local laws by sharing parameters. Coordinated factor analyzers regularize neighboring coordinate systems (Roweis et al., 2002); tree-structured sparse factor models share loadings along branches (Zhang et al., 2011); hierarchical mixtures organize components probabilistically (Williams, 2000); and GEODE combines local principal-axis densities over a dyadic tree (Wang et al., 2016). PPCA and mixtures of PPCA provide classical local linear likelihoods (Tipping & Bishop, 1999b;a), while modern manifold density estimators relax their Gaussian assumptions (Horvat & Pfister, 2023). These models motivate the downstream demonstration but do not define the distributional hypothesis. Lastly, held-out logarithmic score (Gneiting & Raftery, 2007) determines whether one sharing rule can use a relation measured by the model-independent profile.

B.3 How a learner can exploit refinement

One compatible local model is probabilistic PCA (Tipping & Bishop, 1999b):

$$X \mid A = \mu_A + U_A Z_A + E_A, \qquad \Sigma_A = U_A \Lambda_A U_A^\top + \sigma_A^2 I. \tag{7}$$

A recursive conditional atlas writes every node parameter as $\theta_v = \Psi(\theta_{\mathrm{pa}(v)}, \eta_v)$. This changes how terminal parameters are estimated while leaving terminal expressivity unaltered: for fixed leaves, the terminal law remains

$$q_T(x) = \sum_{\ell \in \mathcal{L}(T)} \pi_\ell q(x \mid \theta_\ell).$$

If successive parameter innovations have short conditional descriptions, the model admits the code

$$L(T) + L(\theta_\varnothing) + \sum_{v \neq \varnothing} L(\eta_v \mid \theta_{\mathrm{pa}(v)}). \tag{8}$$

Whether this code is short depends on the chosen learner. The PPCA demonstration below evaluates one concrete parameter-sharing mechanism by held-out predictive log loss.

B.4 A SECONDARY PREDICTIVE QUANTITY

For a learner $\mathcal{A}$, training sample S_n, and independent test pair (X', C'), define held-out predictive log loss by

$$\ell_{\mathcal{A}}(n) = \mathbb{E}[-\log_2 q_{\mathcal{A}}(X' \mid C', S_n)]. \tag{9}$$

Negative log density is strictly proper (Gneiting & Raftery, 2007), and the quantity is measured in bits. We retain every pointwise comparison over $\mathcal{N} = \{2, 4, 8, 12, 16\}$ training observations per terminal law and summarize the curve by

$$A_C = |\mathcal{N}|^{-1} \sum_{n \in \mathcal{N}} [\ell_C(n) - \ell_T(n)]. \tag{10}$$

The expected direction is exact when the stacked terminal means follow a Gaussian tree prior K_T, whose posterior predictive is q_T: for any comparator predictive q_C,

$$\mathbb{E}[-\log_2 q_C(X_{\text{new}} \mid X_{\text{train}})] - \mathbb{E}[-\log_2 q_T(X_{\text{new}} \mid X_{\text{train}})] = \mathbb{E}[\mathrm{KL}_2(q_T \| q_C)] \geq 0.$$

Appendix C derives this Gaussian identity. The implemented PPCA/Student estimator is a finite-sample demonstration of this sharing mechanism, rather than a consequence of the energy profile.

B.5 A PREDICTIVE DEMONSTRATION

Finally, the PPCA/Student demonstration evaluates whether a first-order local density learner can exploit the supplied relations. At every non-root node v, the learner shrinks empirical means and covariances toward their parent:

$$\widetilde{\mu}_v = \frac{n_v \widehat{\mu}_v + \lambda \widetilde{\mu}_{\mathrm{pa}(v)}}{n_v + \lambda}, \tag{11}$$

$$\widetilde{\Sigma}_v = \mathrm{PPCA}_r\Big[(1-\alpha)\widehat{\Sigma}_v + \alpha \widetilde{\Sigma}_{\mathrm{pa}(v)}\Big]. \tag{12}$$

The covariance estimate at an ancestor pools residuals centered at the terminal classes below it. This preserves covariance sharing without absorbing displacement between terminal means. Rank, covariance shrinkage, and the mean pseudocount are selected separately for each model family on validation data. All model families predict the same terminal labels on the same test data; the comparators use global sharing or size-matched randomizations of the supplied relation.

The experiments use five disjoint blocks and fixed 64-dimensional projections of the DINOv2-L and Qwen3-Embedding-4B representations. Matched Gaussian generators compare tangent covariances inherited from their parent with covariances independently rotated relative to the parent's orientation. Appendix D gives the implementation and model-selection details.

B.6 A PARTICULAR LEARNER CAN EXPLOIT SOME SUPPLIED RELATIONS

The preceding results use complete conditional laws and no local density model. The proof-of-concept PPCA/Student demonstration evaluates whether one learner can benefit from sharing tangent parameters along the same relations. The matched Gaussian experiment first isolates the mechanism. Recursive sharing improves prediction when terminal tangent spaces inherit their orientations across scales; global or unrelated fits are favored when those orientations are independently rotated (fig. 5).

On CIFAR-100, the recursive estimate averages 0.357 $[0.313, 0.400]$ fewer bits per observation than global sharing and 1.270 $[1.215, 1.324]$ fewer than size-matched randomized groups over $n \in \{2, 4, 8, 12, 16\}$. The advantage over global sharing approaches zero at the two largest sample sizes, while the advantage over randomized groups remains positive.

For Amazon Reviews, randomizing the fine relation raises held-out loss by 0.853 $[0.752, 0.954]$ bits per observation on average. By contrast, global sharing lowers loss by 0.174 $[0.065, 0.284]$ bits, and randomizing the coarse relation changes it by only -0.016 $[-0.038, 0.005]$ bits. Thus this particular means-and-tangents learner exploits the fine relation but not every relation detected between the complete conditional laws.

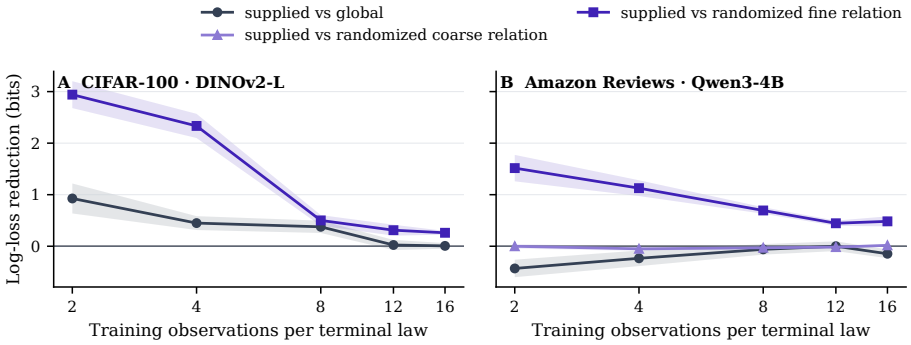


Figure 4: **Predictive value of the supplied relations for a PPCA/Student learner.** Every curve is the log-loss reduction per held-out observation of the supplied hierarchy relative to the named comparator; positive values favor the supplied hierarchy. Curves and paired 95% intervals use five disjoint, terminal-law-stratified blocks. This predictive quantity evaluates a particular sharing mechanism and is separate from the distributional profile.

The distributional profile measures organization in the complete conditional laws. PPCA can exploit only the part expressed through its local means-and-tangents parameterization. The two experiments therefore complement rather than duplicate one another. Exact learning curves are listed in appendix E; the retained result bundle contains all blockwise scores and validation-selected hyperparameters.

C Gaussian inheritance calculation

Collect the leaf parameters in $\Theta \in \mathbb{R}^{mD}$, with prior $\Theta \sim \mathcal{N}(0, K \otimes I_D)$. Suppose a training summary satisfies

$$\widehat{\Theta} \mid \Theta \sim \mathcal{N}(\Theta, V_n),$$

and a future observation from leaf ℓ satisfies

$$X_\ell \mid \Theta \sim \mathcal{N}(S_\ell \Theta, \sigma^2 I_D),$$

where S_ℓ selects the block belonging to leaf ℓ. Gaussian conditioning yields

$$\begin{aligned} M_K &= (K \otimes I_D)\,(K \otimes I_D + V_n)^{-1}\,\widehat{\Theta}, \\ C_K &= (K \otimes I_D) - (K \otimes I_D)\,(K \otimes I_D + V_n)^{-1}\,(K \otimes I_D), \\ X_\ell \mid \widehat{\Theta} &\sim \mathcal{N}\big(S_\ell M_K,\; S_\ell C_K S_\ell^\top + \sigma^2 I_D\big)\,. \end{aligned}$$

When $K = K_T$ is the generating covariance, the final distribution is the true conditional law q_T. Conditional on $\widehat{\Theta}$, the excess cross-entropy of any comparator q_C is

$$\mathbb{E}_{q_T}[-\log_2 q_C] - \mathbb{E}_{q_T}[-\log_2 q_T] = \mathrm{KL}_2(q_T \| q_C).$$

Averaging over $\widehat{\Theta}$ gives the identity used in appendix B.4. If $V_n \to 0$ at every leaf, both regular posterior predictives converge to the observation law centered at the realized Θ. Their per-observation regret consequently vanishes.

D Experimental details

D.1 Energy estimators and profile computation

For empirical laws $\widehat{P} = n^{-1}\sum_i \delta_{x_i}$ and $\widehat{Q} = m^{-1}\sum_j \delta_{y_j}$, the nonnegative estimator is

$$\widehat{\mathcal{E}}_V(\widehat{P}, \widehat{Q}) = \frac{2}{nm}\sum_{i,j}\|x_i - y_j\|_2 - \frac{1}{n^2}\sum_{i,i'}\|x_i - x_{i'}\|_2 - \frac{1}{m^2}\sum_{j,j'}\|y_j - y_{j'}\|_2.$$

This V-statistic is the energy distance between the empirical measures. The unbiased U-statistic replaces the two within-law terms by $\{n(n-1)\}^{-1}\sum_{i\neq i'}$ and $\{m(m-1)\}^{-1}\sum_{j\neq j'}$. We compute both matrices from one matrix of distances among observations. A finite-sample U-statistic may be negative for an individual pair; we do not clip it.

For child probabilities π_a and parent map pa, the implementation substitutes either distance matrix into

$$\begin{aligned}
\widehat{V} &= \frac{1}{2}\sum_{a<b}\pi_a\pi_b\widehat{\mathcal{E}}_{ab},\\
\widehat{I} &= \frac{1}{2}\sum_{v}\frac{1}{\pi_v}\sum_{a<b:\mathrm{pa}(a)=\mathrm{pa}(b)=v}\pi_a\pi_b\widehat{\mathcal{E}}_{ab},\\
\widehat{F} &= \widehat{I}/\widehat{V}, \qquad \widehat{\kappa} = 1-\widehat{F},\\
\widehat{A} &= \frac{\widehat{\kappa}-(K-1)/(m-1)}{1-(K-1)/(m-1)}, \qquad \widehat{R} = \widehat{F}\widehat{A}.
\end{aligned}$$

At each relation, the primary design gives every child law equal probability. This balancing is performed anew at every level and therefore does not induce all internal masses from one terminal measure.

Each observation block produces one U-statistic distance matrix. We average the matrices before computing $\widehat{V}$, $\widehat{I}$, their ratio, and $\widehat{R}$. For a block–child stratum h, let $\widehat{R}_{h,-i}$ be the pooled score after deleting observation i from that stratum while leaving all other strata fixed. The stratified jackknife variance is

$$\widehat{\mathrm{Var}}(\widehat{R}) = \sum_h \frac{n_h-1}{n_h}\sum_{i=1}^{n_h}(\widehat{R}_{h,-i}-\overline{R}_{h,-})^2.$$

The interval is $\widehat{R}$ plus or minus 1.96 jackknife standard errors. It describes repeated sampling of observations from the fixed child laws, not sampling new parents or domains. The zero point of the score is instead the analytical average over all size-matched parent maps.

The asymptotic regime fixes the conditional-law hierarchy and experimental design while the number of independent observations from every child law grows. If the distance kernel has finite second moment and V_j remains positive, the joint U-statistic central limit theorem applies to $(\widehat{V}_j, \widehat{I}_j)$. The delta method then gives asymptotic normality of $\widehat{R}_j$, and the stratified jackknife estimates its first-order variance.

Table 1: The estimand attached to each empirical statement. Every statement is conditional on the declared observation law and the design choices in the last column.

Statement	Observation law and metric	Weighting and hierarchy	Resampled unit / held fixed
Relation profile	Named child-conditional law; energy distance in the stated representation	Uniform child mass; metadata or deterministic WordNet parent map	Observation within block–child stratum / leaves, map, representation, and metric
Non-singleton support	Same law restricted to children whose parent has at least two represented children	Uniform mass over restricted children; same parent map	Observation within restricted strata / support and all design choices
Representation sensitivity	RGB, DINOv2-L, hashed-count, Qwen3, or spike-histogram pushforward with its stated metric	Same child weights and hierarchy within a dataset	Observation in that pushforward / representation map and hierarchy
PPCA exploitation	Terminal conditional feature law; held-out Student log loss	Supplied relation; leaf-balanced train, validation, and test samples	Paired block / learner family and candidate grid

D.2 Matched interval calibration

We calibrated the observation interval using one-dimensional, equal-variance Gaussian child laws. Five designs reproduce the child count, parent count, parent-size vector, block count, and observations per child of the finest Aircraft, SHD, ImageNet-R, Amazon Reviews, and ImageNet-1k relations. In each design, fixed child means place the population score exactly at either the random-parent or the no-innovation boundary. Over 200 samples per setting, coverage at the random-parent

boundary is 95.0–97.5%, with one-sided false-positive rates of 0.5–3.5%. At the no-innovation boundary, coverage is 98.5–99.5%, and the false-positive rate is 0–1.5%.

A third setting contracts the between-law variation toward zero. Coverage then ranges from 85.0% to 98.0%, and 0–12.5% of samples yield a nonpositive estimate of V, for which the ratio is undefined. This stress test locates the regularity limit of the procedure. The reported interval approximates observation sampling for fixed laws with positive variation; it is not uniformly valid at $V = 0$. The distributed 15-row record also reports estimator bias and standard-error comparisons.

D.3 Observation blocks and transformations

FGVC-Aircraft is divided into five blocks of 20 observations per variant, SHD into three blocks of 6 observations per speaker–word law, and ImageNet-R into three blocks of 16 observations per class. CIFAR-100 and Amazon Reviews use five blocks of 48 observations per terminal law; ImageNet-1k uses two blocks of 24. Every selected observation enters the complete-law profile. Only the cross-fitted sensitivities and downstream experiments assign training, validation, or test roles. The machine-readable analysis specification is `experiments/configs/recursive_manifold_study.json`.

For the ImageNet-R partition sensitivity, split seeds 20260805–20260816 use the same three-block design and per-leaf allocation as the primary seed 20260804. A second calculation partitions the same 48 observations per class into two, three, four, or six blocks.

D.4 Cross-fitted location sensitivity

The primary endpoint compares complete conditional laws. A separate analysis measures how much organization remains after removing child-specific location. We estimate each child mean from a disjoint reference sample, subtract it from the observations used to compute distances, and recompute the size-matched pair-dispersion statistic. Cross-fitting permits analysis of population- centered laws instead of in-sample residuals. Since centering changes the conditional laws and a Euclidean mean is not canonical in every space, this calculation is a sensitivity analysis rather than a replacement endpoint.

If R_a is the reference set for child a, the scored observation becomes

$$x_i - |R_a|^{-1} \sum_{r \in R_a} x_r.$$

This operation is absent from the complete-law profile.

D.5 Representations and hierarchy construction

The learned representations are identified by the following versions:

DINOv2-L	`47b73eefe95e8d44ec3623f8890bd894b6ea2d6c`
Qwen3-Embedding-4B	`5cf2132abc99cad020ac570b19d031efec650f2b`.

DINOv2 contributes its 1024-dimensional class token. Qwen3-Embedding-4B contributes the final non-padding token from review text truncated to 512 tokens, without a task prompt. Both features are normalized and used in full in the primary analysis. RGB observations are deterministic 32-by-32 bilinear resizes in $[0, 1]$.

Because DINOv2 was pretrained on a large curated image collection, its profile may express object-semantic regularities learned during pretraining as well as structure in the target observations (Oquab et al., 2024). The RGB rows measure distinct pushforward laws; they do not make the DINO result representation-free.

In ImageNet-1k and ImageNet-R, every class identifier is a WordNet noun synset. When several hypernym paths exist, we choose the longest and resolve ties lexicographically. Ancestors four and two edges above each leaf define the coarse and middle levels. This deterministic two-edge spacing creates two successive relations without asserting a uniquely preferred semantic depth.

For ImageNet-1k, we use all 1,000 classes in the released validation split (Deng et al., 2009). Each class contains 50 images. Before reading image contents, we retain the 48 files with the smallest stable hashes of their official validation filenames. The resulting 48,000-image construction has 118 coarse and 308 middle nodes. Of these, 55 coarse and 153 middle nodes have at least two represented children, retaining 245 middle and 845 leaf laws in the corresponding non-singleton analyses.

Amazon Reviews 2023 supplies product-category paths and review text (Hou et al., 2026). Within Electronics, we take category prefixes at depths two, three, and four after removing the common root. Terminal categories must contain at least 1,000 products and participate in both non-singleton relations. Stable hashing selects one review from each of 240 distinct products per terminal category. This yields 27,120 reviews in 113 terminal, 30 middle, and 8 coarse categories. The Qwen representation is paired with a fixed 64-bin hashed-count representation of the same texts (Zhang et al., 2025).

For ImageNet-R, the resulting hierarchy has 56 coarse nodes, 113 middle nodes, and 200 classes. We repeat the DINO analysis under twelve split seeds and under two, four, and six blocks while holding the budget at 48 observations per class.

FGVC-Aircraft uses the released manufacturer, family, and variant text files. Code checks verify that each variant maps to exactly one family and one manufacturer. SHD terminal identifiers combine the released speaker identity with the word, making the nesting exact. Events are binned into 32 time bins and 35 contiguous cochlear-channel bins. The untransformed relative-frequency law tests sensitivity to the square-root transform used by the primary law.

Table 2: Provenance of the input data. SHA-256 prefixes identify the local archives; the reproducibility record contains the complete hashes.

Dataset	Source artifact	SHA-256 prefix
ImageNet-1k validation	Official `ILSVRC2012_img_val.tar`	`c7e06a6c0bac...`
ImageNet-1k devkit	Official `ILSVRC2012_devkit_t12.tar.gz`	`b59243268c0d...`
Amazon Reviews Electronics	Official `Electronics.jsonl.gz`	`17b2c5f3736d...`
FGVC-Aircraft 2013b	Official `fgvc-aircraft-2013b.tar.gz`	`e4e323d410e2...`
ImageNet-R	Official `imagenet-r.tar`	`18c6bf493b39...`
SHD train	Official `shd_train.h5`	`2bddb4bd4673...`
SHD test	Official `shd_test.h5`	`47d8621a092c...`

D.6 Local-law implementation

Given a sample covariance with ordered eigenpairs $(\widehat{\lambda}_i, \widehat{u}_i)$, the rank-$r$ PPCA projection retains the leading r directions and replaces the remaining eigenvalues by their floored mean. Ancestor covariances are estimated from terminal-leaf-centered residuals. Before fitting, each coordinate undergoes the robust transformation

$$z_j = \operatorname{asinh}\left(\frac{x_j - \operatorname{median}(x_j^{\mathrm{train}})}{s_j}\right),$$

where s_j is estimated from training data and floored at 0.1 times the median scale of the nonconstant coordinates. Test observations enter neither this preprocessing nor hyperparameter selection.

The candidate grid combines ranks $\{4, 8\}$, covariance-inheritance weights $\{0, 0.5, 0.75, 1\}$, mean pseudocounts $\{0, 4, 8\}$, and a Student distribution with four degrees of freedom. For each observation block and training size, every model family is selected independently on eight validation observations per terminal law and evaluated on 24 test observations. The five fixed relation randomizations are averaged within each block. The complete configuration is stored in `experiments/configs/ppca_supplied_relations.json`. These experiments use fixed 64-dimensional projections of the full DINOv2-L and Qwen3-Embedding-4B features; the primary distributional analysis uses the full features. Relation randomization enters only this predictive demonstration, not the model-free hierarchy profile.

D.7 COMPUTE AND PUBLIC ARTIFACTS

Feature extraction used one CUDA GPU; the energy profiles and PPCA fits ran on CPU. The public project includes every blockwise and aggregate model-free row, the non-singleton profiles, ImageNet-R partition sensitivity, source hashes, the analysis configuration, and the scripts used to regenerate every figure. Unit tests cover the energy estimators, weighted identities, degenerate boundaries, observation jackknife, randomization, cross-fitted centering, parent maps, hierarchy parsing, split disjointness, PPCA scoring, and structured-innovation identities.

The accompanying provenance record gives source URLs, complete archive hashes, and usage terms. ImageNet and Aircraft images are offered for non-commercial research, SHD is licensed under CC BY 4.0, and rights to ImageNet-R images remain with their upstream sources. We redistribute neither raw observations nor learned features for individual observations. All quantitative paper figures rebuild from checksum-protected aggregate tables; the hierarchy illustration is generated from the licensed source data. The project code is MIT-licensed.

E SUPPLEMENTARY RESULTS

E.1 MULTISCALE DOMAIN PROFILES

Table 3: Conditional-refinement profiles under the child-balanced law. V denotes absolute law variation in the stated representation, F the innovation fraction, A adjusted parent organization, and R the pooled signed criterion with its 95% observation-jackknife interval.

Distribution	Child / parent	V	F	A	structured innovation R
Aircraft RGB	family / maker	.314	.255	.561	.1428 [.1388,.1467]
Aircraft RGB	variant / family	.260	.038	.874	.0334 [.0155,.0512]
SHD Hellinger	word / language	.065	.984	-.039	-.0383 [-.0410,-.0357]
SHD Hellinger	speaker / word	.182	.633	.312	.1976 [.1953,.1999]
ImageNet-R DINO	middle / coarse	.106	.447	.121	.0540 [.0477,.0603]
ImageNet-R DINO	class / middle	.135	.368	.159	.0584 [.0541,.0626]

E.2 IMAGE AND TEXT PROFILES

Table 4: Conditional-refinement profiles for CIFAR-100, ImageNet-1k, and Amazon Reviews. The columns follow table 3.

Distribution	Child / parent	V	F	A	R
CIFAR-100 pixels	class / superclass	.078	.529	.346	.1828 [.1773,.1883]
CIFAR-100 DINOv2-L	class / superclass	.123	.670	.171	.1144 [.1132,.1155]
ImageNet-1k DINOv2-L	middle / coarse	.141	.566	.086	.0486 [.0453,.0519]
ImageNet-1k DINOv2-L	class / middle	.203	.606	.124	.0755 [.0739,.0770]
Amazon counts	middle / coarse	.0032	.690	.090	.0623 [.0374,.0872]
Amazon counts	category / middle	.0051	.502	.323	.1621 [.1529,.1713]
Amazon Qwen3-4B	middle / coarse	.021	.524	.309	.1621 [.1595,.1646]
Amazon Qwen3-4B	category / middle	.030	.412	.444	.1830 [.1823,.1837]

E.3 NON-SINGLETON-PARENT SUPPORT

A parent represented by one child contributes no within-parent pair. We therefore repeat the analysis using only children whose parent has at least two represented children. Across all 14 relation–law pairs in fig. 3, the restriction preserves every conclusion under the $R_j > 0$ sufficiency condition. Complete profiles are listed in table 5.

Table 5: Conditional-refinement profiles restricted to children whose parent has at least two represented children. The counts give the restricted and complete collection sizes; R includes its 95% observation-jackknife interval.

Distribution	Child / parent	parents kept	children kept	V	F	A	R
CIFAR pixels	class / superclass	20/20	100/100	.078	.529	.346	.1828 [.1773,.1883]
CIFAR DINOv2-L	class / superclass	20/20	100/100	.123	.670	.171	.1144 [.1132,.1155]
ImageNet-1k	middle / coarse	55/118	245/308	.137	.736	.055	.0408 [.0388,.0428]
ImageNet-1k	class / middle	153/308	845/1000	.204	.715	.128	.0916 [.0905,.0927]
Amazon counts	middle / coarse	8/8	30/30	.0032	.690	.090	.0623 [.0374,.0872]
Amazon counts	category / middle	30/30	113/113	.0051	.502	.323	.1621 [.1529,.1713]
Amazon Qwen3-4B	middle / coarse	8/8	30/30	.021	.524	.309	.1621 [.1595,.1646]
Amazon Qwen3-4B	category / middle	30/30	113/113	.030	.412	.444	.1830 [.1823,.1837]
Aircraft RGB	family / maker	16/30	56/70	.228	.440	.396	.1739 [.1640,.1838]
Aircraft RGB	variant / family	16/70	46/100	.081	.266	.601	.1599 [.1334,.1863]
SHD Hellinger	word / language	2/2	20/20	.065	.984	-.039	-.0383 [-.0410,-.0357]
SHD Hellinger	speaker / word	20/20	240/240	.182	.633	.312	.1976 [.1953,.1999]
ImageNet-R DINO	middle / coarse	25/56	82/113	.103	.634	.098	.0625 [.0583,.0666]
ImageNet-R DINO	class / middle	36/113	123/200	.140	.575	.194	.1116 [.1084,.1147]

For ImageNet-R, the restricted scores remain positive at both scales:

$$R_2 = 0.0625\ [0.0583, 0.0666], \qquad R_3 = 0.1116\ [0.1084, 0.1147].$$

These targets contain 82 of 113 child laws under 25 coarse parents and 123 of 200 child laws under 36 middle parents. ImageNet-1k, Amazon Reviews, and Aircraft likewise remain positive at both scales, while the SHD language-to-word relation remains below the random-parent expectation. Thus the structured-refinement results are not produced by singleton branches; as elsewhere, they describe the represented collection rather than every parent individually.

E.4 Representation profiles

Table 6: Profiles for the paired representations. Each row is a distinct pushforward law, and the columns follow table 3.

Distribution	Child / parent	V	F	A	R
Aircraft DINO	family / maker	.214	.335	.421	.1414 [.1407,.1420]
Aircraft DINO	variant / family	.193	.0074	.976	.0072 [.0038,.0106]
SHD frequencies	word / language	.0073	.983	-.037	-.0365 [-.0393,-.0338]
SHD frequencies	speaker / word	.023	.667	.275	.1836 [.1809,.1863]
ImageNet-R RGB	middle / coarse	.258	.414	.186	.0770 [.0140,.1400]
ImageNet-R RGB	class / middle	.278	.262	.402	.1051 [.0898,.1204]

E.5 ImageNet-R partition sensitivity

The finding is insensitive to the computational partition. Across twelve split seeds, the two scores range from 0.0489 to 0.0552 and from 0.0564 to 0.0623. Every interval satisfies the sufficient condition from definition 1. Dividing the same 48 images per class into two, three, four, or six blocks yields the narrower ranges 0.0538–0.0548 and 0.0584–0.0589.

E.6 Cross-fitted location sensitivity

Cross-fitted location removal separates contributions beyond the child means. Parent organization remains detectable in both CIFAR representations, at the fine ImageNet-1k relation, and at the coarse Amazon relation in both representations. The other intervals include zero. The full results are listed in table 7.

Table 7: Relative energy advantage of the declared parent after subtracting child means estimated from disjoint reference observations. These descriptive values use the size-matched pair-dispersion statistic rather than the primary complete-law profile.

Distribution	Child / parent	advantage [95% interval]
CIFAR-100 pixels	class / superclass	2.54 [0.64,4.45]%
CIFAR-100 DINOv2-L	class / superclass	1.54 [1.22,1.86]%
ImageNet-1k DINOv2-L	middle / coarse	7.27 [-0.72,15.27]%
ImageNet-1k DINOv2-L	class / middle	3.37 [1.81,4.93]%
Amazon counts	middle / coarse	9.78 [6.95,12.61]%
Amazon counts	category / middle	0.41 [-1.86,2.69]%
Amazon Qwen3-4B	middle / coarse	9.89 [8.51,11.27]%
Amazon Qwen3-4B	category / middle	-0.45 [-1.82,0.92]%

E.7 Matched-generator learner check

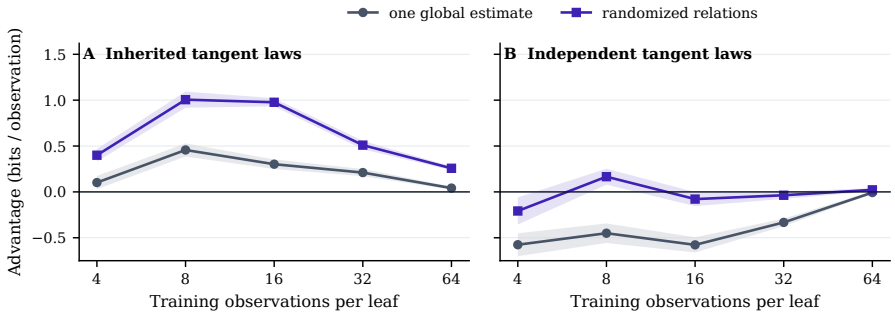


Figure 5: **Mechanism check with learned terminal partitions.** Positive values indicate lower held-out log loss for the recursive learner. Curves average 40 generators, with paired 95% intervals. The two generators differ only in whether terminal tangent laws inherit their orientation across scales.

When tangent laws are related, the recursive model outperforms both global sharing and relation randomization throughout the tested range. Its advantage peaks in the intermediate low-sample regime and diminishes as each terminal law becomes estimable. Independently rotated tangent laws instead favor global sharing, while their randomization contrast fluctuates around zero. Ambient dimensions 16 and 32 show the same sign pattern.

The level-specific experiment holds all eight terminal Gaussians fixed. Under the related generator, randomizing the coarser scale adds 0.045–0.122 bits per observation, while randomizing the finer scale adds 0.344–1.224 bits across 4–32 observations per terminal law. Every corresponding interval is positive. With 64 observations, the coarser contribution approaches zero and the finer contribution persists. Independently rotated terminal laws exhibit no coherent effect across regimes.

E.8 Supplied-relation predictive effects

Table 8: Held-out log-loss difference between each comparator and the recursive model, in bits per observation, over five disjoint blocks. Positive values favor the recursive model.

Dataset	Comparator	$n = 2$	$n = 4$	$n = 8$	$n = 12$	$n = 16$
CIFAR-100	global	0.927	0.449	0.377	0.024	0.007
	randomized coarse groups	2.941	2.334	0.501	0.311	0.261
Amazon Reviews	global	-0.431	-0.234	-0.061	0.000	-0.146
	randomized coarse relation	-0.001	-0.051	-0.030	-0.018	0.017
	randomized fine relation	1.516	1.127	0.694	0.446	0.483

For CIFAR-100, every randomized-relation interval lies above zero; the intervals against global sharing include zero at $n = 12$ and $n = 16$. For Amazon Reviews, every fine-relation randomization interval lies above zero. The global comparison favors global sharing at $n = 2$, $n = 4$, and $n = 16$; its other two intervals include zero. The coarse randomization intervals lie at or below zero. The checksum-protected result bundle contains exact endpoints, per-observation scores, and the validation-selected model for every block.