Zentropy Semantic Configuration Fields: A Cross-Modal Framework for Analyzing Semantic Organization in Learned Representations

Anonymous authors
Paper under double-blind review

Abstract

Representation analysis largely characterizes learned embeddings through feature similarity, linear separability, or pairwise geometry, leaving local semantic organization, i.e., how labels are composed around a point and how that composition changes with depth and neighborhood scale, without a unified measurement language. We introduce **Zentropy Semantic Configuration Fields (SCF)**: a smoothed local class-posterior entropy, resolved as a field over network depth l and neighborhood granularity τ, $\mathcal{S} = \{Z(l, \tau)\}_{l,\tau}$. The statistic itself is a Nadaraya–Watson label posterior followed by Shannon entropy; the contribution is the field, its descriptor calculus, and a permutation-null calibration. Controlled composition experiments show the estimator closely tracks true local semantic entropy at fixed geometry ($R^2 = 0.977$, Pearson $r = 0.994$), and four-quadrant experiments show it is complementary to distance rather than a reparameterization. Instantiating SCF across five modalities including vision, audio, time series, text, physiological signals reveals structured layer-by-granularity organization that often continues to evolve after linear probing has largely saturated, predominantly at fine granularities. A morphology benchmark shows that systems occupy clearly distinct field morphologies, with descriptors that are non-redundant and largely stable under seed, layer, bandwidth, and metric perturbations. A permutation-null calibration separates label-driven organization from a systematic label-free concentration drift, and reveals substantially stronger calibrated organization under label smoothing in a matched training comparison; in low-drift domains the observed organization is entirely label-driven. SCF thus provides a common analytical language for comparing how representation-learning systems organize semantic information.

1 Introduction

Two trained networks can be indistinguishable to a linear probe and yet organize their local neighborhoods in profoundly different ways: one may keep class boundaries sharp at fine scales deep into the network, while another mixes them early and never re-separates them. Standard representation evaluation, downstream accuracy, linear probes, cluster quality, or representation similarity measures such as CKA (Kornblith et al., 2019), compresses the embedding into a scalar or a pairwise relation and cannot see this difference. What it does not express is *how semantics are locally organized*: given a point in representation space, what label composition surrounds it, and how that composition evolves across network depth and across the scale at which "local" is defined.

This paper introduces the **Semantic Configuration Field (SCF)**. The construction has three levels. First, *local semantic composition*: at a point z_i, estimate the label posterior $\hat{p}_i(c)$ of its neighborhood and measure its entropy Z_i. Second, the *multi-granularity signature* $\Gamma_l = [\tilde{Z}_l(\tau_1), \ldots, \tilde{Z}_l(\tau_G)]$: how local semantics resolve as the neighborhood widens. Third,

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.

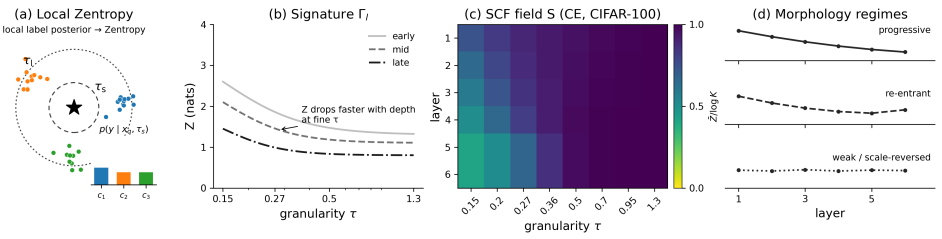

Figure 1: The SCF construction. (a) Kernel-weighted local label posterior → per-point Zentropy. (b) Sweeping granularity τ gives a signature Γ_l per layer. (c) Sweeping layers gives the field $\mathcal{S}$ (shown: CIFAR-100 CE). (d) Morphology descriptors read out the field's organization regime.

the *field* $\mathcal{S} = [Z(l, \tau)]_{L \times G}$: semantic organization as a function of depth l and granularity τ. Figure 1 summarizes the construction. SCF is post-hoc and model-agnostic: any $f_\theta : X \to Z$ produces a field, and the framework trains nothing. The novelty of SCF does not lie in redefining Shannon entropy, but in treating kernel-estimated local semantic entropy as a depth–granularity field with a morphology calculus and a null calibration, which together expose what neither component expresses alone.

Our claims are threefold. (1) **A semantic configuration field**: SCF extends local semantic entropy into a structured field over depth and neighborhood scale, with a mandatory estimator-validity protocol and formal foundations: boundedness and purity ordering, similarity equivariance, and asymptotic consistency with explicit smoothing-bias control (Sec. 4, App. A). (2) **A morphology language**: a compact descriptor calculus (depth organization slope, cross-granularity trend, reconfiguration depth, field curvature, organization magnitude, and deep-level reversal) that quantifies how semantic organization is shaped, not merely whether it exists. (3) **A cross-modal analysis benchmark**: SCF instantiates identically across heterogeneous modalities while producing systematically different morphologies across representation-learning systems.

The evidence is direct: controlled experiments show the estimator tracks local semantic entropy with $R^2 = 0.977$ at fixed geometry; five modalities reveal structured layer–granularity organization; and a four-seed, null-calibrated morphology benchmark separates systems by how much, when, and at which granularity they organize semantics.

2 Related Work

Representation analysis and similarity. The standard toolkit asks what can be *read out* of a representation (linear probes) and how two representations *relate* (CKA and predecessors; Kornblith et al., 2019; Klabunde et al., 2025a), with documented caveats (Davari et al., 2023; Alvarez, 2023). Neural collapse (Papyan et al., 2020; Tan et al., 2026) characterizes *global* class geometry, while recent work probes convex decision regions (Tětková et al., 2025) and local intrinsic dimension (Ruppik et al., 2025).

Geometry and topology. Geo-topological summaries track layerwise structure without labels (Lin & Kriegeskorte, 2024; Ballester et al., 2024; 2023), topology divergence and symmetric-divergence frameworks compare representations across layers and models (Barannikov et al., 2022; Wang & Hu, 2026), cover learning scales topology representation (Scoccola et al., 2025), and topology can serve as a training objective (Trofimov et al., 2023). Closest in spirit, topological alignment spectra separate fine-scale geometric variation from broader semantic reorganization (Tavares et al., 2026), and mechanistic case studies map representation manifolds in controlled settings (Moisescu-Pareja et al., 2026). All of these views are scalar, pairwise, or global-geometric. SCF addresses a complementary object: how labels are *composed* around each individual point, resolved across depth l and granularity τ. A layer can be probe-saturated yet keep reorganizing its local label composition; SCF exposes precisely this regime (Sec. 6).

Local uncertainty and neighborhood statistics. The classical kNN rule (Cover & Hart, 1967) interprets the label distribution of a local neighborhood; the Kraskov estimator (Kraskov et al., 2004) estimates mutual information from kNN distances, a population-level quantity, whereas Zentropy is a kernel-smoothed estimate of local label composition and is never identified with population mutual information. Label-T&C (Jeon et al., 2024) and ReSi (Klabunde et al., 2025b) compress local label agreement into scalar summaries, neighbor-embedding reliability is analyzed through map continuity (Liu et al., 2025), and local entropy of feature maps has been used for architecture selection (Wu et al., 2026). Post-hoc probes require falsification-style validation (Ma et al., 2025), and the manifold assumption itself can fail (Robinson et al., 2025) — all reasons SCF pairs its estimates with a permutation-null calibration rather than reporting raw fields. Semantic-entropy methods (Kuhn et al., 2023) measure ambiguity over generated outputs, a different object from neighborhood label composition in representation space. SCF instead returns a *field* $\mathcal{S} = [\tilde{Z}(l, \tau)]_{L\times G}$ with a morphology calculus: where scalar metrics rank embeddings, SCF characterizes them.

Multi-scale entropy and the Zentropy lineage. SCF's entropy descends from Shannon's measure (Shannon, 1948); the multi-granular reading of entropy comes from statistical physics (Liu et al., 2019; 2022). In machine learning, this lineage is carried by the granular-computing feature-selection series of Yuan and collaborators: entropy evaluated at a *single granular level* is an incomplete description of uncertainty, and integrating across granular levels yields more comprehensive and stable measures (Yuan et al., 2024; 2025; 2026a;b). The thermodynamics-grounded ZENN framework (Wang et al., 2026) shows the same principle improves discriminative objectives. All of these works use Zentropy for feature selection or classification; we bring the multi-granular principle into *representation analysis*: the granularity axis τ plays the role of coarse-graining scale, the objects being coarse-grained are neighborhoods of learned embeddings, and the goal is measurement and characterization rather than selection. To our knowledge, SCF is the first use of the multi-granular Zentropy principle as a layer-by-granularity field for learned representation analysis.

Benchmark objects. SCF trains nothing; it treats representation-learning systems as objects of study (see Gui et al., 2024 for a survey): contrastive instance discrimination (SimCLR; Chen et al., 2020), which is provably close to supervised contrastive learning in the many-class regime (Luthra et al., 2025), self-distillation (DINO; Caron et al., 2021), variance–invariance–covariance regularization (VICReg; Bardes et al., 2022), supervised contrastive learning (SupCon; Khosla et al., 2020), and cross-entropy training, under one identical protocol. Domain benchmarks such as the UCR archive (Dau et al., 2019) and Sleep-EDF Expanded (Kemp et al., 2013; 2000; Goldberger et al., 2000) (whose stage labels carry well-documented inter-scorer disagreement; Rosenberg & Van Hout, 2013) motivate why local semantic ambiguity is worth measuring rather than assumed away.

3 Semantic Configuration Fields

Zentropy here names the *multi-granular principle* of the lineage reviewed above; SCF is its realization for representation analysis.

3.1 Local semantic composition

Let $\mathcal{R} = \{(z_j, y_j)\}_{j=1}^{N}$ be a labeled reference set and let d be a metric on the representation space. Distances are normalized by the reference median squared distance, $\tilde{d}_{ij}^2 = d_{ij}^2 / \operatorname{med}(d^2)$, so bandwidths are dimensionless. For a query point z_i at granularity τ,

$$\hat{p}_i^{(\tau)}(c) = \frac{\sum_j w_{ij}^{(\tau)}\, \mathbf{1}[y_j = c] + \alpha}{1 + K\alpha}, \qquad w_{ij}^{(\tau)} = \frac{\exp(-\tilde{d}_{ij}^2/\tau^2)}{\sum_{j'} \exp(-\tilde{d}_{ij'}^2/\tau^2)}, \tag{1}$$

with Dirichlet smoothing α and leave-self-out normalization (queries never contribute to their own neighborhood; all benchmark results use held-out query sets, App. B). The per-point Zentropy is

$$Z_i^{(\tau)} = -\textstyle\sum_c \hat{p}_i^{(\tau)}(c) \log \hat{p}_i^{(\tau)}(c). \tag{2}$$

With smoothing, the attainable minimum is $Z_{\min}(K,\alpha) > 0$; we report the normalized form $\tilde{Z} = (Z - Z_{\min})/(\log K - Z_{\min})$, for which $\tilde{Z} = 0$ corresponds to the Dirichlet-floor posterior, attained when all positively weighted reference mass belongs to a single class (operationally, a locally label-pure neighborhood), and $\tilde{Z} = 1$ iff uniform.

3.2 Multi-granularity signature and the field

For layer l with representation $z_i^l = f_l(x_i)$,

$$\Gamma_l = [\tilde{Z}_l(\tau_1), \ldots, \tilde{Z}_l(\tau_G)], \qquad \mathcal{S} = [\tilde{Z}(l,\tau)]_{L\times G}. \tag{3}$$

Validity protocol (mandatory). With the effective neighbor count $N_{\text{eff},i} = (\sum_j w_{ij}^2)^{-1}$, a granularity is valid only if the *mean* N_{eff} over queries exceeds $n_{\min} = 10$; granularities failing this are excluded, ruling out 1-NN degeneration. The gate is a mean criterion, so we also audit the lower tail (App. B): within the valid band, the minimum mean N_{eff} over all systems, layers, and seeds is 50.2, five times the gate. All benchmark results use held-out test queries against a train reference; recomputing the field with in-sample leave-self-out queries leaves the depth profile unchanged (Spearman 1.00, max $|\Delta\tilde{Z}| = 0.08$ nats; App. B), so no result is driven by query construction.

3.3 Morphology descriptors

The field is analyzed by a small calculus, computed *per seed* and then aggregated (complete operational definitions, tie-breaking rules, and pseudocode in App. B): **depth organization slope** $M = \rho_s(l, \tilde{Z}(l,\tau_h))$, with τ_h the finest granularity passing the validity gate everywhere; **cross-granularity trend** $D_{\text{scale}} = \rho_s(l, \tilde{Z}(l,\tau_h) - \tilde{Z}(l,\tau_c))$, τ_c the coarsest valid granularity; **reconfiguration depth** $L^* = \arg\max_l |\Delta_l \tilde{Z}|/(L-1)$; **field curvature** κ; **organization magnitude** ΔZ_{total}; the categorical **deep-level reversal** R_{deep}; and **seed consistency** R_{SCF} against layer-permuted references.

4 Theoretical Properties

We maintain a strict estimator-level discipline: $\hat{p}_i^{(\tau)}$ and $Z_i^{(\tau)}$ are kernel-estimated quantities; population quantities (H, I) appear only inside asymptotic statements. Three properties hold at every sample size (proofs in App. A).

Theorem 1 (Dirichlet floor)**.** *With smoothing $\alpha > 0$ and K classes, the minimum attainable per-point Zentropy is $Z_{\min}(K,\alpha) = -\frac{1+\alpha}{1+K\alpha}\log\frac{1+\alpha}{1+K\alpha} - (K-1)\frac{\alpha}{1+K\alpha}\log\frac{\alpha}{1+K\alpha} > 0$, attained exactly at label-pure neighborhoods.*

Theorem 2 (Schur ordering)**.** *If p majorizes q ($p \succ q$), then $Z(p) \le Z(q)$, with equality iff q is a permutation of p: Zentropy respects the canonical purity ordering of neighborhood compositions.*

Theorem 3 (Similarity equivariance)**.** *For $z' = cAz + b$ with $A^\top A = I$ and $c > 0$ (rotation, isotropic scaling, translation), the median-normalization protocol leaves all normalized distances, and hence the entire field $\mathcal{S}$, invariant.*

Theorem 1 justifies the normalized $\tilde{Z}$; Theorem 2 makes Zentropy a purity *ordering* rather than a purity statistic; Theorem 3 makes fields exactly comparable under similarity transformations, while general reparameterizations (whitening, cosine geometry) are audited empirically (Sec. 8, App. C). Asymptotically (App. A), under standard kernel-regression conditions with a vanishing smoothing sequence $\alpha_N \to 0$, the kernel posterior is consistent and the corpus-level Zentropy converges to the population conditional entropy at rate $O_p\big(K\log(1/\alpha_N)(N\tau_N^{d_z})^{-1/2} + K\log(1/\alpha_N)\tau_N^r + \alpha_N\big)$. Consistency requires $\alpha_N \to 0$: at fixed $\alpha > 0$ the posterior converges to $(\eta_c(z) + \alpha)/(1 + K\alpha) \neq \eta_c(z)$, so all experiments treat α as a fixed regularization constant ($\alpha = 0.05$ in the synthetic and physiological experiments; $\alpha = 0.25/K$ in the CIFAR-100/Pets benchmark, i.e., smoothing mass proportional to class count) and the asymptotic statements concern the vanishing-α_N family. The bias–variance

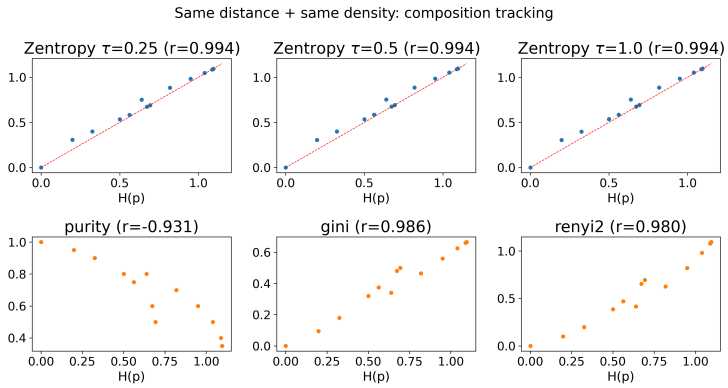

Figure 2: Composition tracking at fixed geometry: Zentropy vs true neighborhood entropy $H(p)$ (left) and alternative statistics (right); kNN distance held constant.

Table 1: Controlled composition contrasts. Same distance: differing compositions on identical rings, kNN distance unchanged. Same purity: identical max-purity, differing minority-class split.

	Same distance	Same purity
Config A	99/1	$(0.9, 0.1, 0)$
Config B	50/50	$(0.9, 0.05, 0.05)$
Z (A, B)	0.304, 0.693	0.398, 0.582
kNN distance	$p = 1.0$	n/a

decomposition is the formal basis of the multi-granularity design: no single bandwidth is uniformly correct, so the signature Γ_l is the reported object, and any single-τ choice is estimator-driven under the N_{eff} gate rather than learned against a task loss.

5 Semantic Validity of Zentropy

Before the field can serve as a measurement instrument, the per-point estimate must itself be validated: does it track true local label entropy at fixed geometry, and does it carry information beyond distance and majority purity?

Composition tracking at fixed geometry. On a ring of m=20 neighbors with fixed radius (identical kNN distance and density) we vary the label composition p over 14 levels. Zentropy tracks the true entropy $H(p)$ almost exactly (Fig. 2): Pearson $r = 0.994$, $R^2 = 0.977$, MAE $= 0.033$ nats, while kNN distance is constant. Gini ($r = 0.986$) and Rényi-2 ($r = 0.980$) track strongly but with larger error; max-purity is anti-correlated by construction ($r = -0.931$).

Geometry-independent semantic sensitivity. Two controls isolate what Zentropy responds to (Table 1). Label compositions 99/1 vs 50/50 on identical rings leave kNN distance statistically unchanged ($p = 1.0$) while Zentropy separates them decisively (0.304 vs 0.693). Purity-controlled variants $(0.9, 0.1, 0)$ vs $(0.9, 0.05, 0.05)$ are purity-identical yet Zentropy-distinct (0.398 vs 0.582). Zentropy therefore measures the full multiclass composition of a neighborhood, not proximity and not majority purity.

Multi-granularity semantic structure. Distance and Zentropy are complementary axes, and their product induces a four-quadrant structure over a reference set (Fig. 3): the *core* (near and label-pure), the *boundary* (mixed-composition neighborhoods among known classes), *coherent-novel* points (far from the reference yet locally pure, a new well-formed cluster), and *ambiguous-novel* points (high-Z mixtures). In the controlled quadrant experiment (N=2000 per zone, 3 seeds), the zone means confirm that the two axes carry

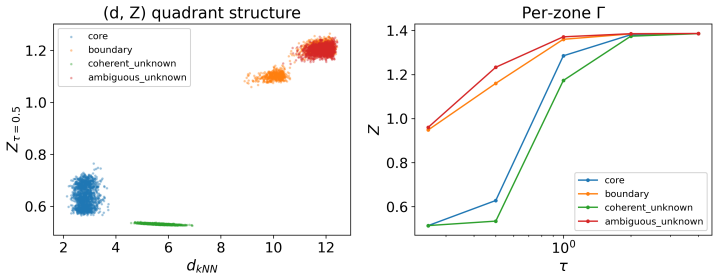

Figure 3: Four-quadrant structure in the (d, Z) plane: core, boundary, coherent-novel (far yet pure), ambiguous-novel (mixed). A classifier on (d, Γ) separates the zones at 0.913 [0.910, 0.916] vs 0.884 [0.880, 0.888] (Γ only) and 0.879 [0.875, 0.883] (distance only); mean and 95% bootstrap CI, B=1000.

Table 2: Cross-modal SCF instantiation.

Modality	Backbone	Layers	Notable morphology
Vision (CIFAR-100)	ResNet/ViT/SSL	6–8	progressive, fine-τ led
Audio (ESC-50)	4-block CNN	4	fine-τ drops, coarse-τ rises
Time series (UCR×3)	1D-CNN	3	late org. (conv2→3)
Text (AG News)	BiLSTM/Transformer	3–5	monotonic, fine-τ carried
Physiol. (Sleep-EDF Exp.)	CNN/LSTM/Trans./SSL	4–6	arch.-specific trajectories

independent information: the coherent-novel zone is distant but sits near the Dirichlet floor ($Z = 0.534$), while the ambiguous-novel zone attains the highest Zentropy ($Z = 1.233$). No scalar novelty score reproduces this structure: a logistic classifier on the joint (d, Γ) representation separates the four zones at 0.913 [0.910, 0.916], above Γ alone (0.884 [0.880, 0.888]) and distance alone (0.879 [0.875, 0.883]); the paired improvement over Γ alone is +0.029 [0.026, 0.032], with the 95% bootstrap interval entirely above zero (B=1000). Each axis misses configurations the other resolves. A manifold-traversal view of the same complementarity is given in App. C.

6 Cross-Modal Representation Analysis

Universal measurement, diverse morphology. SCF instantiates identically across five modalities (Table 2): the protocol is unchanged, yet each domain exhibits its own layer–granularity organization. These are selected exemplar systems per modality, demonstrating that the instrument ports across domains, not a cross-modal law. The recurring observation: semantic organization, carried predominantly by the finest granularities, often continues to evolve after linear probing has largely saturated.

Vision (CIFAR-100). Three encoders (supervised ResNet-18 (He et al., 2016), supervised small ViT (Dosovitskiy et al., 2021), label-free SimCLR (Chen et al., 2020) with a frozen probe), profiled under two nested reference labelings, all reproduce the layer-by-granularity phenomenon: first-to-last Zentropy drops of 0.414, 0.607, and 0.163 nats on fine labels, carried by the finest granularity in every architecture. Reconfiguration depth is architecture-specific (ResNets organize late, half-drop depths 0.63–0.67; the ViT organizes early, 0.12). A nested fine/coarse analysis (App. C) reveals only weak, architecture-dependent relative shifts, not a universal hierarchy.

Audio (ESC-50; Piczak, 2015). A compact 4-block CNN exhibits a *granularity reversal*: fine-grained Zentropy drops with depth (3.405 → 2.548 nats at τ=0.25) while coarse granularities *rise* (τ=1: 3.782 → 3.880), an opposition that a single-bandwidth summary would average away.

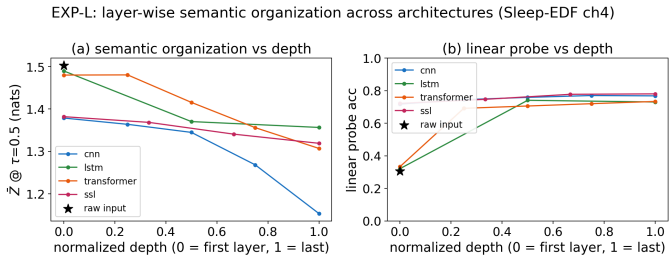

Figure 4: Layer–granularity organization across architectures (Sleep-EDF Expanded): probe accuracy saturates early while Zentropy keeps decreasing through the final layer.

Time series (UCR; Dau et al., 2019). On GunPoint, ECG200, and Wafer, organization concentrates *late* (largest drop at conv2→conv3; Zentropy half-drop depths 0.70–0.72 vs probe half-rise 0.06–0.29).

Text (AG News; Zhang et al., 2015). A BiLSTM (Hochreiter & Schmidhuber, 1997) and a from-scratch Transformer (Vaswani et al., 2017) both show clean monotonic organization carried entirely by the finest granularity (total drops 0.557 and 0.509 nats).

Physiology (Sleep-EDF Expanded; Kemp et al., 2013; 2000; Goldberger et al., 2000). Sleep staging is a natural stress test: it has attracted foundation-model efforts (Thapa et al., 2026), and its stage labels carry documented inter-scorer disagreement (Rosenberg & Van Hout, 2013; Liu et al., 2024). Four encoders (supervised CNN, BiLSTM, patch Transformer, SimCLR-SSL) exhibit distinct trajectories (Fig. 4). The CNN organizes late and keeps organizing after the probe saturates (half-drop depth 0.76 vs probe half-rise 0.23); the SSL encoder matches the supervised probe (0.780) while retaining a less organized final field (Z=1.319 vs 1.153 nats): linear readout recovers the label, but the neighborhood composition differs. Sleep-EDF Expanded provides a single nominal label per epoch, so SCF measures estimated local label entropy, not inter-rater uncertainty itself.

Fig. 4 makes the point precisely: probe accuracy saturates early, but Zentropy keeps decreasing through the final layer. Which granularity carries the change, and at which depth, is a fingerprint of the modality and architecture, not of the measurement.

7 SCF Morphology Benchmark

This section presents the paper's main empirical result: under one fixed measurement protocol, different representation-learning systems occupy distinct and reproducible field morphologies.

7.1 Benchmark design

The benchmark fixes the measurement and varies the learning system. Six representation-learning systems are trained on CIFAR-100 (Krizhevsky, 2009) with a common ResNet-18 backbone (He et al., 2016): cross-entropy (CE), CE with label smoothing (CE-LS, 0.1), SupCon (Khosla et al., 2020), SimCLR (Chen et al., 2020), a DINO-lite self-distillation variant (Caron et al., 2021), and VICReg (Bardes et al., 2022), **each with four seeds**. Because the original CE reference used a shorter schedule and standard augmentation, we additionally train **CE***: plain CE under the exact 35-epoch, strong-augmentation protocol of the other systems (four seeds), making the CE*/CE-LS pair an exactly matched comparison in which only label smoothing varies. A final system, ResNet-18 on Oxford-IIIT Pets (Parkhi et al., 2012) (37 breeds, four seeds), serves as a cross-dataset reference. All systems are evaluated post-hoc under the identical SCF protocol: stratified train reference, held-out test queries, shared granularity grid, median-distance normalization, $N_{\text{eff}} \geq 10$ gate. Full training protocols, accuracies, and per-seed descriptor matrices are in App. B. The benchmark spans a

Table 3: SCF morphology matrix (four seeds, seed mean; headline τ=0.27, the finest granularity passing the validity gate for every system, layer, and seed). Left: raw-field descriptors (M: depth slope; D: cross-granularity trend; ΔZ: raw magnitude; Rev.: deep reversal, fraction of seeds; R: seed consistency). Right: null-calibrated (ΔZ_{cal}: label-driven depth drop; O_{last}: organization beyond null at the final layer). Bold: the raw markers traced to drift in Sec. 7.4.

System	M	D	ΔZ	Rev.	R	ΔZ_{cal}	O_{last}
CE	−0.99	+0.93	0.211	0/4	0.97	+0.181	+0.229
CE* (matched budget)	−0.93	+0.93	0.114	0/4	0.95	+0.091	+0.137
CE-LS	−0.99	+0.93	0.245	0/4	0.97	+0.199	+0.246
SupCon	−0.99	+0.99	0.184	0/4	0.95	+0.149	+0.198
SimCLR	−0.99	+0.77	0.074	**4/4**	0.95	+0.055	+0.102
DINO-lite	−0.62	+0.64	0.018	**4/4**	0.50	+0.016	+0.063
VICReg	−0.99	+0.87	0.088	**4/4**	0.92	+0.037	+0.085
ResNet (Pets)	−0.99	**−0.78**	0.127	**4/4**	0.84	+0.099	+0.116

wide quality range (accuracy 0.712 down to 0.199), so descriptors can be decoupled from raw downstream performance.

7.2 Distinct semantic field morphologies

Table 3 shows the morphology matrix (mean over four seeds; per-seed matrix in App. B). Because M and D are rank correlations over five to eight layers, we use them as ordinal morphology markers; the continuous descriptors (ΔZ_{cal}, O_{last}, κ) carry the quantitative content. Systems sharing a backbone and dataset nevertheless occupy clearly different field morphologies.

7.3 Semantic field regimes

Three total-field morphologies emerge consistently across the benchmark. **Progressive total-field organization** (CE, CE*, CE-LS, SupCon): saturated depth slope ($M = -0.99$), positive cross-granularity trend, the largest magnitudes, and no deep reversal in any seed. **Re-entrant total-field morphology** (SimCLR, VICReg, Pets): fine-grained re-emergence at the deepest layer (4/4 seeds) with intermediate magnitudes; Pets is additionally *scale-reversed* on the raw field ($D = -0.78$). **Weak total-field organization** (DINO-lite): an order of magnitude smaller ($\Delta Z = 0.018$). These morphologies characterize the observed total field; we next separate its label-driven and geometric components.

7.4 Null calibration: separating label-driven organization from geometric concentration

Permutation-null calibration measures label-driven organization in excess of a label-permuted geometric baseline. Permuting the reference labels (80 null fields per system) reveals that baseline directly: null fields still exhibit a systematic depth trend (null depth slope ≤ -0.76 for six of seven systems), small in magnitude (≤ 0.05 vs up to 0.25 real) but structured. The mechanism is geometric: deeper layers tighten clusters, lowering N_{eff}, so chance label imbalance among fewer effective neighbors lowers entropy even with random labels. Raw descriptor space barely separates real from null fields (507 of 560 null vectors fall inside the real regime clusters); the raw scale-reversed and deep-reversal markers belong to the drift component.

Subtracting the permutation-null mean field everywhere measures the *organization beyond null*, $O(l) = \bar{Z}_{\text{null}}(l) - \bar{Z}_{\text{real}}(l)$: label-driven organization in excess of the label-permuted baseline (Table 3, right; Fig. 5). Every benchmarked system exhibits label-driven organization beyond its permutation null (depth drop and final-layer organization exceed all 80 null realizations, $p_{\text{perm}} \leq 1/81$). The supervised family organizes most strongly ($O_{\text{last}} = 0.198$–0.246), self-supervised systems more weakly (0.063–0.102), and DINO-lite weakest (0.063). Under the matched protocol, CE-LS exhibits substantially stronger calibrated organization

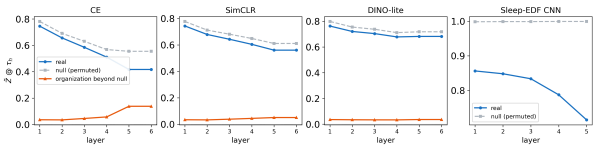

Figure 5: Null calibration (seed 0, fine level). Grey: label-permutation null; blue: real; orange: organization beyond null. The null drifts on CIFAR-100 but is flat on Sleep-EDF Expanded.

than CE* (ΔZ_{cal} 0.199 vs 0.091). After calibration, fine-led organization is universal on CIFAR-100 ($D_{\text{cal}} = +0.99$), while Pets' coarse-led trend is not significant against its own null ($p = 0.41$). The raw deep-reversal markers vanish after calibration (0/4 seeds in every system). The calibration is domain-dependent: on Sleep-EDF Expanded the null field is flat ($\bar{Z}_{\text{null}} = 0.999$, $N_{\text{eff}} \approx 4000$), so the physiological depth organization of Fig. 4 is entirely label-driven.

Reconfiguration depth and seed consistency. CE-based systems reconfigure at mid-to-late layers ($L^* = 0.6$) while several self-supervised systems reconfigure near the input side ($L^* = 0$); seed consistency is system-dependent (App. C).

Established local statistics and a multiscale kNN-entropy baseline reproduce the coarse regimes (App. C.4); SCF's distinctive content is composition fidelity, the (d, Γ) field, and the validity-gated, null-calibrated protocol.

8 Reliability of SCF Morphology

Descriptor complementarity. Across all 28 system–seed units, no descriptor pair exceeds the prespecified redundancy threshold ($|\rho| > 0.9$; Fig. 8, App. C): e.g., CE and SupCon share the saturated depth slope yet differ sharply in reconfiguration depth.

Perturbation, scale, and metric robustness. Under three structural perturbations per unit (band averaging, first/last-layer removal), the reversal type persists in 26/28 units and the cross-granularity trend sign in 24/28; across the valid band, the sign holds in 21/28; and under cosine/unit-normalized/whitened metrics all drop signs agree (ordering: Spearman 1.00 whitened, 0.77 cosine). Details in App. C.

9 Discussion and Conclusion

Representation-specific morphology. The protocol is shared, but the resulting organization is system-specific, and null calibration makes the comparison semantic by removing the geometric component; the per-point field also carries information about prediction errors (App. C).

Conclusion. We introduced Zentropy Semantic Configuration Fields: local semantic entropy, validity-gated and resolved as a field over depth and granularity. Controlled experiments establish what the field measures; five modalities show it ports unchanged; and a null-calibrated benchmark separates systems by how much, when, and at which granularity they organize semantics. SCF thus provides a foundation for auditing the semantic side of representation learning alongside accuracy-side evaluation.

Reproducibility statement

All estimators, protocols, and descriptor definitions are specified in Sec. 3 and App. B (including pseudocode); complete proofs appear in App. A; per-seed descriptor matrices, training protocols, and all reported statistics are tabulated in App. B and App. C. Source code and cached representations will be released as supplementary material.

AI use statement

We used generative AI tools for code assistance and manuscript polishing. All experiments were run and verified by the authors, all numerical claims were checked against the stored experiment outputs. All citations were manually verified. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

S. A. Alvarez. Gaussian rbf centered kernel alignment (cka) in the large-bandwidth limit. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):6587–6593, 2023.

Rubén Ballester, Carles Casacuberta, and Sergio Escalera. Topological data analysis for neural network analysis: A comprehensive survey, 2023.

Rubén Ballester, Xavier Arnal Clemente, Carles Casacuberta, Meysam Madadi, Ciprian A. Corneanu, and Sergio Escalera. Predicting the generalization gap in neural networks using topological data analysis. *Neurocomputing*, 596:127787, 2024. doi: 10.1016/j.neucom.2024.127787.

Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations. In *Proc. ICML (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1607–1626. PMLR, 2022.

A. Bardes, J. Ponce, and Y. LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *Proc. ICLR*, 2022.

M. Caron et al. Emerging properties in self-supervised vision transformers. In *Proc. ICCV*, pp. 9650–9660, 2021.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, pp. 1597–1607, 2020.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13 (1):21–27, 1967.

H. A. Dau et al. The ucr time series archive. *IEEE/CAA J. Autom. Sinica*, 6(6):1293–1305, 2019.

M. R. Davari, S. Horoi, A. Natik, et al. Reliability of cka as a similarity measure in deep learning. In *Proc. ICLR*, 2023.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, NY, USA, 1996.

A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*, 2021.

A. L. Goldberger et al. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9052–9071, 2024. doi: 10.1109/TPAMI.2024.3415112.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pp. 770–778, 2016.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

H. Jeon, Y.-H. Kuo, M. Aupetit, K.-L. Ma, and J. Seo. Classes are not clusters: Improving label-based evaluation of dimensionality reduction. *IEEE Trans. Vis. Comput. Graphics*, 30(1):781–791, 2024.

B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the eeg. *IEEE Trans. Biomed. Eng.*, 47(9):1185–1194, 2000.

Bob Kemp, Aeilko H. Zwinderman, Bert Tuk, H. A. C. Kamphuisen, and J. J. L. Oberyé. Sleep-EDF Database Expanded v1.0.0. PhysioNet, 2013. URL `https://physionet.org/content/sleep-edfx/1.0.0/`. Deposited October 24, 2013.

P. Khosla et al. Supervised contrastive learning. In *Proc. NeurIPS*, volume 33, pp. 18661–18673, 2020.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Comput. Surv.*, 57(9):1–52, 2025a. doi: 10.1145/3728458.

Max Klabunde, Tassilo Wald, Tobias Schumacher, Klaus Maier-Hein, Markus Strohmaier, and Florian Lemmerich. Resi: A comprehensive benchmark for representational similarity measures. In *Proc. ICLR (ICLR)*, 2025b. URL `https://openreview.net/forum?id=PRvd03nfFi`.

S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *Proc. ICML*, pp. 3519–3529, 2019.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, 2004. Art. no. 066138.

A. Krizhevsky. Learning multiple layers of features from tiny images, 2009. Tech. Rep., Univ. of Toronto, Toronto, ON, Canada.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proc. ICLR (ICLR)*, 2023.

B. Lin and N. Kriegeskorte. The topology and geometry of neural representations. *Proc. Nat. Acad. Sci. USA*, 121(42):e2317881121, 2024. Art. no. e2317881121.

Peng Liu, Wei Qian, Hua Zhang, Yabin Zhu, Qi Hong, Qiang Li, and Yudong Yao. Automatic sleep stage classification using deep learning: Signals, data representation, and neural networks. *Artif. Intell. Rev.*, 57(11):301, 2024. doi: 10.1007/s10462-024-10926-9.

Z.-K. Liu, B. Li, and H. Lin. Multiscale entropy and its implications to critical phenomena, emergent behaviors, and information. *J. Phase Equilibria Diffusion*, 40(4):508–521, 2019.

Z.-K. Liu, Y. Wang, and S.-L. Shang. Zentropy theory for positive and negative thermal expansion. *J. Phase Equilibria Diffusion*, 43(6):598–605, 2022.

Zhexuan Liu, Rong Ma, and Yiqiao Zhong. Assessing and improving reliability of neighbor embedding methods: a map-continuity perspective. *Nature Communications*, 16(1):5037, 2025. doi: 10.1038/s41467-025-60434-9. arXiv:2410.16608.

Achleshwar Luthra, Tianbao Yang, and Tomer Galanti. Self-supervised contrastive learning is approximately supervised contrastive learning. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

George Ma, Samuel Pfrommer, and Somayeh Sojoudi. Revising and falsifying sparse autoencoder feature explanations. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Gabriela Moisescu-Pareja, Gavin McCracken, Harley Wiltzer, Vincent Létourneau, Colin Daniels, Doina Precup, and Jonathan Love. On the geometry and topology of representations: the manifolds of modular addition. In *Proc. ICLR (ICLR)*, 2026.

E. A. Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9(1):141–142, 1964.

V. Papyan, X. Y. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proc. Nat. Acad. Sci. USA*, 117(40):24652–24663, 2020.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *Proc. CVPR*, pp. 3498–3505, 2012.

K. J. Piczak. Esc: Dataset for environmental sound classification. In *Proc. ACM MM*, pp. 1015–1018, 2015.

Michael Robinson, Sourya Dey, and Tony Chiang. Token embeddings violate the manifold hypothesis. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

R. S. Rosenberg and S. Van Hout. The american academy of sleep medicine inter-scorer reliability program: Sleep stage scoring. *J. Clin. Sleep Med.*, 9(1):81–87, 2013.

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, and Milica Gašić. Less is more: Local intrinsic dimensions of contextual language models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Luis Scoccola, Uzu Lim, and Heather A. Harrington. Cover learning for large-scale topology representation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53728–53756. PMLR, 2025. arXiv:2503.09767.

C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423, 623–656, 1948.

C. J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–620, 1977.

H. Tan, K. Gai, and S. Zhang. Beyond neural collapse: Task-intrinsic geometry governs neural representations in modular arithmetic, 2026.

Tiago F. Tavares, Fabio Ayres, and Paris Smaragdis. Diagnosing neural convergence with topological alignment spectra. In *Proc. IJCNN (IJCNN), IEEE World Congress on Computational Intelligence (WCCI)*, 2026.

Lenka Tětková, Thea Brüsch, Teresa Dorszewski, Fabian Martin Mager, Rasmus Ørtoft Aagaard, Jonathan Foldager, Tommy Sonne Alstrøm, and Lars Kai Hansen. On convex decision regions in deep network representations. *Nature Communications*, 16:5419, 2025. doi: 10.1038/s41467-025-60809-y.

Rahul Thapa, Magnus Ruud Kjær, Bryan He, Ian Covert, Hyatt Moore, Umaer Hanif, Gauri Ganjoo, M. Brandon Westover, Poul Jennum, Andreas Brink-Kjær, Emmanuel Mignot, and James Zou. A multimodal sleep foundation model developed with 500K hours of sleep recordings for disease predictions. *Nature Medicine*, 2026. doi: 10.1038/s41591-025-04133-4.

Ilya Trofimov, Daniil Cherniavskii, Eduard Tulchinskii, Nikita Balabin, Evgeny Burnaev, and Serguei Barannikov. Learning topology-preserving data representations. In *Proc. ICLR (ICLR)*, 2023.

A. Vaswani et al. Attention is all you need. In *Proc. NeurIPS*, volume 30, 2017.

Shun Wang, Shun-Li Shang, Zi-Kui Liu, and Wenrui Hao. ZENN: A thermodynamics-inspired computational framework for heterogeneous data-driven modeling. *Proc. Nat. Acad. Sci. USA*, 123(1):e2511227122, 2026. doi: 10.1073/pnas.2511227122.

Yan Wang and Tianyang Hu. Symmetric divergence and normalized similarity: A unified topological framework for representation analysis. *Transactions on Machine Learning Research (TMLR)*, 2026. arXiv:2606.06342 [stat.ML].

G. S. Watson. Smooth regression analysis. *Sankhyā A*, 26(4):359–372, 1964.

Ning Wu, Han Huang, Yueting Xu, and Zhifeng Hao. Neural architecture selection via maximizing local entropy of feature maps. *Neurocomputing*, 690:133783, 2026. doi: 10.1016/j.neucom.2026.133783.

K. Yuan, D. Miao, Y. Yao, H. Zhang, and X. Zhao. Feature selection using zentropy-based uncertainty measure. *IEEE Trans. Fuzzy Syst.*, 32(4):2246–2260, Apr. 2024.

Kehua Yuan, Duoqian Miao, Witold Pedrycz, Hongyun Zhang, and Liang Hu. Multigranularity data analysis with zentropy uncertainty measure for efficient and robust feature selection. *IEEE Trans. Cybern.*, 55(2):740–752, 2025. doi: 10.1109/TCYB.2024.3499952.

Kehua Yuan, Duoqian Miao, Weiping Ding, Witold Pedrycz, and Yiyu Yao. Robust semi-supervised feature selection with multi-granularity zentropy modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(4):4587–4604, 2026a. doi: 10.1109/TPAMI.2025.3647921.

Kehua Yuan, Duoqian Miao, Witold Pedrycz, and Yiyu Yao. Zentropy-enhanced multigranularity knowledge modeling for robust feature selection. *IEEE Trans. Cybern.*, 2026b. doi: 10.1109/TCYB.2026.3665802. Early Access.

X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Proc. NeurIPS*, volume 28, 2015.

A Proofs and Asymptotic Properties

We work at a fixed layer l (suppressed), with reference set $\mathcal{R} = \{(z_j, y_j)\}_{j=1}^{N}$, $y_j \in \{1, \dots, K\}$, and normalized weights $w_{ij} \geq 0$, $\sum_j w_{ij} = 1$, so that $\hat{p}_i^{(\tau)}(c) = \big(\sum_j w_{ij}^{(\tau)} \mathbf{1}[y_j = c] + \alpha\big)/(1 + K\alpha)$.

A.1 Finite-sample properties

Proof of Theorem 1. Every coordinate of $\hat{p}_i^{(\tau)}$ obeys $\hat{p}_i^{(\tau)}(c) \geq a$ with $a = \alpha/(1 + K\alpha)$, since the weights are nonnegative and sum to one. Consider the simplex region $\Delta_a = \{p : p_c \geq a \ \forall c, \ \sum_c p_c = 1\}$ and the point $q_0 = (1 - (K-1)a, a, \dots, a)$. We show q_0 majorizes every $p \in \Delta_a$. Sort p in nonincreasing order, $p_{(1)} \geq \cdots \geq p_{(K)}$. For the first partial sum, $p_{(1)} = 1 - \sum_{c \geq 2} p_{(c)} \leq 1 - (K-1)a = q_{0,(1)}$. For $k \geq 2$, $\sum_{c=1}^{k} p_{(c)} = 1 - \sum_{c>k} p_{(c)} \leq 1 - (K-k)a = \sum_{c=1}^{k} q_{0,(c)}$, since the $K - k$ remaining coordinates of p are each at least a. Hence $p \prec q_0$ for all $p \in \Delta_a$. Shannon entropy is Schur-concave, so $H(p) \geq H(q_0)$ on Δ_a, with equality only at permutations of q_0. Direct substitution of q_0 into the entropy functional yields $Z_{\min}(K, \alpha)$, and q_0 is attained exactly when every neighbor shares one label. The maximum statement is the standard fact that entropy on the simplex is uniquely maximized at the uniform point. The normalized measure is affine in Z and maps $Z_{\min} \mapsto 0$ and $\log K \mapsto 1$.

Proof of Theorem 2. By the Hardy–Littlewood–Pólya theorem, $p \succ q$ iff there exists a doubly stochastic matrix D with $q = Dp$. Let $\phi(t) = t \log t$, strictly convex on $[0, 1]$. Row c of D is a convex combination vector, so by Jensen's inequality $\phi(q_c) = \phi(\sum_r D_{cr} p_r) \leq \sum_r D_{cr}\, \phi(p_r)$. Summing over c and using that the columns of D also sum to one, $\sum_c \phi(q_c) \leq \sum_r \big(\sum_c D_{cr}\big)\phi(p_r) = \sum_r \phi(p_r)$, i.e., $-Z(q) \leq -Z(p)$. Strict convexity of ϕ makes the inequality strict unless each row of D places all mass on entries of equal value, which under $p \succ q$ forces q to be a permutation of p.

Proof of Theorem 3. For any pair, $\|z'_i - z'_j\|_2^2 = \|cA(z_i - z_j)\|_2^2 = c^2 (z_i - z_j)^\top A^\top A (z_i - z_j) = c^2 \|z_i - z_j\|_2^2$: translation by b cancels in the difference and orthogonality of A preserves the norm. The pairwise squared distances are thus all multiplied by the same constant c^2, so their median is multiplied by c^2 as well, and $\tilde{d}'^2_{ij} = c^2 d_{ij}^2 / \operatorname{med}(c^2 d^2) = \tilde{d}^2_{ij}$. Kernels,

weights, posteriors, entropies, effective neighbor counts, and all field-level and descriptor-level quantities derived from them are unchanged pointwise. (Equivalently, on raw distances the same invariance is obtained by jointly rescaling the bandwidth so that $\tau'^2 = c^2\tau^2$.) Nonlinear reparameterizations (whitening, per-component PCA scaling) are not covered by the theorem and are audited empirically instead (Sec. 8).

A.2 Asymptotic behavior

For the asymptotic statements we adopt the standard kernel-regression setting: the reference set consists of N i.i.d. draws from a design density f supported on a compact set $\mathcal{Z} \subset \mathbb{R}^{d_z}$, and $\eta_c(z) = P(Y{=}c \mid Z{=}z)$ denotes the population conditional class probability.

Theorem 4 (Kernel posterior consistency)**.** *Assume (i) $\mathcal{Z}$ is compact and $f(z) > 0$ at the evaluation point; (ii) each η_c is Lipschitz; (iii) the kernel is bounded, symmetric, and integrable; (iv) $\tau_N \to 0$; (v) $N\tau_N^{d_z} \to \infty$; and (vi) the smoothing sequence satisfies $\alpha_N \to 0$. Then the leave-one-out kernel posterior satisfies, at every continuity point z of f, $\hat{p}_N(c \mid z) \xrightarrow{p} \eta_c(z)$.*

Proof sketch. The unnormalized numerator $\frac{1}{N\tau_N^{d_z}} \sum_j K\big(\frac{z-z_j}{\tau_N}\big)\mathbf{1}[y_j = c]$ is a kernel estimate of $\eta_c(z)f(z)$, and the denominator is a kernel estimate of $f(z)$; under (i)–(v) both converge in probability by the classical consistency theory of Nadaraya–Watson regression (Nadaraya, 1964; Watson, 1964; Stone, 1977; Devroye et al., 1996). Since $f(z)$ is bounded away from zero in a neighborhood of a continuity point with $f(z) > 0$, the ratio converges to $\eta_c(z)$ by Slutsky's theorem. Removing the leave-one-out term changes numerator and denominator by $O\big(1/(N\tau_N^{d_z})\big)$, which vanishes under (v). The Dirichlet shift contributes a deterministic bias of order α_N, which vanishes under (vi). Condition (vi) is essential: at any fixed $\alpha > 0$ the estimator converges instead to $(\eta_c(z) + \alpha)/(1 + K\alpha) \neq \eta_c(z)$.

Theorem 5 (Entropy error decomposition, corrected rate)**.** *Under the conditions of Theorem 4, let r be the smoothness (Hölder) order of the functions η_c. Then, at a continuity point z,*

$$\big|\hat{Z}_N(Y \mid Z{=}z) - H(Y \mid Z{=}z)\big| = O_p\left(L_N\Big(\frac{1}{\sqrt{N\tau_N^{d_z}}} + \tau_N^r\Big) + \alpha_N\right), \qquad L_N = K\log(1/\alpha_N), \tag{4}$$

so consistency follows whenever $\alpha_N \to 0$ with $L_N\big((N\tau_N^{d_z})^{-1/2} + \tau_N^r\big) \to 0$, e.g. $\alpha_N \to 0$ polynomially while $(N\tau_N^{d_z})^{-1/2}$ decays polynomially.

Proof sketch. On the clipped simplex $\{p : p_c \geq \alpha_N\}$ the entropy functional is Lipschitz in $\|\cdot\|_1$ with constant of order $L_N = K\log(1/\alpha_N)$, because $|\log p|$ is then bounded; unlike in the unclipped case, entropy is not uniformly Lipschitz as probabilities approach zero, so the factor L_N cannot be omitted without further conditions. Hence $|\hat{H}(\hat{p}) - \hat{H}(\eta)| \leq L_N\|\hat{p} - \eta\|_1$ pointwise. The kernel regression error for each class decomposes into a stochastic term of order $O_p\big((N\tau_N^{d_z})^{-1/2}\big)$ and a deterministic term of order $O(\tau_N^r)$ under the Hölder condition on η_c; the Dirichlet shift adds the deterministic term $O(\alpha_N)$ of Theorem 4. The stated condition on α_N makes L_N slowly growing relative to the decaying terms, so the right-hand side vanishes.

Remark 1 (Scope of the rate statements)**.** *Two qualifications delimit Theorem 5. First, the rate is pointwise. The corpus-level Zentropy $\hat{Z}_N = \frac{1}{N}\sum_i Z_i^{(\tau_N)}$ averages per-query estimates that share the same reference set and are therefore dependent; we report the corpus average as an empirical summary with bootstrap confidence intervals rather than claiming an averaged-rate theorem, and all empirical comparisons in the paper are made at matched protocol and bandwidth. Second, the condition $N\tau_N^{d_z} \to \infty$ is restrictive in high-dimensional ambient spaces and should be read in the intrinsic dimension of the representation; we do not claim the asymptotic regime is reached in our finite experiments, which is precisely why the validity gate and the empirical reliability analyses of Sec. 8 carry the evidential weight.*

Proposition 1 (Conditional information gain)**.** *Let Z_S and Z_m denote representations computed from a set S of inputs and an additional input m, and define the estimator-level gain $\Delta\hat{Z}_{m|S} = \hat{Z}(Y \mid Z_S) - \hat{Z}(Y \mid Z_S, Z_m)$ with matched bandwidths. Under the conditions of Theorem 5 applied to both estimates, $\Delta\hat{Z}_{m|S} \xrightarrow{p} H(Y \mid Z_S) - H(Y \mid Z_S, Z_m) = I(Y; Z_m \mid Z_S)$.*

Proof sketch. Each term converges in probability to its population counterpart by Theorem 5; the difference of two consistent estimators is consistent for the difference; and $I(Y; Z_m \mid Z_S) = H(Y \mid Z_S) - H(Y \mid Z_S, Z_m)$ is the chain-rule identity for conditional mutual information.

Proposition 1 must be read at the estimator level: $\Delta\hat{Z}_{m|S}$ is an *estimate* of a conditional information contribution, not a population mutual information, and at finite N it is not guaranteed nonnegative per estimate; small empirical gains are interpreted through bootstrap confidence intervals with matched bandwidths. This distinguishes the quantity from dedicated kNN mutual-information estimators (Kraskov et al., 2004).

Metric-space generality. Nothing in the construction requires $\mathcal{Z} \subset \mathbb{R}^{d_z}$: the estimator needs only a metric space $(\mathcal{Z}, d)$ and labeled references. The full profile (posterior, Zentropy, signature, field, descriptors) is defined unchanged on graphs, manifolds, or hyperbolic embeddings; Theorems 1, 2, and 3 are metric-agnostic, and only Theorems 4 and 5 use the Euclidean structure.

B Estimation and Evaluation Protocol

B.1 Estimator protocol (pseudocode)

All reported SCF fields use the following fixed protocol.

1. **Reference and query construction.** The labeled reference set $\mathcal{R}$ is a stratified subset of the training split (fixed across systems within a modality; CIFAR-100: 5000 images, 50 per fine class; Pets: 1850 images, 50 per breed; Sleep-EDF Expanded and other modalities: stratified train subsets of the sizes reported per experiment). Queries are the *held-out test split*; queries never vote and never appear in the reference set, so no self-neighbor inclusion is possible in any benchmark result. Where in-sample queries are used (synthetic validity experiments of Sec. 5), the query's own index is explicitly excluded from its neighborhood (leave-self-out).
2. **Distance normalization.** Compute squared Euclidean distances d_{ij}^2 between query i and reference j; normalize by the reference median squared distance, $\tilde{d}_{ij}^2 = d_{ij}^2 / (\mathrm{med}_{j<j'}\, d_{jj'}^2 + 10^{-12})$. All bandwidths live on this normalized, dimensionless scale.
3. **Kernel weights.** For each granularity τ in the grid: $w_{ij}^{(\tau)} = \exp(-\tilde{d}_{ij}^2/\tau^2)$, row-normalized over reference points.
4. **Validity gate.** Compute $N_{\mathrm{eff},i} = (\sum_j w_{ij}^2)^{-1}$; a granularity is valid only if the mean N_{eff} over queries exceeds $n_{\min} = 10$. Invalid granularities are excluded from all analyses (they are never silently interpolated).
5. **Posterior and entropy.** $\hat{p}_i^{(\tau)}(c) = (\sum_j w_{ij} \mathbf{1}[y_j = c] + \alpha)/(1 + K\alpha)$, with $\alpha = 0.05$ in the synthetic and physiological experiments and $\alpha = 0.25/K$ in the CIFAR-100/Pets benchmark; $Z_i^{(\tau)} = -\sum_c \hat{p}_i^{(\tau)}(c) \log \hat{p}_i^{(\tau)}(c)$; normalized $\tilde{Z}$ against $Z_{\min}(K, \alpha)$ and $\log K$ (Theorem 1).
6. **Field assembly.** $\tilde{Z}(l, \tau)$ is the corpus mean of per-query $\tilde{Z}_i$ at layer l, computed per seed; seed means are then averaged. Confidence intervals, where reported, are query bootstraps ($B = 1000$) within each seed, then combined across seeds.

Distance concentration and scaling. Median normalization fixes the global scale but not distance concentration in high-dimensional early layers: the ratio of per-query distance spread to the median is audited per layer, and the N_{eff} gate rejects granularities where concentration collapses the kernel. On scaling: reference-set size N enters the error decomposition as $(N\tau^{d_z})^{-1/2}$, so large-label-space or massively imbalanced datasets need stratified reference sampling (which we use throughout) and, at scale, approximate-nearest-neighbor kernels; both are engineering extensions, not changes to the estimator.

N_{eff} **tail audit.** The valid granularity band is computed jointly across systems, layers, and seeds; within it, the minimum *mean* N_{eff} is 50.2 (gate: 10), so no reported cell approaches the 1-NN regime. Per-query lower quantiles at the headline granularity are reported in `results/neff_report.json`.

Held-out vs in-sample queries. Recomputing the CIFAR-100 CE field with in-sample leave-self-out queries (the query's own reference row excluded via `ref_self_idx`) instead of held-out test queries yields depth profiles with Spearman correlation 1.00 and maximum absolute deviation 0.08 nats at the headline granularity (three seeds; `results/heldout_check.json`). The layerwise trends reported throughout are therefore not artifacts of query construction.

B.2 Descriptor computation

All descriptors are computed *per seed* from that seed's field $\tilde{Z}(l,\tau)$ (layers indexed $0,\dots,L{-}1$ from input side; input/embedding layers are included as layer 0), then reported as seed means; the categorical reversal is reported as the fraction of seeds exhibiting it. Let τ_h denote the headline granularity (the finest τ passing the validity gate for every system, layer, and seed in the comparison; $\tau_h = 0.27$ for the CIFAR/Pets benchmark, the finest granularity at which every system, layer, and seed passes the validity gate) and τ_c the coarsest valid granularity.

- $M = \rho_s\big(l, \tilde{Z}(l,\tau_h)\big)$, Spearman correlation over the L layer indices. Ties are broken by average ranks (the `scipy` convention).
- $D_{\text{scale}} = \rho_s\big(l, \tilde{Z}(l,\tau_h) - \tilde{Z}(l,\tau_c)\big)$.
- $L^* = \arg\max_l |\tilde{Z}(l{+}1,\tau_h) - \tilde{Z}(l,\tau_h)|/(L-1)$; ties broken toward the smaller index.
- $\kappa = \max_l |\tilde{Z}_{l+1} - 2\tilde{Z}_l + \tilde{Z}_{l-1}|/\operatorname{range}(\tilde{Z})$ at τ_h, over both label hierarchy levels where applicable.
- $\Delta Z_{\text{total}} = \tilde{Z}(0,\tau_h) - \tilde{Z}(L{-}1,\tau_h)$.
- R_{deep}: let $l^\dagger$ be the last layer with $|\Delta_l \tilde{Z}| > 10^{-6}$ at τ_h; reversal holds iff $\Delta\tilde{Z}(l^\dagger) < 0.5\,\Delta\tilde{Z}(0)$ and $\tilde{Z}_{\text{fine}}(l^\dagger) < \tilde{Z}_{\text{coarse}}(l^\dagger)$.
- $R_{\text{SCF}} = 1 - \|\mathcal{S}^{(a)} - \mathcal{S}^{(b)}\|_F \,/\, \mathbb{E}_{\text{perm}}\|\mathcal{S}^{(a)} - \mathcal{S}^{(b)}_{\text{perm}}\|_F$ with 50 random layer permutations, averaged over all $\binom{4}{2}$ seed pairs.

Because M and D_{scale} are rank correlations over five to eight layers, their values are coarsely quantized (e.g., -0.99, ± 0.93, ± 0.81); we read them as ordinal regime markers, not as precise continuous measurements.

B.3 Training protocols

Table 4 reports the training configuration of every benchmarked system. Budgets are comparable but not identical: the CE reference was trained with a shorter schedule and standard augmentation, and the label-free systems are evaluated with frozen-feature linear probes. CE* is plain CE retrained under the exact shared protocol (35 epochs, strong augmentation), making CE*/CE-LS an exactly matched pair in which only label smoothing varies; under this control, label smoothing more than doubles calibrated organization (ΔZ_{cal} 0.199 vs 0.091).

Table 4: Training protocols and test/probe accuracies (mean over four seeds).

System	Epochs	Batch	Augmentation	Eval	Acc.
CE	25	512	standard	test acc	0.712
CE* (matched)	35	512	strong (SimCLR-style)	test acc	0.495
CE-LS (0.1)	35	512	strong (SimCLR-style)	test acc	0.691
SupCon	35	512	strong	linear probe	0.628
SimCLR	35	512	strong	linear probe	0.372
DINO-lite	35	512	strong	linear probe	0.206
VICReg	35	512	strong	linear probe	0.402
ResNet (Pets)	40	512	standard	test acc	0.311

B.4 Per-seed morphology matrix

Table 5 reports the full descriptor matrix for every system and seed (the seed-level resolution underlying Table 3).

Table 5: Per-seed SCF morphology descriptors (headline τ=0.27). Note the coarse quantization of the rank-based descriptors M and D, which we interpret as ordinal regime markers only.

System	Seed	M	D	L^*	κ	ΔZ	Rev.
CE	0	−0.99	+0.93	0.60	0.162	0.216	no
CE	1	−0.99	+0.93	0.60	0.164	0.211	no
CE	2	−0.99	+0.93	0.60	0.156	0.205	no
CE	3	−0.99	+0.93	0.60	0.161	0.213	no
CE* (matched)	0	−0.93	+0.93	0.60	0.091	0.110	no
CE* (matched)	1	−0.93	+0.93	0.60	0.091	0.115	no
CE* (matched)	2	−0.93	+0.93	0.60	0.091	0.113	no
CE* (matched)	3	−0.93	+0.93	0.60	0.095	0.118	no
CE-LS	0	−0.99	+0.93	0.60	0.207	0.242	no
CE-LS	1	−0.99	+0.93	0.60	0.209	0.246	no
CE-LS	2	−0.99	+0.93	0.60	0.202	0.247	no
CE-LS	3	−0.99	+0.93	0.60	0.204	0.246	no
DINO-lite	0	−0.64	+0.64	0.00	0.015	0.015	yes
DINO-lite	1	+0.06	+0.58	0.00	0.019	0.017	yes
DINO-lite	2	−0.93	+0.41	0.60	0.014	0.015	yes
DINO-lite	3	−0.99	+0.93	0.40	0.007	0.026	yes
ResNet (Pets)	0	−0.99	−0.99	0.60	0.054	0.113	yes
ResNet (Pets)	1	−0.99	−0.81	0.60	0.100	0.138	yes
ResNet (Pets)	2	−0.99	−0.52	0.60	0.117	0.142	yes
ResNet (Pets)	3	−0.99	−0.81	0.60	0.060	0.115	yes
SimCLR	0	−0.99	+0.81	0.60	0.038	0.071	yes
SimCLR	1	−0.99	+0.75	0.60	0.040	0.074	yes
SimCLR	2	−0.99	+0.75	0.60	0.038	0.075	yes
SimCLR	3	−0.99	+0.75	0.60	0.040	0.077	yes
SupCon	0	−0.99	+0.99	0.60	0.115	0.186	no
SupCon	1	−0.99	+0.99	0.60	0.119	0.190	no
SupCon	2	−0.99	+0.99	0.60	0.128	0.183	no
SupCon	3	−0.99	+0.99	0.60	0.103	0.178	no
VICReg	0	−0.99	+0.99	0.60	0.060	0.089	yes
VICReg	1	−0.99	+0.93	0.60	0.056	0.085	yes
VICReg	2	−0.99	+0.75	0.60	0.055	0.091	yes
VICReg	3	−0.99	+0.81	0.60	0.054	0.089	yes

C Extended Results

C.1 Manifold traversal

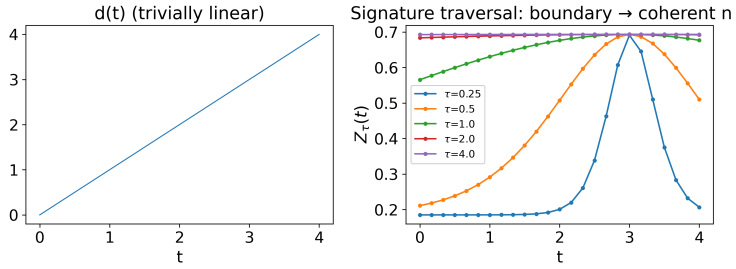

Figure 6: Signature traversal: $Z(t)$ rises through the semantic boundary and falls into the coherent-novel region; $d(t)$ is trivially monotone.

A continuous path from the core through a semantic boundary into a coherent-novel region produces a non-monotonic Zentropy profile (Fig. 6): $Z(t)$ rises through the boundary and falls as the path enters the novel but coherent cluster, while the distance profile is trivially monotone. The four-quadrant structure is therefore not an artifact of discrete zone construction.

C.2 Fine/coarse hierarchy on CIFAR-100

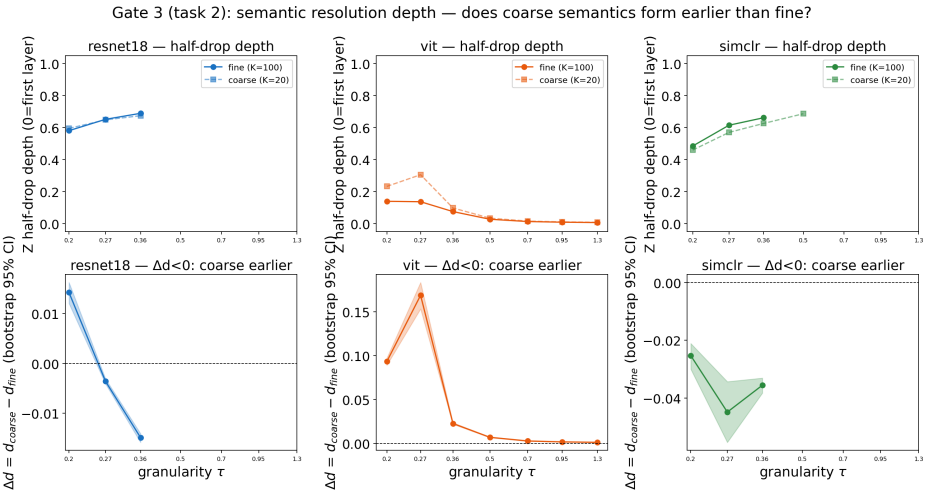

Figure 7: Fine-vs.-coarse semantic resolution on CIFAR-100. Fine neighborhoods are purer at every layer (paired Wilcoxon, $p \approx 0$), yet $\Delta\tilde{Z}$ increases with depth in all three architectures: a weak, architecture-dependent shift resolvable only at fine granularities.

Under two nested reference labelings (100 fine classes, 20 superclasses), the purity advantage of fine neighborhoods shrinks with depth in all three architectures (depth-wise Spearman correlation of $\Delta\tilde{Z} = \tilde{Z}_{\text{fine}} - \tilde{Z}_{\text{coarse}}$: +0.99 ResNet, +0.64 ViT, +0.99 SimCLR), yet fine neighborhoods remain strictly purer than coarse ones at every layer, and the effect is small ($|\Delta\tilde{Z}| \leq 0.02$) and resolvable only at fine granularities. We read this as one observable field morphology, not a universal semantic hierarchy.

C.3 Label-permutation null control

For each system and seed we permute the reference labels 20 times (identical permutation across hierarchy levels) and recompute the full field, yielding 80 null fields per system. Three findings. **(i) Concentration drift.** Null fields are not flat: at the headline granularity their depth slope is ≤ -0.76 for six of seven systems, though the null raw magnitude stays small (≤ 0.05 versus up to 0.25 real). The drift tracks cluster tightening with depth (lower N_{eff}), which lowers entropy through chance imbalance; the $N_{\text{eff}} \geq 10$ gate reduces but does not eliminate it at K=100. **(ii) Raw descriptors barely separate real from null.** In the standardized raw descriptor space, 507 of 560 null vectors fall inside the real regime clusters, and the null deep-reversal rule fires at rate 1.00. This motivates the calibration of Sec. 7.4. **(iii) After calibration**, every system's depth drop and final-layer organization are significant against the permutation null ($p = 0.012$, the floor of 80 nulls per system); only Pets' cross-granularity trend is not ($p = 0.41$). The calibrated fine-vs-coarse contrast is $D_{\text{cal}} = +0.99$ for every CIFAR-100 system. On Sleep-EDF Expanded the null is flat ($\bar{Z}_{\text{null}} = 0.999$, $N_{\text{eff}} \approx 4000$), so no drift correction is needed there. The descriptor PCA of Fig. 8 has explained-variance ratios PC1 = 55.2%, PC2 = 25.4%.

C.4 Alternative local statistics

Established local statistics (kNN label entropy, Gini impurity, max-purity, trustworthiness/continuity, Fisher ratio) separate the coarse regimes equally well, and a *multiscale* kNN-entropy baseline reproduces SCF's depth profiles for well-organized systems (correlation 1.00; 0.65 for DINO-lite): the depth trends are generic to multiscale local entropy (App. C.4). What these baselines do not provide is composition fidelity (Table 1), the joint (d, Γ) quadrant structure, and the validity-gated, null-calibrated protocol with its estimator theory (Sec. 4).

Under the identical protocol (same features, reference, queries, bandwidth band), we recompute the layerwise analysis with: kNN label entropy (k=10, no smoothing), Gini impurity of the unsmoothed kernel posterior, max-purity, trustworthiness/continuity (Jeon et al., 2024), and a class-conditional Fisher ratio (this last uses true query labels, unlike SCF). Leave-one-out 1-NN regime classification over the 28 system-seed units (chance $\approx 1/3$): Zentropy 1.00, kNN entropy 1.00, max-purity 1.00, Fisher 1.00, Gini 0.96: the coarse regimes are visible to every local statistic, so regime separation alone does not justify SCF. What the simpler statistics do not provide is what the paper builds on: composition fidelity (Table 1; max-purity is provably blind to minority-class splits), the multi-granularity field (the quadrant structure, traversal, and the audio granularity reversal live in the (l, τ) plane, not in any scalar), and the estimator theory (Dirichlet floor, Schur ordering, null calibration). On the CE/CE-LS intervention, effect sizes $|\Delta|/\sigma_{\text{pool}}$ on the total change are: Zentropy 9.0, Gini 8.8, max-purity 8.2, Fisher 11.4, kNN entropy 2.3; Fisher is strongest but uses query labels and conflates geometry with semantics.

C.5 Reconfiguration depth and seed consistency (extended)

CE-based models reconfigure at mid-to-late layers ($L^* = 0.6$) while several self-supervised systems reconfigure near the input side ($L^* = 0$); L^* is an argmax over few layers (a coarse locator), and SupCon, though supervised, shares the early pattern, so no paradigm-level claim follows. Seed consistency is system-dependent (R_{SCF} from 0.50 to 0.97).

C.6 Per-point field and prediction errors

As a small downstream sanity signal, last-layer $\tilde{Z}$ on CIFAR-100 detects test-time misclassifications at AUROC 0.76–0.79 (CE and CE-LS, four seeds), below the model's own softmax confidence (0.85) but correlated with rather than identical to it (Spearman ≈ -0.7). We report this as an implication, not as a task.

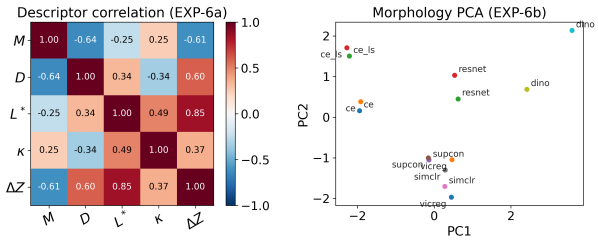

Figure 8: Reliability: (left) descriptor correlation matrix (no pair above $|\rho|$=0.9); (right) descriptor PCA (PC1 + PC2 = 80.6% variance), distinct fields occupy distinct regions.

C.7 PCA DETAILS

The descriptor PCA of Fig. 8 is computed on the standardized five-descriptor vectors of the 28 system-seed units; explained variance: PC1 62.8%, PC2 22.9% (App. C.3 discusses the null overlay).