Geometric Stability: The Missing Axis of Representations

Prashant C. Raju
rajuprashant@gmail.com

Significance Statement

Current methods for comparing representations in neural networks and biological systems measure similarity: whether two systems encode the same information. None measure whether that encoding is reliable. We introduce geometric stability, a distinct property quantifying how consistently a representation's distance structure holds under perturbation. A key mathematical property gives stability a complementary blind spot to existing metrics: it detects compression-induced damage that similarity cannot see. Across 2,463 representations spanning language models, vision systems, protein sequences, single-cell profiles, and neural recordings, stability and similarity share less than 0.1% of variance, an independence that holds for systems engineered and evolved, trained and recorded, optimized and simply measured. Geometric stability provides a universal diagnostic for representational reliability across computational and biological sciences.

Abstract

Representational similarity analysis and related methods have become standard tools for comparing the internal geometries of neural networks and biological systems. These methods measure what is represented, the alignment between two representational spaces, but not whether that structure is robust. We introduce geometric stability, a distinct dimension of representational quality that quantifies how reliably a representation's pairwise distance structure holds under perturbation. Our metric, Shesha, measures self-consistency through split-half correlation of representational dissimilarity matrices constructed from complementary feature subsets. A key formal property distinguishes stability from similarity: Shesha is not invariant to orthogonal transformations of the feature space, unlike Centered Kernel Alignment (CKA) and Procrustes distance, enabling it to detect compression-induced damage to manifold structure that similarity metrics cannot see. Spectral analysis reveals the mechanism: similarity metrics collapse after removing the top principal component, while stability retains sensitivity across the eigenspectrum. Across 2,463 encoder configurations in seven domains—language, vision, audio, video, protein sequences, molecular profiles, and neural population recordings—stability and similarity are empirically uncorrelated (Spearman $\rho = -0.01$, 95% CI $[-0.06, +0.03]$). A regime analysis shows this independence arises from opposing effects: geometry-preserving transformations make the metrics redundant, while compression makes them anti-correlated, canceling in aggregate. Applied to 94 pretrained vision models across six datasets, stability exposes a "geometric tax": DINOv2, the top-performing model for transfer learning, ranks last in geometric stability on five of six benchmarks. Contrastive alignment and hierarchical architecture predict stability, providing actionable guidance for model selection in deployment contexts where representational reliability matters.

Geometrical characterizations of internal representations have become central to understanding computation in both artificial and biological systems (Edelman 1998; Kriegeskorte, Mur, and Bandettini 2008). In neural networks, pairwise distance structures among stimulus-evoked activity patterns provide a model-agnostic summary of what a population code represents (Diedrichsen and Kriegeskorte 2017; Nili et al. 2014; Schütt 2025; Schütt et al. 2023; Walther et al. 2016), enabling systematic comparisons across architectures (Kornblith, Shlens, and Le 2019; T. Nguyen, Raghu, and Kornblith 2021; Schrimpf et al. 2018), training regimes (Mehrer et al. 2020; Zhuang et al. 2021), species (Kriegeskorte, Mur, Ruff, et al. 2008), and brain regions (Freiwald and Tsao 2010).

The same geometric perspective now pervades fields far beyond neuroscience. In computational biology, protein language models encode amino acid sequences as high-dimensional embeddings whose geometry predicts structure and function (Abramson et al. 2024; Ahdritz et al. 2024; Jumper et al. 2021; Z. Lin et al. 2023). In single-cell genomics, transcriptomic

profiles of individual cells(Zheng et al. 2017) define points in gene expression space whose pairwise distances(Luecken and Fabian J Theis 2019) reflect cell type identity(Wolf, Angerer, and Fabian J. Theis 2018), developmental trajectory (Trapnell et al. 2014), and perturbation response (Butler et al. 2018). In systems neuroscience, simultaneous recordings from hundreds of neurons (Siegle et al. 2021; Steinmetz et al. 2019) yield population activity vectors(Pandarinath et al. 2018; Saxena and Cunningham 2019) whose representational geometry (Kriegeskorte and Kievit 2013; Sorscher, Ganguli, and Sompolinsky 2022) encodes sensory stimulii (Ding et al. 2023; Nogueira et al. 2023), decisions (Mante et al. 2013), and motor plans(Churchland et al. 2012).

In each case, the analytical strategy is the same: abstract from the specific identity of measured features (genes, neurons, embedding dimensions) to characterize the information structure of a high-dimensional representation through its pairwise distance geometry. The universality of this strategy is what makes the question we raise here consequential: if there is a blind spot in how we evaluate representational geometry, it affects every field that relies on it. Across these domains, the core analytical paradigm is representational similarity: given two systems exposed to the same inputs, do they produce representations with the same pairwise distance structure? This paper is not about similarity between representations. It is about the stability within a single representation.

Representational similarity analysis (RSA Kriegeskorte, Mur, and Bandettini 2008) and Centered Kernel Alignment (CKA Kornblith, Norouzi, et al. 2019) exemplify the current paradigm. RSA compares representational dissimilarity matrices (RDMs) constructed from pairwise distances among response patterns, abstracting from the roles of individual neurons or feature dimensions to capture what a population code represents (Kriegeskorte, Mur, and Bandettini 2008; Nili et al. 2014).CKA provides a normalized measure of alignment between kernel matrices that is invariant to isotropic scaling and orthogonal transformation of the feature space (Kornblith, Norouzi, et al. 2019).Additional methods have enriched this framework: Singular Vector Canonical Correlation Analysis (SVCCA Raghu et al. 2017) and Projection Weighted CCA (PWCCA Morcos, Raghu, and S. Bengio 2018) compare representations through canonical subspace alignment; Procrustes analysis (Dryden and Mardia 1998; Masarotto, Panaretos, and Zemel 2018; Rohlf and Slice 1990; Schönemann 1966) measures alignment after optimal rotation. More recently, topological representational similarity analysis (tRSA (B. Lin and Kriegeskorte 2024)), was introduced, which demonstrates that geotopological summary statistics, which compress uninformative variation in very small and very large representational distances, provide more robust signatures of computational function across brain regions and deep neural network layers. By defining a family of nonlinear monotonic transforms of the RDM parameterized by lower and upper distance thresholds, tRSA enables researchers to calibrate the balance between geometric sensitivity and topological sensitivity, revealing that intermediate settings often outperform pure geometry for identifying corresponding brain regions across individuals. A recent work (Sucholutsky et al. 2025) surveyed this full landscape across cognitive science, neuroscience, and machine learning, proposing a unifying framework for representational alignment. Each of these methods, in its own way, answers the same question: do two representations encode similar structure? None answers whether that structure is reliable.

These methods share a common thread: they all measure the relationship *between* two representational spaces. Each takes as input a pair of representations and quantifies how well the pairwise distance structure of one predicts the other. An ideal characterization of a representation, however, should also address a logically prior question: is the structure within a single representational space reliable? A representation whose internal geometry rearranges when the measurement basis is perturbed, when features are resampled, inputs are shifted, or the encoding is projected onto a different subspace, may score well on every between-system comparison and still be fundamentally fragile.

Stability as an independent axis. The distinction between content and reliability is not new in measurement science. It recapitulates the classical separation of validity and reliability in psychometrics (Cohen 1988): a test may measure the right construct (high validity, analogous to high similarity) yet produce inconsistent scores across administrations (low reliability, analogous to low stability). A parallel principle appears in data science, where Yu and Kumbier (Yu and Kumbier 2020) argue that a result is trustworthy only if it is stable to reasonable perturbations of the data and analytical pipeline, establishing stability alongside predictability and computability as a foundational requirement for veridical inference. In the context of neural representations, a closely related idea was formalized through the noise ceiling (Nili et al. 2014), which uses split-half correlation across observations (trials or subjects) to bound how well any model could account for an empirical RDM given measurement noise. The noise ceiling is a diagnostic of data quality: it asks whether the measured geometry is replicable across independent observations of the same system. The question we pose is complementary. We ask not whether the measured geometry is replicable across *observations*, but whether the representational architecture distributes geometric information redundantly across the *feature space*, such that any sufficiently large random subset

of features recovers the same pairwise distance structure. This is a diagnostic of representational architecture, not data quality.

An analogy clarifies the distinction. Two libraries may hold identical collections: the same books, organized by the same classification scheme. But suppose one library maintains strict physical ordering, where nearby books remain topically related and retrieval degrades gracefully under minor displacements, while the other has brittle shelving, where a reader returning a few books to the wrong positions triggers a cascade that makes neighboring titles unreliable guides to content. A content audit, the analog of a similarity comparison, would confirm equivalence. However, a structural audit, the analog of a stability assessment, would reveal the difference. For deployed models, geometric stability measures exactly this resilience: whether the representational geometry that similarity metrics evaluate is itself robust to perturbation of the measurement basis. This practical urgency extends across domains. A vision model whose internal geometry is fragile may appear equivalent to a robust model by every similarity benchmark (CKA, RSA, Procrustes, tRSA), yet fail unpredictably under distribution shift (Kumar et al. 2022; Wortsman et al. 2022), adversarial perturbation (Ilyas et al. 2019), or post-training fine-tuning (Aghajanyan, Zettlemoyer, and S. Gupta 2020; Li et al. 2025). The fragility is invisible to similarity because similarity compares representations between systems, while fragility is a property within a single system.

We formalize geometric stability as a dimension of representational quality that is empirically distinct from representational similarity. Our metric, Shesha, named for the serpent deity of Hindu cosmology on whom Vishnu rests, representing the invariant remainder of the cosmos Daniélou 1964; Dimmitt and Buitenen 1978; Vogel 1995, quantifies the self-consistency of a representation's pairwise distance structure through split-half correlation of RDMs constructed from complementary random partitions of the feature space. The approach adapts the split-half methodology of noise ceiling estimation (Nili et al. 2014), but along a different axis and for a different purpose. Where the noise ceiling splits observations to assess data quality, Shesha splits features to assess representational architecture. Where the noise ceiling asks "given measurement noise, how well could any model predict this RDM?", Shesha asks "given the distribution of geometric information across feature dimensions, how reliably does this representation maintain its distance structure?" This reframing requires no repeated measurements, enabling stability assessment for systems, including pretrained embeddings, single-cell molecular profiles, and neural population recordings, where observation-level replication is unavailable or irrelevant.

A key formal property distinguishes stability from similarity and explains why the two must be distinct. CKA depends on the Gram matrix $XX^\top$. Any transformation that preserves $XX^\top$, including arbitrary orthogonal rotation of the feature space, is invisible to CKA. Shesha depends on per-axis structure: it splits feature dimensions into halves and checks whether each half recovers the same pairwise geometry. Transformations that redistribute geometric information across axes, including the very rotations CKA cannot see, change what each half captures and thus change Shesha. This creates a pair of metrics with complementary blind spots. CKA sees through rotations that Shesha detects; Shesha sees through compressions that CKA misses. The compression case is particularly consequential. PCA and related dimensionality reduction techniques concentrate geometric information into a few dominant components, leaving one random feature-split half uninformative. While CKA, which is dominated by the top eigenvalues, registers minimal change, Shesha drops sharply (**Fig. 2A**). This asymmetry is the mechanism of what we call the "geometric tax": the cost, paid in stability, of concentrating representational geometry into fewer dimensions to optimize downstream task performance.

Overview of results. We validate this framework across 2,463 encoder configurations spanning seven domains: language models, vision systems, audio and video encoders, protein sequence representations (Z. Lin et al. 2023), single-cell molecular profiles (Wolf, Angerer, and Fabian J. Theis 2018; Zheng et al. 2017), and neural population recordings from Neuropixels probes (Steinmetz et al. 2019). Across this breadth, four computational domains and three biological domains encompassing systems as different as GPT-2 hidden states and mouse V1 spike trains, stability and similarity are empirically uncorrelated (Spearman $\rho = -0.01$, 95% CI $[-0.06, +0.03]$), falling entirely within the negligible range ($|\rho| < 0.10$; (Cohen 1988)). This near-zero aggregate correlation is not noise masking a hidden relationship. A regime analysis reveals that geometry-preserving transformations (scaling, permutation) make the two metrics positively correlated ($\rho = +0.90$ to $+0.92$), while compression makes them negatively correlated ($\rho = -0.47$), and these opposing effects cancel in aggregate. The Johnson–Lindenstrauss lemma (Dasgupta and A. Gupta 2002; Johnson and Lindenstrauss 1984) explains the positive regime: random projections preserve pairwise distances, making both metrics redundant. PCA explains the negative regime: by concentrating all variance into a few components, it maximizes similarity while minimizing the distributional redundancy that stability requires (Tenenbaum, Silva, and Langford 2000). Spectral deletion experiments pinpoint the

mechanism empirically: CKA collapses to near-zero after removing the top principal component, while Shesha retains sensitivity across the first 26 components of the eigenspectrum. The two metrics probe different scales of the representational geometry.

Applied to 94 pretrained vision models on six benchmark datasets, this independence has a striking practical consequence. DINOv2 (Oquab et al. 2024), the model with the highest transfer learning performance (assessed via LogME; (You, Liu, Wang, et al. 2021)) on five of six benchmarks, ranks last or near last in geometric stability on those same benchmarks. This dissociation is not idiosyncratic: it reflects a systematic trade-off. Contrastive learning objectives (Radford et al. 2021), which train representations to concentrate discriminative information into a few dominant dimensions, produce high transferability but low stability. Hierarchical architectures that maintain information flow across scales produce higher stability. The geometric tax is invisible to every existing representational benchmark, including transferability assessments C. V. Nguyen et al. 2020; You, Liu, Wang, et al. 2021, visual task adaptation suites (Zhai et al. 2019), and holistic evaluation frameworks (Liang et al. 2023). It becomes visible only when stability and similarity are measured independently. The result suggests that the representations powering state-of-the-art transfer learning are optimized for content at the expense of robustness, a trade-off that current evaluation frameworks do not measure and therefore cannot manage.

The structure of this paper follows three movements. First, we establish that stability and similarity are formally distinct, using the complementary invariance properties of CKA and Shesha under rotation and compression to build a visual taxonomy of when each metric detects change and when it is blind (**Fig. 1**). This figure is designed to make the formal contribution self-contained: a reader who understands **Fig. 1** understands why the two metrics must be independent, without following an algebraic proof. Second, we demonstrate that this formal distinction holds empirically across seven domains, with a spectral deletion analysis revealing the mechanism at the level of the eigenspectrum and a regime analysis explaining exactly when the metrics agree, disagree, and cancel (**Fig. 2, Table 2**). Third, we show that the distinction has consequences: in pretrained vision models, the geometric tax dissociates the best models for transfer from the best models for reliability, exposing a trade-off that has been present but invisible throughout the model evaluation literature (**Fig. 3**). Together, these results establish geometric stability as a complementary axis for evaluating information encoding in artificial and biological systems. Evaluating what a representation encodes without measuring how reliably it encodes it leaves a critical blind spot in representational analysis. Geometric stability fills it.

Results

Geometric stability captures structure invisible to similarity metrics

Formal framework.

Let $X \in \mathbb{R}^{n \times d}$ denote a matrix of n samples with d-dimensional representations. A representational dissimilarity matrix (RDM) $D \in \mathbb{R}^{n \times n}$ captures pairwise dissimilarities, computed here as cosine distance:

$$D_{ij} = 1 - \frac{x_i^\top x_j}{\|x_i\| \, \|x_j\|}. \tag{1}$$

Shesha operates by constructing two RDMs from complementary views of X and measuring their agreement:

$$\text{Shesha}(X) = \rho_s\big(\text{vec}(D^{(1)}),\ \text{vec}(D^{(2)})\big), \tag{2}$$

where ρ_s denotes Spearman's rank correlation and $\text{vec}(\cdot)$ extracts the upper triangular elements. The choice of how to construct the complementary views $D^{(1)}$ and $D^{(2)}$ defines distinct Shesha variants (SI Appendix). The primary variant, Feature-Split Shesha ($\text{Shesha}_{\text{FS}}$), partitions feature dimensions $\{1, \ldots, d\}$ into two disjoint random halves $F_k^{(1)}, F_k^{(2)}$ and computes an RDM from each half, averaging over $K = 30$ random partitions for stable estimation:

$$\text{Shesha}_{\text{FS}}(X) = \frac{1}{K} \sum_{k=1}^{K} \rho_s\big(\text{vec}(D_{F_k^{(1)}}),\ \text{vec}(D_{F_k^{(2)}})\big). \tag{3}$$

$\text{Shesha}_{\text{FS}}$ answers a specific question about representational architecture: is the pairwise distance structure redundantly encoded across the feature space, such that any sufficiently large random subset of dimensions recovers the same geometry? A high score indicates that geometric information is distributed broadly across the basis; a low score indicates that

Table 1: **Invariance properties of stability and similarity metrics.** Shesha's non-invariance to orthogonal transformations is the formal mechanism by which it captures geometric properties invisible to CKA and Procrustes. [a]CKA depends only on $XX^\top$; [b]Procrustes explicitly optimizes over orthogonal alignment.

	Scale	Orthogonal rotation	Feature permutation	Monotonic distance	Isotropic scale
$\text{Shesha}_{\text{FS}}$	✓	×	✓	✓	✓
Linear CKA[a]	✓	✓	✓	×	✓
Procrustes[b]	✓	✓	✓	×	✓
PWCCA	×	×	×	×	✓

it is concentrated in a fragile subspace. This question is distinct from the one addressed by the noise ceiling in RSA Nili et al. 2014, which splits observations (trials or subjects) to bound data quality. Shesha splits features (neurons or embedding dimensions) to assess whether geometric structure is an architectural property of the representation rather than an artifact of how it was measured. The mathematical machinery is the same; the interpretive axis is orthogonal. A low noise ceiling diagnoses unreliable measurements. A low Shesha score diagnoses brittle geometry, which may reflect architectural choices (e.g., sparse coding), training dynamics (e.g., feature collapse), or intrinsic properties of the domain (e.g., high-dimensional but low-rank structure). The metric requires no labels and no repeated measurements, making it applicable to pretrained embeddings, single-cell molecular profiles, and neural population recordings where observation-level replication is unavailable or undefined.

Invariance properties: why stability and similarity diverge.

Table 1 and **Fig. 1** characterize the transformations under which Shesha and standard similarity metrics are invariant, following the framework of Kornblith et al. (Kornblith, Norouzi, et al. 2019). $\text{Shesha}_{\text{FS}}$ is invariant to global scaling (cosine distance normalizes magnitudes), monotonic distance transformations (Spearman correlation operates on ranks), and feature permutation (random equipartition is exchangeable over coordinates). These invariances ensure that Shesha is insensitive to nuisance variation in scale, distance metric choice, and arbitrary labeling of feature axes.

The critical property is what Shesha is *not* invariant to. If $Y = XQ$ for an orthogonal matrix Q, then $\text{Shesha}_{\text{FS}}(Y) \neq \text{Shesha}_{\text{FS}}(X)$ in general. The reason is transparent: Q redistributes geometric information across coordinate axes, altering what each feature-split half captures. But the Gram matrix is preserved ($YY^\top = XQQ^\top X^\top = XX^\top$), so CKA, which depends only on $XX^\top$ Kornblith, Norouzi, et al. 2019, is invariant: $\text{CKA}(X, Y) = \text{CKA}(X, X) = 1$. Procrustes distance, which explicitly optimizes over orthogonal alignment Schönemann 1966, is similarly invariant. This creates a pair of metrics with complementary blind spots (**Fig. 1**):

- *Rotation* (**Fig. 1b**): CKA is invariant; Shesha detects the redistribution of geometric information across the basis.
- *Compression* **(Fig. 1c)**: PCA concentrates all geometric information into a few dominant components. CKA, dominated by the top eigenvalues, registers minimal change. Shesha drops sharply, because one random feature-split half now contains most of the variance while the other contains mostly noise.
- *Scaling* (**Fig. 1a**) and *feature permutation* (**Fig. 1d**): both metrics are invariant, for different formal reasons.

The non-invariance to orthogonal transformations is the structural mechanism underlying the empirical dissociation reported below. Full proofs are in the **SI Appendix**; **Fig. 1** is designed to make the argument visual and self-contained.

Construct validation with known ground truth.

Before testing whether stability and similarity dissociate in real representations, we verified that Shesha measures what it claims to measure using synthetic representations with known properties. Three complementary tests establish construct validity, discriminant validity, and the mechanistic basis of the dissociation (full details in **SI Appendix**).

Sensitivity to known stability levels. We generated representations with parametrically controlled stability by mixing a low-rank signal component with isotropic noise:

$$X = \alpha \cdot \frac{ZW}{\|ZW\|_F} + (1 - \alpha) \cdot \epsilon, \tag{4}$$

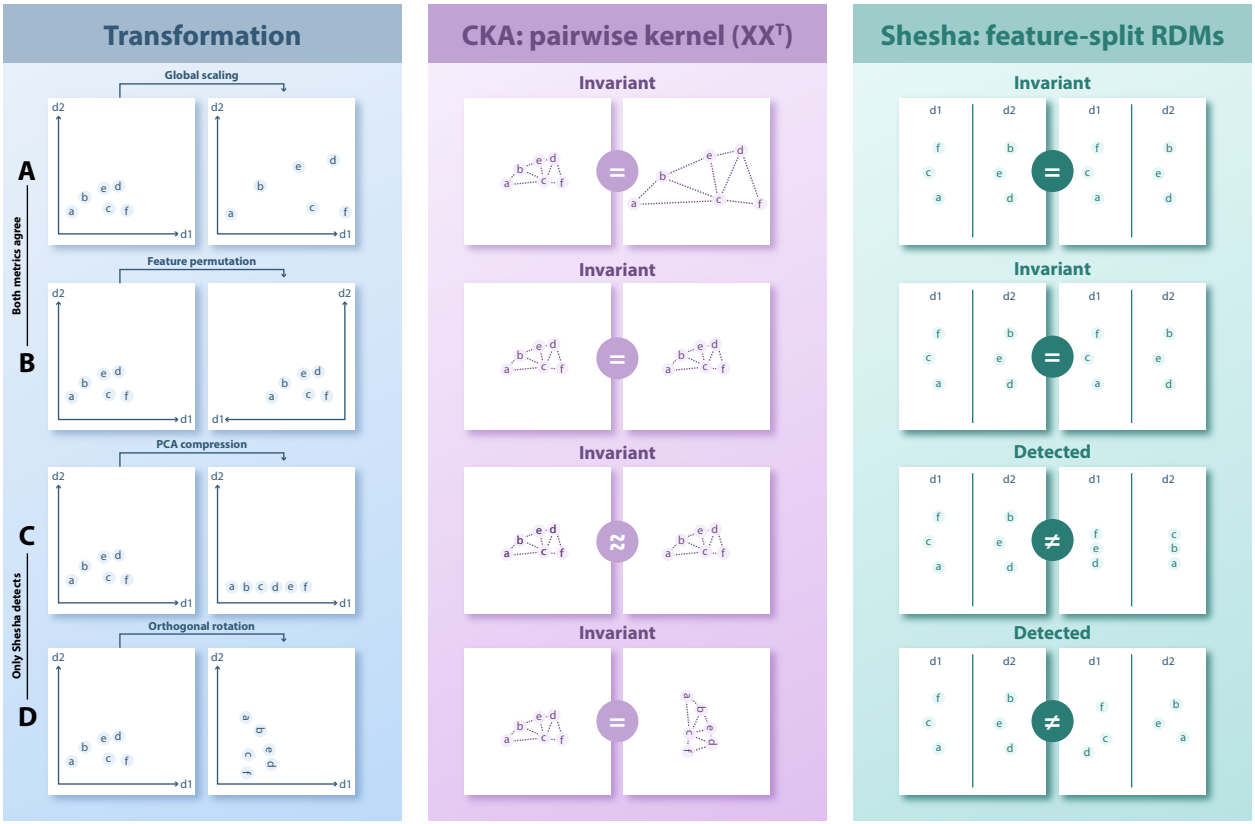

Figure 1: **CKA and Shesha have complementary blind spots under geometric transformations.** Each row applies a transformation to the same six-point representation, then shows how CKA (center) and Shesha (right) respond. CKA computes pairwise kernel alignment from the Gram matrix $XX^\top$; Shesha splits feature dimensions into two halves (dashed line) and compares the resulting RDMs. Green borders indicate the metric is unchanged; red borders indicate a detected change. **A. Global scaling** preserves cosine distances, leaving both metrics invariant. **B. Feature permutation** relabels coordinate axes without altering content; random equipartition is exchangeable over relabeled indices, so both metrics are invariant. Rows b and c reveal complementary blind spots: CKA is insensitive to how geometry is distributed across the representation's basis, while Shesha is sensitive to exactly this property. **C. PCA compression** retains dominant variance (CKA approximately unchanged) but concentrates all geometric information into fewer coordinates, collapsing one feature half to noise (Shesha drops). This is the mechanism underlying the geometric tax. **D. Orthogonal rotation** preserves $XX^\top$ (CKA unchanged) but redistributes geometric information across coordinate axes, altering which structure each feature half captures (Shesha detects the change). This is the key formal dissociation.

where $Z \in \mathbb{R}^{n \times k}$ is a latent matrix (n=200 samples, k=50 latent dimensions), $W \in \mathbb{R}^{k \times d}$ is a random projection (d=256 features), $\epsilon \sim \mathcal{N}(0, I)$ is isotropic noise, and $\alpha \in [0, 1]$ controls ground-truth stability. Across 21 levels from $\alpha = 0$ to $\alpha = 1$, Shesha recovered the ground truth with near-perfect accuracy (Spearman $\rho = 0.990$, $p < 10^{-86}$; Fig. S1). This confirms that the metric responds monotonically and with high fidelity to the underlying geometric consistency of a representation.

Four-quadrant dissociation. A metric that merely tracks signal strength would correlate positively with CKA, since high-signal representations tend to score highly on both. To establish that stability and similarity are separable properties, we constructed balanced samples (15 pairs each) from all four quadrants of the stability × similarity space: (Q1) high stability, high similarity (same latent structure with small noise; Shesha = 0.701 ± 0.003, CKA = 0.998 ± 0.000); (Q2) high stability, low similarity (independent high-signal representations; Shesha = 0.701 ± 0.004, CKA = 0.001 ± 0.010); (Q3) low stability, low similarity (independent noise; Shesha = 0.001 ± 0.003, CKA = −0.001 ± 0.010); and (Q4) low stability, high similarity (adversarial construction: aligned sample geometry with inconsistent feature-split structure; Shesha = −0.001 ± 0.005, CKA = 0.978 ± 0.000). Quadrant Q4 is the critical case: representations can exhibit CKA > 0.97 while Shesha is near zero, demonstrating that high similarity does not entail high stability. With balanced sampling across all four quadrants, the correlation between Shesha and debiased CKA (Song et al. 2012) was $\rho = 0.20$ (**Fig. 5**), confirming that these metrics assess largely different properties.

Spectral sensitivity: the mechanistic explanation. The invariance analysis (**Fig. 1**) predicts that CKA should be dominated by the top eigenvalues while Shesha should retain sensitivity across the spectral tail. We tested this prediction by generating representations with power-law eigenspectra (mimicking trained neural networks; $S_{ii} = 100/(i + 1)$) and progressively removing the top k principal components (**Fig. 2A**). All similarity metrics, debiased CKA, PWCKA, and Procrustes, collapsed below 0.4 after removing just the top principal component (k=1). Shesha remained above 0.4 until k=26 components were removed. At k=30, Shesha retained 110× higher signal than CKA. This divergence was robust across preprocessing conditions (raw, centered, normalized, whitened), though whitening caused CKA to recover sensitivity by artificially equalizing the spectrum (**Fig. 6**).

The spectral sensitivity experiment confirms the formal prediction: CKA tracks dominant variance concentrated in the top principal components, while Shesha measures whether geometric structure is distributed across the full eigenspectrum. These are different properties of the same representation, and they produce different diagnostic conclusions. A representation whose geometry resides entirely in a single dominant direction will score maximally on CKA (all pairwise distances are preserved in the Gram matrix) but minimally on Shesha (a random feature split has a 50% chance of placing that direction entirely in one half). Conversely, a representation whose geometry is distributed uniformly across all dimensions will score equivalently on both metrics. The divergence arises precisely when geometry is concentrated, and it is this concentration that the geometric tax (reported below) exploits.

Stability and similarity are distinct across seven domains

Cross-domain validation.

We validated the formal prediction of independence across seven representational domains spanning four computational and three biological systems (**Table 2**). For each domain, we extracted base representations from pretrained models or structured data sources and applied a standardized set of geometric interventions: PCA at various ranks, random projections at multiple target dimensions, top-variance feature selection, random feature subsets, Gaussian noise injection, and normalization variants (SI Appendix). This design produces 2,463 encoder configurations in total, each characterized by a $\text{Shesha}_{\text{FS}}$ score (geometric stability) and a debiased CKA score (similarity to domain-specific reference representations, averaged across three references per domain). All results were aggregated across 15 random seeds for estimation stability.

The machine learning domains comprise 448 configurations: language embeddings from sentence transformers (MiniLM, MPNet, DistilBERT, RoBERTa; N=127), vision embeddings from ViT, CLIP, DeiT, and ResNet50 (N=129), audio embeddings from Wav2Vec2 and HuBERT (N=64), and video embeddings from TimeSformer, VideoMAE, and frame-level encoders (N=128). The biological domains comprise 2,015 configurations: protein sequence encoders including compositional, spectral, and physicochemical features applied to Swiss-Prot human sequences (Z. Lin et al. 2023) (N=402), molecular embeddings of single-cell RNA-seq features from the pbmc3k dataset (Wolf, Angerer, and Fabian J. Theis 2018; Zheng

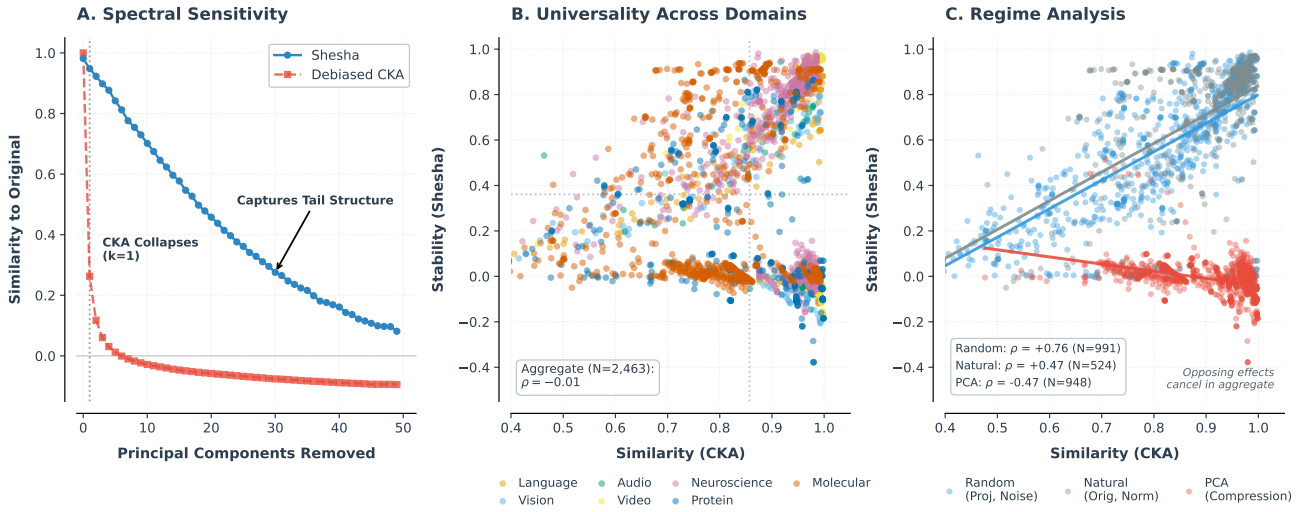

Figure 2: **Stability and similarity are independent dimensions of representational geometry. (a) Spectral Sensitivity:** CKA (red) collapses after removing just the single top principal component, while Shesha (blue) retains sensitivity to the spectral tail. CKA measures dominant variance; Shesha measures full manifold geometry. **(b) Universality:** Across 2,463 encoder configurations spanning seven domains, Shesha and CKA show negligible net correlation ($\rho = -0.01$, 95% CI $[-0.06, +0.03]$), confirming they capture distinct geometric properties. **(c) Regime Analysis:** Aggregate near-zero correlation emerges from opposing effects: random transformations yield positive correlation ($\rho = +0.76$), while PCA compression yields negative correlation ($\rho = -0.47$). These cancel in aggregate, revealing that Shesha specifically detects compression-induced damage invisible to CKA.

Table 2: **Domain-level correlations between stability and similarity.** Aggregate correlation is negligible ($\rho = -0.01$, CI within ± 0.06); four domains show negligible correlations ($|\rho| < 0.10$). [a]Protein shows moderate negative correlation driven by PCA on low-dimensional sequence encoders (20–500 dims); see SI Appendix.

Domain	N	ρ	95% CI	p
Machine Learning				
Language	127	+0.03	[−0.18, +0.24]	0.77
Vision	129	−0.03	[−0.23, +0.18]	0.72
Audio	64	−0.26	[−0.52, +0.02]	0.04
Video	128	−0.24	[−0.45, −0.02]	0.006
Biology				
Neuroscience	846	+0.01	[−0.06, +0.09]	0.67
Protein[a]	402	−0.36	[−0.45, −0.28]	<0.001
Molecular	767	+0.06	[−0.02, +0.13]	0.13
Aggregate	**2463**	**−0.01**	**[−0.06, +0.03]**	**0.57**

et al. 2017) (N=767), and neural population activity representations extracted from the Steinmetz recordings (Steinmetz et al. 2019) containing 26 sessions across multiple brain areas (N=846).

Aggregate distinctness.

The aggregate analysis yields $\rho = -0.01$ (95% CI $[-0.06, +0.03]$), a correlation indistinguishable from zero ($p = 0.57$) and entirely within the negligible range ($|\rho| < 0.10$; Cohen 1988). Stability and similarity share less than 0.1% of variance ($R^2 < 0.001$). Six of seven domains show distinctness ($|\rho| < 0.30$); four show negligible correlations ($|\rho| < 0.10$). The highest-powered domain, neural population recordings (N=846), provides the strongest individual evidence for distinctness ($\rho = +0.01$, tightest CI). Only the protein domain shows a moderate correlation ($\rho = -0.36$), which is driven specifically by PCA's interaction with low-dimensional sequence encoders (20 to 500 dimensions); when PCA configurations are excluded, the protein domain correlation drops to $\rho = +0.15$ (CI $[-0.00, +0.30]$; **SI Appendix**).

To control for dependencies among encoder configurations derived from the same base model (e.g., multiple perturbed versions of ResNet50), we fitted a linear mixed-effects model (Stability ~ Similarity + (1 | BaseModel)). The fixed effect of similarity on stability remains negligible ($\beta = 0.10$, 95% CI $[0.06, 0.15]$), and the intraclass correlation coefficient (ICC = 0.10) indicates that base model identity explains less than 10% of variance in stability scores. The remaining 90% is attributable to encoder-specific properties and residual variation, confirming that the distinctness is not an artifact of model family clustering.

The cancellation mechanism.

The aggregate near-zero correlation is not noise. It arises from structured opposing effects across transformation regimes (**Fig. 2C**).

Regime I: Geometry-preserving transformations (redundant). Random projections ($\rho = +0.90$) and feature selection ($\rho = +0.92$) yield near-perfect positive correlation between stability and similarity. This follows from the Johnson–Lindenstrauss lemma (Dasgupta and A. Gupta 2002; Johnson and Lindenstrauss 1984): random projections approximately preserve all pairwise distances, maintaining both the Gram matrix structure that CKA tracks and the per-axis distributional redundancy that Shesha tracks. Noise injection behaves similarly ($\rho = +0.58$). In this regime, the two metrics are redundant and Shesha adds no unique diagnostic value beyond CKA.

Regime II: Compression (dissociated). PCA yields a strong negative correlation ($\rho = -0.47$). Dimensionality reduction preserves dominant variance (maintaining high CKA) while concentrating all geometric information into a few components and discarding fine-grained manifold structure (reducing Shesha; Tenenbaum, Silva, and Langford 2000). This is the regime predicted by the invariance analysis (**Fig. 1c**) and confirmed by the spectral deletion experiment (**Fig. 2A**): CKA indexes on principal subspaces, while Shesha indexes on the full eigenspectrum. In this regime, Shesha captures geometric damage that CKA, by construction, cannot detect.

Regime III: Natural encoders (complementary). Original, untransformed representations show weak positive correlations ($\rho \approx +0.31$ to $+0.34$). Shesha contributes approximately 90% unique information beyond CKA ($1 - \rho^2 \approx 0.90$). Real-world encoders occupy this intermediate regime, consistent with the manifold hypothesis Y. Bengio, Courville, and Vincent 2013: learned representations balance geometry-preserving and compressive operations, placing them between the fully redundant and fully dissociated regimes. Both metrics offer complementary diagnostic value in this setting.

These opposing effects (positive under geometry-preserving transforms, negative under compression) cancel in aggregate, producing the near-zero correlation as a structured consequence of complementary sensitivities rather than an absence of relationship (**Fig. 2C**). This decomposition clarifies precisely when Shesha adds value: it detects compression-induced damage to manifold structure that remains invisible to similarity metrics optimized for dominant variance.

Robustness to metric.

The distinctness result is robust across multiple dimensions. Excluding any single domain preserves $|\rho| < 0.10$ in aggregate (Table S4, SI Appendix). Restricting to transformer-based domains only (N=448) yields $\rho = -0.05$ (CI $[-0.16, +0.07]$); restricting to biological domains only (N=2,015) yields $\rho = +0.01$ (CI $[-0.04, +0.06]$). To verify that the result generalizes beyond CKA, we evaluated two alternative similarity metrics in the language domain (N=127): effective-rank Projection-Weighted CKA (PWCKA; $\rho = -0.22$, $p = 0.012$) and Procrustes similarity ($\rho = +0.28$, $p = 0.001$). All three similarity metrics maintain $|\rho| < 0.30$ with Shesha, confirming that the distinctness is a property of the stability-similarity relationship itself, not an idiosyncrasy of CKA (Fig. S6).

Stability extends to biological representations

The biological domains in **Table 2** deserve brief interpretive comment, as they demonstrate that geometric stability is not restricted to engineered computational systems.

In the protein domain (N=402), sequence encoders ranging from compositional amino acid frequencies to physicochemical feature vectors applied to Swiss-Prot sequences (Bateman et al. 2022) show distinctness that is strongest for high-dimensional encoders and weakest under aggressive PCA compression. The moderate aggregate correlation ($\rho = -0.36$)

Table 3: **The DINOv2 paradox.** DINOv2 family averages across six datasets. DINOv2 achieves the highest LogME on multiple datasets while ranking last or near-last in $\text{Shesha}_{\text{FS}}$, except on EuroSAT.

Dataset	LogME	LogME Rank	$\text{Shesha}_{\text{FS}}$	FS Rank
CIFAR-10	1.020	1/29	0.378	29/29
CIFAR-100	1.360	1/29	0.273	28/29
Flowers-102	2.466	1/29	0.337	29/29
DTD	0.878	9/29	0.466	24/29
EuroSAT	0.572	2/29	0.948	3/29
Oxford Pets	1.280	9/29	0.530	25/29

reflects the prevalence of PCA configurations among low-dimensional sequence encoders (20–500 features), where compression disproportionately concentrates geometric structure.

In the molecular domain (N=767), single-cell RNA-seq profiles from the pbmc3k dataset (Zheng et al. 2017) show negligible correlation between stability and similarity ($\rho = +0.06$). This is notable because single-cell transcriptomic representations are high-dimensional (~2,000 highly variable genes) yet exhibit the same independence pattern as language and vision embeddings with far more parameters.

In the neural population recording domain (N=846), Steinmetz Neuropixels recordings (Steinmetz et al. 2019) across 26 sessions (each with ≥20 simultaneously recorded neurons and ≥50 trials) produce the tightest confidence interval of any domain ($\rho = +0.01$, CI $[-0.06, +0.09]$). The near-zero correlation holds despite substantial variation in recording quality, brain region, and session-level properties, providing the strongest individual evidence that stability and similarity are distinct dimensions of representational geometry.

Together, the biological results confirm that the formal distinction between stability and similarity is not an artifact of computational architectures or training objectives. It holds for systems that were not designed but evolved. Not trained but recorded. And not optimized but simply measured.

A geometric tax in pretrained vision models

The independence of stability and similarity established above raises a practical question: does the distinction matter for model selection? We tested this by evaluating 94 pretrained vision models on six benchmark datasets spanning four visual domains: natural images (CIFAR-10, CIFAR-100 (Krizhevsky 2009)), fine-grained recognition (Flowers-102 (Nilsback and Zisserman 2008), Oxford Pets Parkhi et al. 2012), texture (DTD (Cimpoi et al. 2014)), and remote sensing (EuroSAT (Helber et al. 2018)). The model selection covers four axes of architectural variation: training objectives (supervised, self-supervised, contrastive, generative), architectural families (columnar transformers, hierarchical transformers, hybrids, convolutional networks), model scales (from MobileNetV3-Small to ViT-Giant/14), and training paradigms (standard, distillation, augmentation, foundation model pretraining). For each model and dataset, we computed $\text{Shesha}_{\text{FS}}$ (geometric stability) from penultimate-layer features and LogME (You, Liu, Wang, et al. 2021) (transferability). Models were grouped into 29 architectural families for aggregate analysis (**SI Appendix**).

The DINOv2 paradox.

DINOv2 (Oquab et al. 2024) achieves the highest family-level LogME on four of six datasets (CIFAR-10, CIFAR-100, Flowers-102, EuroSAT) and ranks in the top ten on the remaining two. On the same benchmarks, DINOv2 ranks last or near-last in geometric stability on five of six datasets (**Fig. 3A**, **Table 3**). On CIFAR-10: LogME rank 1/29 families, $\text{Shesha}_{\text{FS}}$ rank 29/29. On CIFAR-100: LogME rank 1/29, $\text{Shesha}_{\text{FS}}$ rank 28/29. On Flowers-102: LogME rank 1/29, $\text{Shesha}_{\text{FS}}$ rank 29/29. The pattern is not limited to family averages: at the individual model level, DINOv2-giant ranks 1/94 in LogME but 88–92/94 in $\text{Shesha}_{\text{FS}}$ across CIFAR-10, CIFAR-100, and Flowers-102 (**Table 5**).

The sole exception is EuroSAT, where DINOv2 achieves both high transfer (LogME rank 2/29) and high stability ($\text{Shesha}_{\text{FS}}$ = 0.948, rank 3/29).This exception is informative: EuroSAT consists of satellite imagery with relatively uniform texture distributions, a domain where self-distillation may preserve geometric redundancy that is sacrificed in more heterogeneous visual domains. On all other datasets, the dissociation is complete. A model can be the best choice for transfer learning and simultaneously the worst choice for geometric reliability.

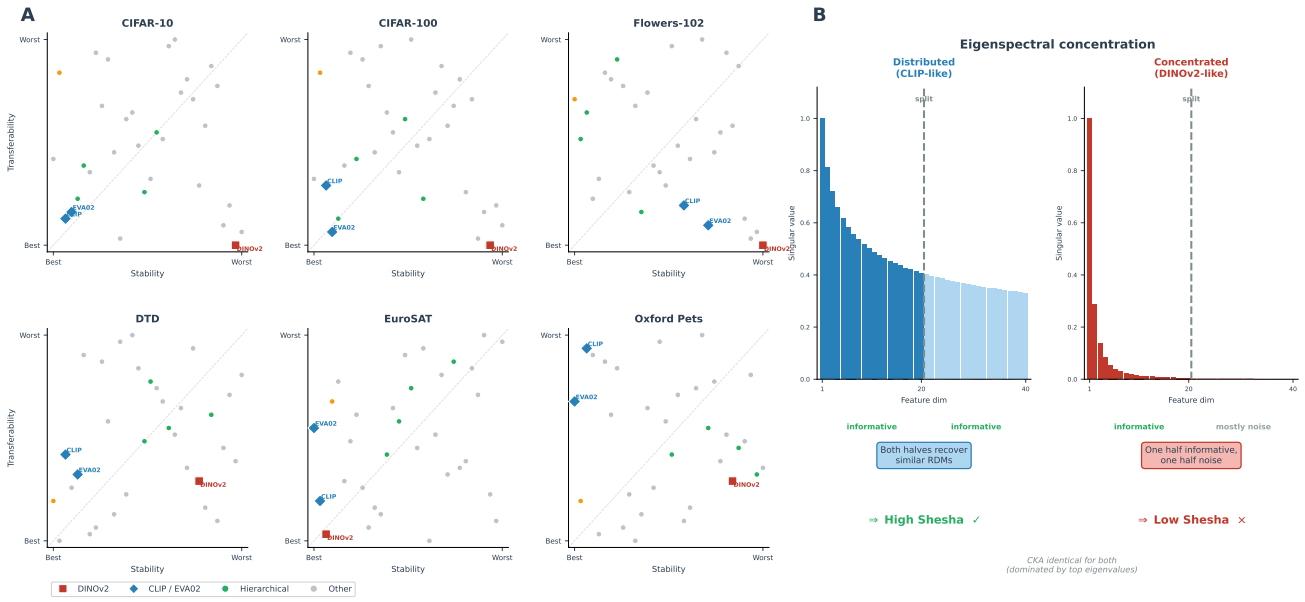

Figure 3: **The geometric tax in pretrained vision models.** (A) Family-level stability rank (Shesha-FS) versus transferability rank (LogME) for 32 architectural families across six datasets spanning four visual domains: natural images (CIFAR-10, CIFAR-100), fine-grained recognition (Flowers-102, Oxford Pets), texture (DTD), and remote sensing (EuroSAT). Each point represents one family; rank 0 is best. DINOv2 (red) achieves the highest transferability on 4/6 datasets while ranking last or near-last in geometric stability on 5/6, revealing a previously invisible trade-off. CLIP and EVA02 (blue) maintain high stability with competitive transfer. The sole exception is EuroSAT, where DINOv2 achieves both high transfer and high stability, possibly because uniform texture distributions align with the self-distillation objective. (B) The mechanism: eigenspectral concentration explains the dissociation. A representation with distributed singular values (left, CLIP-like) yields two informative feature-split halves, producing high Shesha. A representation with concentrated singular values (right, DINOv2-like) yields one informative and one uninformative half, producing low Shesha. Both representations produce identical CKA because CKA operates on the Gram matrix, which is dominated by the top eigenvalues regardless of how the remaining variance is distributed across coordinates. The geometric tax is the cost of concentrating representational geometry into fewer dimensions to maximize downstream adaptability.

The mechanism: eigenspectral concentration.

The DINOv2 paradox is not an isolated curiosity; it is a direct consequence of the compression mechanism identified in the regime analysis (Fig. 2C) and the invariance properties (**Fig. 1c**). DINOv2's self-distillation training objective concentrates representational geometry into a few dominant dimensions, producing features that are rich and highly adaptable (high LogME) but that fail the feature-split test because a random partition of dimensions has a high probability of placing most geometric information in one half and mostly noise in the other (low $\text{Shesha}_{\text{FS}}$). This is precisely the compression regime where CKA and Shesha dissociate: the Gram matrix is dominated by the top eigenvalues (high CKA, high LogME), while the per-axis distributional redundancy that Shesha measures is destroyed.

Fig. 3B illustrates this mechanism schematically. A representation with a flat eigenspectrum (distributed geometry, as in CLIP) yields two informative halves under any random feature split, producing high Shesha. A representation with a steep, power-law eigenspectrum (concentrated geometry, as in DINOv2) yields one informative and one uninformative half, producing low Shesha. Both representations may achieve identical CKA with a reference, because CKA is dominated by the top eigenvalues that both share. The geometric tax is the cost of this concentration: features optimized for downstream adaptability pay a price in structural redundancy that is invisible to every transferability metric but measurable through stability.

Training objectives and architecture determine stability.

Two factors consistently predict geometric stability across datasets (**SI Appendix**, **Tables 6–7**).

Contrastive alignment. CLIP models (Radford et al. 2021) outperform self-supervised alternatives on four of six datasets (Mann–Whitney $p < 0.05$; **Table 6**). The stability advantage of contrastive training is substantial: on CIFAR-100, CLIP achieves $\text{Shesha}_{\text{FS}} = 0.83 \pm 0.06$ compared to 0.48 ± 0.24 for self-supervised models ($\Delta = +0.34$). EVA-02 achieves the highest stability of any family by reconstructing CLIP features rather than raw pixels, confirming that alignment targets, not training mechanisms, determine geometric structure. The consistent family rankings (CLIP, EVA, and Inception families maintaining high stability; DINOv2, DeiT3, and ViT families showing low stability) suggest that the geometric tax is a systematic property of training objectives, not a stochastic artifact of individual model runs.

Hierarchical architecture. Hierarchical transformers (Swin, PVT, CoAtNet; n=18) outperform columnar architectures (ViT, DeiT; n=23) on CIFAR-10 ($\Delta = +0.12$, $p = 0.011$), CIFAR-100 ($\Delta = +0.15$, $p = 0.007$), and Flowers-102 ($\Delta = +0.26$, $p < 0.001$; **Table 7**). Multi-scale processing, which maintains information flow across spatial resolutions rather than compressing it into a single token sequence, acts as implicit geometric regularization. The advantage disappears on DTD and EuroSAT, where the visual structure is more spatially uniform.

Stability is an intrinsic architectural property.

A model's geometric stability is not task-specific; it generalizes across datasets within a visual domain. The cross-dataset rank correlation between CIFAR-10 and CIFAR-100 (which share the same image distribution but differ in the number of classes) is $\rho = 0.92$ for $\text{Shesha}_{\text{FS}}$, indicating that stability rankings are highly reproducible across task complexity levels (Fig. S8). Cross-domain rank correlations are weaker but consistently positive, confirming that stability reflects an intrinsic property of the representational architecture rather than a task-dependent interaction.

Practical implications.

The geometric tax reframes model selection as a two-dimensional problem. Current practice ranks models by transferability alone, using metrics such as LogME (You, Liu, Wang, et al. 2021; You, Liu, Zhang, et al. 2022), LEEP (C. V. Nguyen et al. 2020), or downstream accuracy on adaptation benchmarks Zhai et al. 2019. This implicitly assumes that the best model for transfer is also the best model for deployment. The stability-transferability dissociation shows that this assumption is false: the model that adapts most effectively to downstream tasks (DINOv2) is also the model whose internal geometry is most fragile under perturbation of the feature basis.

For deployment contexts where representational reliability matters, such as safety-critical applications, zero-shot inference without fine-tuning, or multi-task pipelines where the same features serve diverse downstream heads, stability provides a complementary selection criterion. CLIP and EVA families offer a concrete alternative: their contrastive alignment

produces representations that score highly on both axes, avoiding the worst of the geometric tax while retaining competitive transfer performance. Whether stability-aware training objectives can reduce the tax further, achieving DINOv2-level transferability without sacrificing geometric redundancy, remains an open question.

Discussion

Representational analysis has operated along a single axis: similarity. RSA, CKA, SVCCA, PWCCA, Procrustes analysis, and the recent topological extension tRSA (B. Lin and Kriegeskorte 2024) all answer the same class of question, whether two representations encode the same pairwise structure, while varying in their sensitivity to geometric versus topological features of that structure. The results presented here establish a second, independent axis: stability, which asks whether a single representation's pairwise structure is robust to perturbation of the measurement basis. These axes are formally distinct (complementary invariance properties under rotation and compression), empirically independent (Spearman $\rho = -0.01$ across 2,463 configurations in seven domains), and practically consequential (the geometric tax dissociates the best models for transfer from the best models for reliability). Together, similarity and stability span a two-dimensional diagnostic space for representational quality. Existing tools have been mapping one dimension of this space; the second has been invisible.

Existing tools have been mapping one dimension of this space; the second has been invisible. Recent work independently underscores that global alignment measures leave important structure uncharacterized: Muttenthaler et al. (Muttenthaler et al. 2025) demonstrate that vision models fail to capture human-like hierarchical abstraction despite high overall alignment scores, while Mahner et al. (Mahner et al. 2025) show that the latent dimensions underlying human and DNN similarity judgments diverge in ways scalar measures cannot detect. Geometric stability exposes a distinct gap: representations may share the same content, organized along similar dimensions, yet differ in whether that organization is robust to perturbation of the measurement basis.

The analogy to RSA's own intellectual trajectory is instructive. Kriegeskorte et al. (Kriegeskorte, Mur, and Bandettini 2008) abstracted from individual neurons to pairwise dissimilarity matrices, enabling cross-system comparison. Sch"utt et al. (Schütt et al. 2023) developed statistical inference methods for comparing representational geometries. Lin and Kriegeskorte (B. Lin and Kriegeskorte 2024) abstracted further, from geometry to topology, compressing uninformative distance variation to focus on neighborhood relationships. Each step refined the characterization of *what* is represented. Geometric stability adds a different kind of axis entirely: not a further abstraction of content, but a measure of *how reliably* that content is encoded. Shesha and RSA/CKA are jointly informative, not competing. A representation that scores highly on both axes encodes the right content and does so robustly; a representation that scores highly on one but not the other has a diagnostic blind spot that neither axis alone can identify.

The geometric tax as a design principle.

The DINOv2 paradox is not a curiosity specific to one model family. It reveals that current training objectives implicitly trade stability for transferability, and that this trade-off is invisible to every benchmark in the current evaluation ecosystem, including LogME (You, Liu, Wang, et al. 2021; You, Liu, Zhang, et al. 2022), LEEP (C. V. Nguyen et al. 2020), visual task adaptation suites (Zhai et al. 2019), and holistic evaluation frameworks Liang et al. 2023. The mechanism is clear: self-distillation and related objectives concentrate representational geometry into a few dominant dimensions to maximize downstream adaptability, paying a cost in distributional redundancy that only becomes visible when features are probed along the coordinate basis rather than through the Gram matrix. This is the compression regime of the regime analysis, now observed in the wild rather than under controlled geometric interventions.

Making the tax visible enables principled model selection. For transfer learning pipelines where a model will be fine-tuned on labeled downstream data, transferability (LogME, LEEP) remains the appropriate criterion; the geometric tax is acceptable because fine-tuning can reshape the representation to suit the target task. For deployment contexts where representational reliability matters directly, such as safety-critical applications, zero-shot inference, multi-task pipelines, or settings where the same features serve diverse downstream heads without further adaptation, stability provides a complementary selection criterion. Whether stability-aware training objectives can reduce the tax, achieving DINOv2-level transferability without sacrificing geometric redundancy, remains an important open question. The observation that contrastive alignment (CLIP) and hierarchical architecture (Swin) independently predict higher stability suggests that the tax is not an inevitable consequence of learning powerful representations, but a contingent property of specific training

regimes that can, in principle, be optimized.

When does stability add value?

The regime analysis provides explicit guidance. Under geometry-preserving operations (random projection, feature selection, noise injection), stability and similarity are highly correlated ($\rho > +0.58$) and Shesha adds no unique diagnostic information beyond CKA. Under compression (PCA, dimensionality reduction), the two metrics are anti-correlated ($\rho = -0.47$) and Shesha captures manifold damage that CKA, by construction, cannot detect. For natural, untransformed representations, the metrics are weakly correlated ($\rho \approx +0.31$ to $+0.34$) and Shesha contributes approximately 90% unique variance ($1 - \rho^2 \approx 0.90$). The practical implication is that stability is most informative precisely when representations have undergone compression-like transformations, whether through explicit dimensionality reduction, through training objectives that concentrate features, or through post-training interventions such as quantization or pruning that reduce the effective dimensionality of the representation.

The framework established here naturally suggests several extensions. The unsupervised, feature-split variant of Shesha studied in this paper measures intrinsic geometric redundancy without reference to any downstream task. Supervised variants that condition the split-half assessment on task-relevant structure could address settings where intrinsic stability alone may be insufficient, such as predicting whether a representation will support linear intervention along a specific semantic direction, or assessing whether post-training alignment has preserved task-relevant geometry. Similarly, the split-half principle can be adapted beyond feature partitions: splitting observations rather than features yields a measure of representational robustness to input variation, while splitting across time points yields a measure of temporal drift. In biological systems, where perturbation experiments produce populations of cells responding to the same intervention, an analogous construction measuring the directional coherence of cellular responses could quantify whether a genetic perturbation engages a robust regulatory program or scatters cells stochastically. These directions share a common principle: geometric stability, whether measured across features, observations, time, or perturbation responses, provides an axis of representational quality that is orthogonal to the content-focused metrics that currently dominate the field.

Boundaries and limitations.

Several boundaries of the present work deserve explicit statement.

First, Shesha is a global metric: it operates on the full representational dissimilarity matrix and returns a single scalar characterizing the entire geometry. It may miss localized instabilities, such as fragile subspaces within an otherwise stable representation or token-level dynamics in language models where stability varies across sequence positions. Per-subspace, per-layer, and per-token variants are natural extensions.

Second, the vision analysis reported here uses a single random seed (320) for feature extraction due to computational constraints. Preliminary runs with additional seeds showed consistent family rankings, and the cross-domain analysis averages over 15 seeds, but single-seed results should be interpreted with appropriate caution.

Third, the cross-domain validation relies on controlled encoder transformations (PCA, random projection, noise injection, normalization) applied to base representations. These interventions are designed to test metric behavior under known geometric operations, not to simulate the full range of natural model variation. The regime analysis clarifies which results generalize to untransformed representations (Regime III) and which are specific to controlled interventions (Regimes I and II).

Fourth, the computational cost of Shesha is nontrivial: K=30 split-half iterations each require construction of an $n \times n$ RDM, making the metric $O(Kn^2d)$ in time. For the sample sizes used here ($n \leq 5{,}000$), this is manageable; for larger datasets, subsampling to $n_{\max} = 1{,}600$ yields estimates with mean absolute deviation below 0.008 from full-sample values (**SI Appendix**).

Finally, the present work establishes that unsupervised stability and similarity are distinct and that the distinction has consequences for vision model evaluation. It does not establish a causal relationship between stability and downstream performance; the geometric tax is an observed dissociation, not a proven mechanism of failure. Whether low stability causes fragile deployment behavior, or merely correlates with architectural properties that independently cause fragility, is a question for future interventional studies.

Conclusion.

Evaluating what a representation encodes without measuring how reliably it encodes it leaves a critical blind spot in representational analysis. The tools developed over the past fifteen years for comparing representational geometries, from RSA through CKA to tRSA, have focused on content: whether two systems represent the same information. Geometric stability adds the complementary dimension of reliability: whether a single system's representation is structurally robust. Across seven domains spanning artificial and biological systems, these dimensions are empirically independent. In pretrained vision models, this independence exposes a previously invisible trade-off between transferability and structural integrity. By measuring how consistently representations maintain their pairwise distance structure rather than what that structure encodes, geometric stability provides a necessary complement to similarity for auditing the representations on which both computational and biological systems rely.

Materials and Methods

The detailed description of all analyses, data sources, preprocessing protocols, and statistical procedures is reported in **SI Appendix**. Briefly, geometric stability was quantified using Feature-Split Shesha ($\text{Shesha}_{\text{FS}}$), which partitions feature dimensions into random complementary halves, computes cosine-distance representational dissimilarity matrices (RDMs) from each half, and correlates the vectorized upper triangles via Spearman rank correlation, averaged over K=30 random partitions (**SI Appendix, Shesha Computation**). Representational similarity was measured using debiased linear CKA with the unbiased HSIC estimator (Kornblith, Norouzi, et al. 2019; Song et al. 2012), computed against three domain-specific reference representations per encoder configuration and averaged (**SI Appendix, Similarity Metrics**). Cross-domain validation spanned seven domains (language, vision, audio, video, protein, molecular, neural) comprising 2,463 encoder configurations generated by applying standardized geometric interventions (PCA, random projection, feature selection, noise injection, normalization) to base representations from pretrained models or structured data sources (**SI Appendix, Data Sources and Encoder Transformations**). All results were aggregated across 15 random seeds; all computations used Float64 precision and fixed seed 320 for reproducibility. The vision benchmark evaluated 94 pretrained models across six datasets (CIFAR-10, CIFAR-100, Flowers-102, Oxford Pets, DTD, EuroSAT), with transferability assessed via LogME (You, Liu, Wang, et al. 2021; You, Liu, Zhang, et al. 2022) and models grouped into 29 architectural families (**SI Appendix, Vision Benchmark**). Distinctness was assessed via Spearman correlation with 10,000 bootstrap replicates; mixed-effects models controlled for base model identity; architectural comparisons used Mann–Whitney U tests with exact p-values reported (**SI Appendix, Statistical Methods**).

Code Availability

All custom code is available on GitHub (https://github.com/prashantcraju/geometric-stability). We have also released an open source Python library through PYPI (shesha-geometry) Raju 2026.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature* 630.8016 (May 2024), pp. 493–500. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07487-w. URL: http://dx.doi.org/10.1038/s41586-024-07487-w.

[2] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. "Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning". In: *Annual Meeting of the Association for Computational Linguistics*. 2020.

[3] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Peter K Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. "OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization". In: *Nature Methods* 21.8 (May 2024), pp. 1514–1524. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02272-z. URL: http://dx.doi.org/10.1038/s41592-024-02272-z.

[4] Scott Allyn. *Jellyfish video*. https://test-videos.co.uk/vids/jellyfish/mp4/h264/360/Jellyfish_360_10s_1MB.mp4. 360p resolution version, with a duration of 10 seconds. 2016.

[5] Alex Bateman et al. "UniProt: the Universal Protein Knowledgebase in 2023". In: *Nucleic Acids Research* (2022).

[6] Y. Bengio, A. Courville, and P. Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828. ISSN: 2160-9292. DOI: 10.1109/tpami.2013.50.

[7] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". In: *Nature Biotechnology* 36.5 (Apr. 2018), pp. 411–420. ISSN: 1546-1696. DOI: 10.1038/nbt.4096. URL: http://dx.doi.org/10.1038/nbt.4096.

[8] Mark M. Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. "Neural population dynamics during reaching". In: *Nature* 487.7405 (June 2012), pp. 51–56. ISSN: 1476-4687. DOI: 10.1038/nature11129. URL: http://dx.doi.org/10.1038/nature11129.

[9] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. "Describing Textures in the Wild". In: *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[10] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2nd. Routledge Member of the Taylor and Francis Group, Aug. 1988. ISBN: 978-0805802832.

[11] Alain Daniélou. *Hindu Polytheism*. Bollingen Series. Later republished as 'The Myths and Gods of India'. Princeton University Press, Mar. 1964. ISBN: 978-0691097459.

[12] Sanjoy Dasgupta and Anupam Gupta. "An elementary proof of a theorem of Johnson and Lindenstrauss". In: *Random Structures & Algorithms* 22.1 (2002), pp. 60–65. ISSN: 1098-2418. DOI: 10.1002/rsa.10073.

[13] Jörn Diedrichsen and Nikolaus Kriegeskorte. "Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis". In: *PLOS Computational Biology* 13.4 (2017), e1005508. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005508.

[14] Cornelia Dimmitt and Johannes Adrianus Bernardus van Buitenen. *Classical Hindu Mythology: A Reader in the Sanskrit Puranas*. Philadelphia, PA: Temple University Press, 1978. ISBN: 978-0877221227.

[15] Xuehao Ding, Dongsoo Lee, Joshua Brendan Melander, George Sivulka, Surya Ganguli, and Stephen Baccus. "Information Geometry of the Retinal Representation Manifold". In: *Advances in Neural Information Processing Systems*. 2023.

[16] I L Dryden and K V Mardia. *Statistical analysis of shape*. Wiley Series in Probability and Statistics. Chichester, England: John Wiley & Sons, July 1998.

[17] Shimon Edelman. "Representation is representation of similarities". In: *Behavioral and Brain Sciences* 21.4 (Aug. 1998), pp. 449–467. ISSN: 1469-1825. DOI: 10.1017/s0140525x98001253. URL: http://dx.doi.org/10.1017/s0140525x98001253.

[18] Winrich A. Freiwald and Doris Y. Tsao. "Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System". In: *Science* 330.6005 (Nov. 2010), pp. 845–851. ISSN: 1095-9203. DOI: 10.1126/science.1194908. URL: http://dx.doi.org/10.1126/science.1194908.

[19] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. "Introducing EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification". In: *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2018.

[20] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. "Adversarial Examples are not Bugs, they are Features". In: *Advances in Neural Information Processing Systems* (2019).

[21] William B. Johnson and Joram Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space". In: *Conference on Modern Analysis and Probability* (1984), pp. 189–206. ISSN: 0271-4132. DOI: 10.1090/conm/026/737400.

[22] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas

Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (July 2021), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: http://dx.doi.org/10.1038/s41586-021-03819-2.

[23] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. "Similarity of Neural Network Representations Revisited". In: *International Conference on Machine Learning*. 2019.

[24] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. "Do Better ImageNet Models Transfer Better?" In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[25] Nikolaus Kriegeskorte and Rogier A. Kievit. "Representational geometry: integrating cognition, computation, and the brain". In: *Trends in Cognitive Sciences* 17.8 (Aug. 2013), pp. 401–412. ISSN: 1364-6613. DOI: 10.1016/j.tics.2013.06.007. URL: http://dx.doi.org/10.1016/j.tics.2013.06.007.

[26] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. "Representational similarity analysis – connecting the branches of systems neuroscience". In: *Frontiers in Systems Neuroscience* (2008). ISSN: 1662-5137. DOI: 10.3389/neuro.06.004.2008.

[27] Nikolaus Kriegeskorte, Marieke Mur, Douglas A. Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A. Bandettini. "Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey". In: *Neuron* 60.6 (Dec. 2008), pp. 1126–1141. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2008.10.043. URL: http://dx.doi.org/10.1016/j.neuron.2008.10.043.

[28] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. Tech. rep. University of Toronto, Toronto, Ontario, 2009.

[29] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. "Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution". In: *International Conference on Learning Representations*. 2022.

[30] Melody Zixuan Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Kumar Teru, Adam Santoro, Guillaume Lajoie, and Blake A. Richards. "Tracing the Representation Geometry of Language Models from Pretraining to Post-training". In: *ICML Workshop on High-dimensional Learning Dynamics*. 2025.

[31] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. "Holistic Evaluation of Language Models". In: *Transactions on Machine Learning Research* (2023). ISSN: 2835-8856.

[32] Baihan Lin and Nikolaus Kriegeskorte. "The topology and geometry of neural representations". In: *Proceedings of the National Academy of Sciences* 121.42 (Oct. 2024). ISSN: 1091-6490. DOI: 10.1073/pnas.2317881121. URL: http://dx.doi.org/10.1073/pnas.2317881121.

[33] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (2023). DOI: 10.1126/science.ade2574.

[34] Malte D Luecken and Fabian J Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular Systems Biology* 15.6 (June 2019). ISSN: 1744-4292. DOI: 10.15252/msb.20188746. URL: http://dx.doi.org/10.15252/msb.20188746.

[35] Florian P. Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N. Hebart. "Dimensions underlying the representational alignment of deep neural networks with humans". In: *Nature Machine Intelligence* 7.6 (June 2025), pp. 848–859. ISSN: 2522-5839. DOI: 10.1038/s42256-025-01041-7. URL: http://dx.doi.org/10.1038/s42256-025-01041-7.

[36] Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. "Context-dependent computation by recurrent dynamics in prefrontal cortex". In: *Nature* 503.7474 (Nov. 2013), pp. 78–84. ISSN: 1476-4687. DOI: 10.1038/nature12742. URL: http://dx.doi.org/10.1038/nature12742.

[37] Valentina Masarotto, Victor M. Panaretos, and Yoav Zemel. "Procrustes Metrics on Covariance Operators and Optimal Transportation of Gaussian Processes". In: *Sankhya A* 81.1 (2018), pp. 172–213. ISSN: 0976-8378. DOI: 10.1007/s13171-018-0130-1.

[38] Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. "Individual differences among deep neural network models". In: *Nature Communications* 11.1 (Nov. 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-19632-w. URL: http://dx.doi.org/10.1038/s41467-020-19632-w.

[39] Ari Morcos, Maithra Raghu, and Samy Bengio. "Insights on representational similarity in neural networks with canonical correlation". In: *Advances in Neural Information Processing Systems*. 2018.

[40] Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C. Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K. Lampinen. "Aligning machine and human visual representations across abstraction levels". In: *Nature* 647.8089 (Nov. 2025), pp. 349–355. ISSN: 1476-4687. DOI: 10.1038/s41586-025-09631-6. URL: http://dx.doi.org/10.1038/s41586-025-09631-6.

[41] Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. "LEEP: A New Measure to Evaluate Transferability of Learned Representations". In: *International Conference on Machine Learning*. 2020.

[42] Thao Nguyen, Maithra Raghu, and Simon Kornblith. "Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network Representations Vary with Width and Depth". In: *International Conference on Learning Representations*. 2021.

[43] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. "A Toolbox for Representational Similarity Analysis". In: *PLoS Computational Biology* 10.4 (2014), e1003553. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003553.

[44] Maria-Elena Nilsback and Andrew Zisserman. "Automated Flower Classification over a Large Number of Classes". In: *Indian Conference on Computer Vision, Graphics and Image Processing*. Dec. 2008.

[45] Ramon Nogueira, Chris C. Rodgers, Randy M. Bruno, and Stefano Fusi. "The geometry of cortical representations of touch in rodents". In: *Nature Neuroscience* 26.2 (Jan. 2023), pp. 239–250. ISSN: 1546-1726. DOI: 10.1038/s41593-022-01237-9. URL: http://dx.doi.org/10.1038/s41593-022-01237-9.

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. "DINOv2: Learning Robust Visual Features without Supervision". In: *Transactions on Machine Learning Research* (2024). ISSN: 2835-8856.

[47] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. "Librispeech: an ASR corpus based on public domain audio books". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 5206–5210.

[48] Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, L. F. Abbott, and David Sussillo. "Inferring single-trial neural population dynamics using sequential auto-encoders". In: *Nature Methods* 15.10 (Sept. 2018), pp. 805–815. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0109-9. URL: http://dx.doi.org/10.1038/s41592-018-0109-9.

[49] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. "Cats and Dogs". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". In: *International Conference on Machine Learning*. 2021.

[51] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. "SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability". In: *Advances in Neural Information Processing Systems*. 2017.

[52] Prashant C. Raju. *Shesha: Self-Consistency Metrics for Representational Stability*. 2026. DOI: 10.5281/zenodo.18227453. URL: https://doi.org/10.5281/zenodo.18227453.

[53] F. James Rohlf and Dennis Slice. "Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks". In: *Systematic Zoology* 39.1 (1990), p. 40. ISSN: 0039-7989. DOI: 10.2307/2992207.

[54] Shreya Saxena and John P Cunningham. "Towards the neural population doctrine". In: *Current Opinion in Neurobiology* 55 (Apr. 2019), pp. 103–111. ISSN: 0959-4388. DOI: 10.1016/j.conb.2019.02.002. URL: http://dx.doi.org/10.1016/j.conb.2019.02.002.

[55] Peter H. Schönemann. "A Generalized Solution of the Orthogonal Procrustes Problem". In: *Psychometrika* 31.1 (1966), pp. 1–10. ISSN: 1860-0980. DOI: 10.1007/bf02289451.

[56] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo.

"Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?" In: *bioRxiv* (Sept. 2018). DOI: 10.1101/407007. URL: http://dx.doi.org/10.1101/407007.

[57] Heiko H Schütt. "Bayesian Comparisons Between Representations". In: *Conference on Cognitive Computational Neuroscience*. 2025.

[58] Heiko H Schütt, Alexander D Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. "Statistical inference on representational geometries". In: *eLife* 12 (2023). ISSN: 2050-084X. DOI: 10.7554/elife.82566.

[59] Joshua H. Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K. Ramirez, Hannah Choi, Jennifer A. Luviano, Peter A. Groblewski, Ruweida Ahmed, Anton Arkhipov, Amy Bernard, Yazan N. Billeh, Dillan Brown, Michael A. Buice, Nicolas Cain, Shiella Caldejon, Linzy Casal, Andrew Cho, Maggie Chvilicek, Timothy C. Cox, Kael Dai, Daniel J. Denman, Saskia E. J. de Vries, Roald Dietzman, Luke Esposito, Colin Farrell, David Feng, John Galbraith, Marina Garrett, Emily C. Gelfand, Nicole Hancock, Julie A. Harris, Robert Howard, Brian Hu, Ross Hytnen, Ramakrishnan Iyer, Erika Jessett, Katelyn Johnson, India Kato, Justin Kiggins, Sophie Lambert, Jerome Lecoq, Peter Ledochowitsch, Jung Hoon Lee, Arielle Leon, Yang Li, Elizabeth Liang, Fuhui Long, Kyla Mace, Jose Melchior, Daniel Millman, Tyler Mollenkopf, Chelsea Nayan, Lydia Ng, Kiet Ngo, Thuyahn Nguyen, Philip R. Nicovich, Kat North, Gabriel Koch Ocker, Doug Ollerenshaw, Michael Oliver, Marius Pachitariu, Jed Perkins, Melissa Reding, David Reid, Miranda Robertson, Kara Ronellenfitch, Sam Seid, Cliff Slaughterbeck, Michelle Stoecklin, David Sullivan, Ben Sutton, Jackie Swapp, Carol Thompson, Kristen Turner, Wayne Wakeman, Jennifer D. Whitesell, Derric Williams, Ali Williford, Rob Young, Hongkui Zeng, Sarah Naylor, John W. Phillips, R. Clay Reid, Stefan Mihalas, Shawn R. Olsen, and Christof Koch. "Survey of spiking in the mouse visual system reveals functional hierarchy". In: *Nature* 592.7852 (2021). DOI: 10.1038/s41586-020-03171-x.

[60] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Empirical Methods in Natural Language Processing*. 2013.

[61] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. "Feature Selection via Dependence Maximization". In: *Journal of Machine Learning Research* 13.47 (2012), pp. 1393–1434.

[62] Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. "Neural representational geometry underlies few-shot concept learning". In: *Proceedings of the National Academy of Sciences* 119.43 (Oct. 2022). ISSN: 1091-6490. DOI: 10.1073/pnas.2200800119. URL: http://dx.doi.org/10.1073/pnas.2200800119.

[63] Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. "Distributed coding of choice, action and engagement across the mouse brain". In: *Nature* 576.7786 (2019), pp. 266–273. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1787-x.

[64] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine Hermann, Kerem Oktar, Klaus Greff, Martin N Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus Robert Muller, Mariya Toneva, and Thomas L. Griffiths. "Getting aligned on representational alignment". In: *Transactions on Machine Learning Research* (2025). ISSN: 2835-8856.

[65] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction". In: *Science* 290.5500 (2000), pp. 2319–2323. ISSN: 1095-9203. DOI: 10.1126/science.290.5500.2319.

[66] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nature Biotechnology* 32.4 (Mar. 2014), pp. 381–386. ISSN: 1546-1696. DOI: 10.1038/nbt.2859. URL: http://dx.doi.org/10.1038/nbt.2859.

[67] Jean Philippe Vogel. *Indian Serpent-Lore: Or, The Nāgas in Hindu Legend and Art*. Reprint of the 1926 edition. New Delhi: Asian Educational Services, 1995. ISBN: 978-8120610712.

[68] Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. "Reliability of dissimilarity measures for multi-voxel pattern analysis". In: *NeuroImage* 137 (2016), pp. 188–200. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2015.12.012.

[69] Ross Wightman. *PyTorch Image Models*. https://github.com/rwightman/pytorch-image-models. 2019. DOI: 10.5281/zenodo.4414861.

[70] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biology* 19.1 (2018). DOI: 10.1186/s13059-017-1382-0.

[71] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. "Robust fine-tuning of zero-shot models". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.00780.

[72] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. "LogME: Practical Assessment of Pre-trained Models for Transfer Learning". In: *International Conference on Machine Learning*. 2021.

[73] Kaichao You, Yong Liu, Ziyang Zhang, Jianmin Wang, Michael I. Jordan, and Mingsheng Long. "Ranking and Tuning Pre-trained Models: A New Paradigm for Exploiting Model Hubs". In: *Journal of Machine Learning Research* 23.1 (2022). ISSN: 1532-4435.

[74] Bin Yu and Karl Kumbier. "Veridical data science". In: *Proceedings of the National Academy of Sciences* 117.8 (Feb. 2020), pp. 3920–3929. ISSN: 1091-6490. DOI: 10.1073/pnas.1901326117. URL: http://dx.doi.org/10.1073/pnas.1901326117.

[75] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. "A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark". In: *arXiv* (2019).

[76] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. "Massively parallel digital transcriptional profiling of single cells". In: *Nature Communications* 8.1 (2017). DOI: 10.1038/ncomms14049.

[77] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. "Unsupervised neural network models of the ventral visual stream". In: *Proceedings of the National Academy of Sciences* 118.3 (Jan. 2021). ISSN: 1091-6490. DOI: 10.1073/pnas.2014196118. URL: http://dx.doi.org/10.1073/pnas.2014196118.

A SI Text

A.1 Shesha Variants

The main text presents Feature-Split Shesha ($\text{Shesha}_{\text{FS}}$), the primary unsupervised variant. The general Shesha framework admits additional variants, each probing a different aspect of geometric stability by constructing the complementary RDM views $D^{(1)}$ and $D^{(2)}$ through different partitioning strategies. The present paper uses only $\text{Shesha}_{\text{FS}}$.

Feature-Split Shesha ($\text{Shesha}_{\text{FS}}$). The primary variant, described in the main text. Feature dimensions $\{1, \ldots, d\}$ are randomly partitioned into two disjoint halves $F_k^{(1)}, F_k^{(2)}$; an RDM is computed from each half using cosine distance; and Spearman rank correlation between the two vectorized upper triangles is averaged over K=30 random partitions. This variant measures whether geometric structure is redundantly distributed across the feature basis and requires no labels or repeated measurements.

Sample-Split Shesha ($\text{Shesha}_{\text{SS}}$). Data points (rather than features) are partitioned into two disjoint subsets $S_k^{(1)}, S_k^{(2)} \subset \{1, \ldots, n\}$. RDMs are computed within each subset, and correlation is evaluated on the overlapping pairs (those where both samples appear in both partitions) or through anchor-based approaches. This variant measures robustness to input variation across subsets. A low value may indicate that the representation is excessively sensitive to sampling noise or relies on spurious input-specific information. Sample-Split Shesha is not used in the present paper but is included here for completeness, as the feature-split and sample-split variants represent complementary axes of the same split-half principle (features vs. observations).

A.2 Full Invariance Proofs

We verify each invariance property of $\text{Shesha}_{\text{FS}}$ stated in Table 1 of the main text. Let $X \in \mathbb{R}^{n \times d}$ and let $\text{Shesha}_{\text{FS}}$ be computed using cosine distance RDMs and Spearman rank correlation, averaged over K random equipartitions of $\{1, \ldots, d\}$.

Global scaling. For any $\alpha > 0$, cosine distance satisfies

$$D_{ij}(\alpha X) = 1 - \frac{(\alpha x_i)^\top (\alpha x_j)}{\|\alpha x_i\| \, \|\alpha x_j\|} = 1 - \frac{x_i^\top x_j}{\|x_i\| \, \|x_j\|} = D_{ij}(X). \tag{5}$$

Since all RDMs are unchanged, Shesha is unchanged.

Isotropic scaling. Follows identically from the above, since isotropic scaling $X \mapsto \alpha X$ does not change cosine distances.

Feature permutation. A permutation P relabels feature indices. Random equipartition of $\{1, \ldots, d\}$ is uniform and therefore invariant under relabeling: the distribution over partitions $(F^{(1)}, F^{(2)})$ of the permuted indices equals the distribution over partitions of the original indices. Hence $\mathbb{E}[\text{Shesha}_{\text{FS}}(XP)] = \mathbb{E}[\text{Shesha}_{\text{FS}}(X)]$.

Monotonic distance invariance. Spearman correlation depends only on ranks. A strictly monotone transformation g preserves all pairwise orderings: $D_{ij} < D_{kl} \Leftrightarrow g(D_{ij}) < g(D_{kl})$. The rank vectors are therefore identical, leaving ρ_s unchanged.

Non-invariance to orthogonal transformations. We construct a counterexample. Let $d = 4$ and

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{pmatrix}.$$

The geometric structure is uniformly distributed across all four coordinates: any split into two pairs of coordinates yields RDMs that agree, producing high Shesha. Now let $\mathbf{Q}$ be an orthogonal matrix that concentrates the row energy of $\mathbf{X}$ into coordinates 1 and 2 (e.g., a block rotation that maps columns 3 and 4 into the span of columns 1 and 2). Set $\mathbf{Y} = \mathbf{XQ}$. After

rotation, coordinates 3 and 4 carry negligible geometric information. Splits that separate $\{1, 2\}$ from $\{3, 4\}$ now yield one informative and one nearly uninformative RDM, reducing Spearman correlation. Since such splits occur with positive probability under random equipartition, the expected Shesha decreases. Meanwhile, $\mathbf{Y}\mathbf{Y}^\top = \mathbf{X}\mathbf{Q}\mathbf{Q}^\top\mathbf{X}^\top = \mathbf{X}\mathbf{X}^\top$, so linear CKA satisfies $\text{CKA}(\mathbf{X}, \mathbf{Y}) = \text{CKA}(\mathbf{X}, \mathbf{X}) = 1$.

This non-invariance is the structural mechanism underlying the empirical dissociation from similarity metrics. It enables Shesha to detect compression-induced damage to manifold structure (PCA concentrates geometry into fewer coordinates, making feature splits asymmetric) that CKA, operating on the rotation-invariant Gram matrix, cannot see.

A.3 Connection to RSA Noise Ceiling

The noise ceiling in RSA, introduced by Nili et al. (2014), bounds how well any model RDM can correlate with an empirical brain RDM given measurement noise. It is computed by splitting observations (trials or subjects) into two groups, computing an RDM from each, and correlating the resulting RDM vectors. The upper bound uses the mean of one group correlated with the other; the lower bound uses one group correlated with the grand mean.

Shesha adapts the same split-half correlation machinery but applies it along the feature axis rather than the observation axis. Where the noise ceiling asks "given measurement noise across trials, how replicable is the observed RDM?", Shesha asks "given the distribution of geometric information across features, how consistently is the RDM recovered from arbitrary feature subsets?"

The key differences are:

(i) *Axis of splitting.* Noise ceiling: observations (trials, subjects). Shesha: features (neurons, embedding dimensions).

(ii) *Diagnostic target.* Noise ceiling: data quality (is the measurement reliable?). Shesha: representational architecture (is the geometry redundantly encoded?).

(iii) *Requirements.* Noise ceiling: requires repeated measurements of the same conditions. Shesha: requires only a single measurement matrix $X \in \mathbb{R}^{n \times d}$, enabling assessment of pretrained embeddings, single-cell profiles, and other systems where observation-level replication is unavailable.

Despite these differences, the mathematical structure is identical: both compute Spearman correlation between vectorized upper triangles of RDMs derived from complementary partitions of the data. This shared structure means that the statistical properties of split-half RDM correlation (convergence rates, bias correction, bootstrap inference) established for the noise ceiling apply directly to Shesha.

B SI Methods

B.1 Shesha Computation

All $\text{Shesha}_{\text{FS}}$ computations followed a standardized protocol. Feature dimensions were randomly partitioned into two disjoint halves of equal size (for odd d, one half received $(d + 1)/2$ features). Cosine distance RDMs were computed from each half. Spearman rank correlation between the vectorized upper triangles of the two RDMs was computed. This procedure was repeated for K=30 random partitions and averaged.

When n^2 RDM computation was prohibitive, samples were subsampled to $n_{\max} = 1{,}600$ (stratified by available labels when present, random otherwise). Convergence analysis confirmed that subsampled estimates deviate from full-sample values by a mean absolute difference of 0.0077, with maximum deviation below 0.02 across all tested conditions.

All computations used fixed random seed 320 for reproducibility. Float64 precision was used throughout for ranking and correlation computations to avoid numerical artifacts from tied ranks.

CKA was computed as debiased linear CKA using the unbiased estimator of HSIC (Song et al., 2012), which zeros the Gram matrix diagonals. This correction eliminates the positive bias (~0.4 for independent random matrices) present in standard linear CKA.

B.2 Cross-Domain Validation: Data Sources and Preprocessing

Language (N=127). Sentences from the SST-2 validation set (Socher et al. 2013) were tokenized using each model's default tokenizer with padding and truncation (max length: 64 tokens). Representations were extracted from the final hidden layer and mean-pooled across tokens using attention masks. 500 sentences; base models: `all-MiniLM-L6-v2`, `all-mpnet-base-v2`, `distilbert-base-nli-stsb-mean-tokens`, and `paraphrase-distilroberta-base-v1`.

Vision (N=129). Images from CIFAR-100 (Krizhevsky 2009) were preprocessed using each model's default image processor (resized to 224×224, ImageNet normalization). Representations were extracted from the final layer with global average pooling. 400 images; base models: `google/vit-base-patch16-224`, `openai/clip-vit-base-patch32`, `facebook/deit-base-patch` and ResNet50 (ImageNet-V2 weights).

Audio (N=64). Audio samples from LibriSpeech dev-clean (Panayotov et al. 2015) were resampled to 16 kHz and truncated/padded to 1 second duration. Representations were extracted from the final encoder layer and mean-pooled across time. 200 samples; base models: `facebook/wav2vec2-base-960h` and `facebook/hubert-base-ls960`.

Video (N=128). Video clips from the Jellyfish sample (Allyn 2016) were uniformly sampled at 16 frames per clip and preprocessed to 224 × 224 spatial resolution with ImageNet normalization. 100 clips; base models: temporal transformers (`facebook/timesformer-base-finetuned-k400`, `MCG-NJU/videomae-base`) and frame-level encoders (ViT on mean frame, CLIP multi-frame averaging).

Protein (N=402). Protein sequences from Swiss-Prot (UniProt reviewed human proteins; Bateman et al. 2022), filtered to lengths between 50 and 2,000 residues. 200 sequences; multiple encoding schemes: amino acid composition (20-dim), dipeptide frequency (400-dim), hydrophobicity and charge profiles at multiple resolutions (25, 50, 100 bins), and 3-mer spectra (500-dim hashed).

Molecular (N=767). Single-cell RNA-seq data from the pbmc3k dataset (Zheng et al. 2017), loaded with Scanpy (Wolf et al., 2018). Genes with fewer than 3 expressing cells were filtered. 1,000 cells; multiple preprocessing strategies: log-transformation, various PCA dimensions, top-variance gene selection, CPM normalization, and binarization (presence/absence).

Neural population recordings (N=846). Neuropixels recordings from the (Steinmetz et al. 2019) dataset, comprising high-density recordings from 29,134 neurons across 42 brain areas in awake mice. Sessions were filtered to include only those with at least 20 neurons and 50 trials (N=26 qualified sessions). Spike counts were binned at 20 ms resolution and averaged across time bins.

B.3 Encoder Transformations

For each base representation in each domain, we applied a standardized set of geometric interventions, resulting in 2,463 unique encoder configurations across all seven domains, aggregated across 15 seeds ($\{3, 7, 9, 11, 12, 18, 103, 108, 320, 411, 724, 1754, 1991, 2222, 7258\}$). Table 4 lists all transformation categories and their parameter ranges.

PCA. Principal component projection to k dimensions, with $k \in \{5, 10, \ldots, 300\}$ (capped at $\min(n, d) - 1$).

Random projection. Gaussian random projection to k dimensions, $k \in \{16, 32, \ldots, 256\}$.

Top-variance feature selection. Selection of k features with highest marginal variance, $k \in \{50, 100, \ldots, 800\}$.

Random feature subsets. Random subset of k features without replacement, $k \in \{50, 100, 200\}$.

Gaussian noise injection. Additive Gaussian noise scaled by $\sigma \cdot \text{std}(X)$, with $\sigma \in \{0.05, 0.1, \ldots, 1.0\}$.

Normalization. Z-score (per-feature zero mean, unit variance) and L2 (per-sample unit norm).

B.4 Similarity Metrics

For each encoder configuration, CKA was computed between the transformed representation and three domain-specific reference representations: the original untransformed base representation, a PCA projection at k=100 (or the closest available rank), and a z-scored version. The three CKA values were averaged to produce a single similarity score per configuration, minimizing single-reference artifacts.

Alternative similarity metrics were evaluated in the language domain (N=127):

Effective-rank PWCKA. Both representations are projected to a shared dimensionality determined by the minimum effective rank (number of components explaining 99% of variance). CKA is then computed on the truncated projections.

Procrustes similarity. After centering and Frobenius normalization, the optimal orthogonal alignment matrix $R^* = UV^\top$ is obtained via SVD of the cross-covariance matrix. Procrustes similarity: $1 - \|\tilde{X} - \tilde{Y}R^*\|_F^2 / (\|\tilde{X}\|_F^2 + \|\tilde{Y}R^*\|_F^2)$.

B.5 Statistical Methods

Bootstrap inference. Distinctness was assessed via Spearman rank correlation with 10,000 bootstrap replicates, resampling encoder configurations within each domain. 95% confidence intervals were computed as bootstrap percentile intervals.

Mixed-effects models. To control for dependencies among configurations derived from the same base model, we fitted a linear mixed-effects model: Stability ~ Similarity + (1 | BaseModel). The intraclass correlation coefficient (ICC) quantifies the proportion of variance attributable to base model identity.

Mann-Whitney U tests. Architectural comparisons (contrastive vs. self-supervised; hierarchical vs. columnar) used two-sided Mann-Whitney U tests on $\text{Shesha}_{\text{FS}}$ scores, reported with exact p-values.

Multiple comparisons. Per-dataset statistical tests in the vision benchmark are reported without multiplicity correction, as each dataset represents an independent evaluation domain rather than a repeated test of the same hypothesis.

B.6 Vision Benchmark: Extended Methods

Model selection. 94 pretrained vision models were drawn from the PyTorch Image Models (timm) library(Wightman 2019). Selection ensured broad coverage across four axes: (i) training objectives (supervised ImageNet-1k/21k, self-supervised DINO/DINOv2/MAE, contrastive CLIP, generative EVA-02/BEiT); (ii) architectural families (columnar ViT/DeiT, hierarchical Swin/SwinV2/PVT-v2, hybrid CoAtNet/MaxViT, convolutional ResNet/ConvNeXt/EfficientNet/RegNet/DenseNet); (iii) model scales (MobileNetV3-Small to ViT-Giant/14); (iv) training paradigms (standard, distillation, augmentation, foundation model pretraining). Models were grouped into 29 semantic families for aggregate analysis. When training objective and architecture conflicted, training objective was prioritized for family assignment (e.g., ViT-CLIP assigned to "CLIP" rather than "ViT").

Feature extraction. Penultimate-layer features were extracted from fixed random subsets of each dataset (seed 320): 5,000 images for CIFAR-10, CIFAR-100, and EuroSAT; 5,000 for Flowers-102 (with replacement where the dataset is smaller); 1,500 for Oxford Pets; 1,600 for DTD. All images were preprocessed using each model's standard transform (resize, center crop, normalization).

Transferability metrics. LogME (You, Liu, Wang, et al. 2021; You, Liu, Zhang, et al. 2022) was computed on the same features using the authors' implementation. LEEP (C. V. Nguyen et al. 2020) was computed for models with classification heads.

C SI Figures

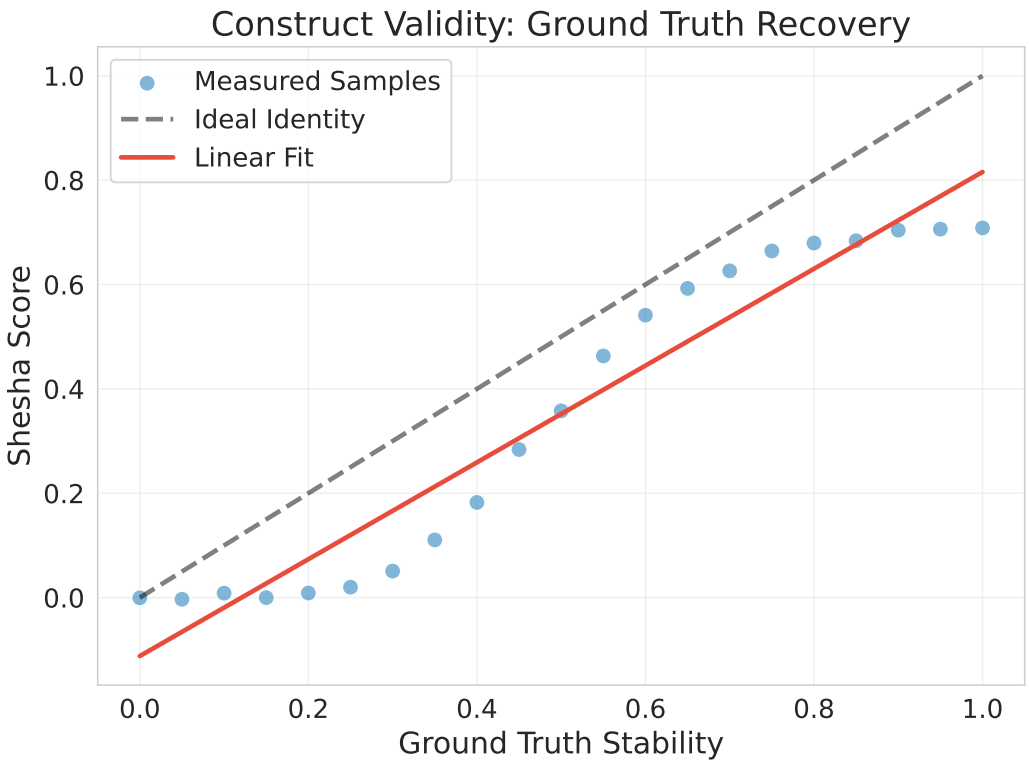

Figure 4: **Construct Validity: Ground Truth Recovery.** Shesha scores plotted against parametrically controlled stability levels (signal-to-noise ratio) in synthetic representations. The metric shows a near-perfect monotonic response ($\rho = 0.990$) to the underlying ground truth, confirming high sensitivity to geometric consistency.

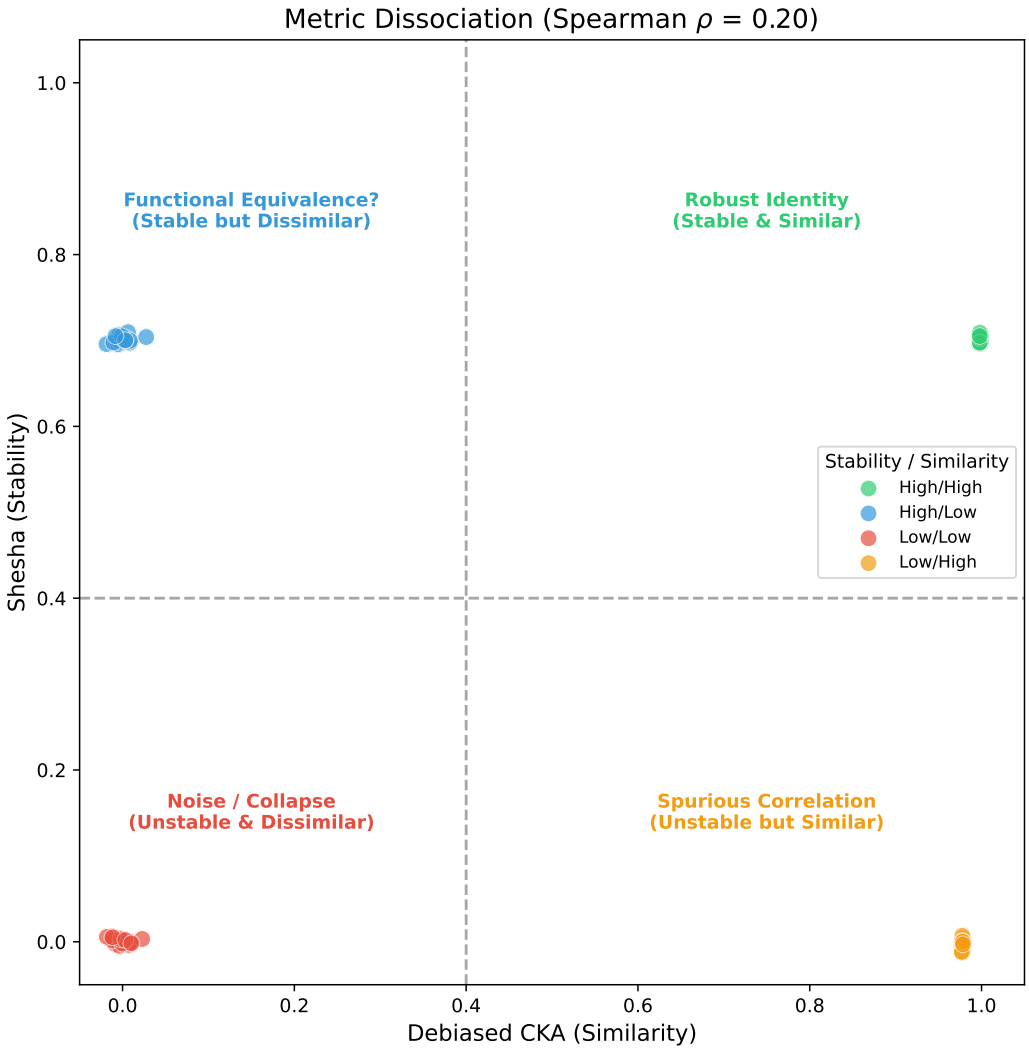

Figure 5: **Four-quadrant dissociation.** Shesha vs. debiased CKA for 60 representation pairs sampled equally from four quadrants of the stability × similarity space. Q1 (high/high): Shesha = 0.701 ± 0.003, CKA = 0.998 ± 0.000. Q2 (high/low): Shesha = 0.701 ± 0.004, CKA = 0.001 ± 0.010. Q3 (low/low): Shesha = 0.001 ± 0.003, CKA = -0.001 ± 0.010. Q4 (low/high, adversarial): Shesha = -0.001 ± 0.005, CKA = 0.978 ± 0.000. Balanced Spearman $\rho = 0.20$.

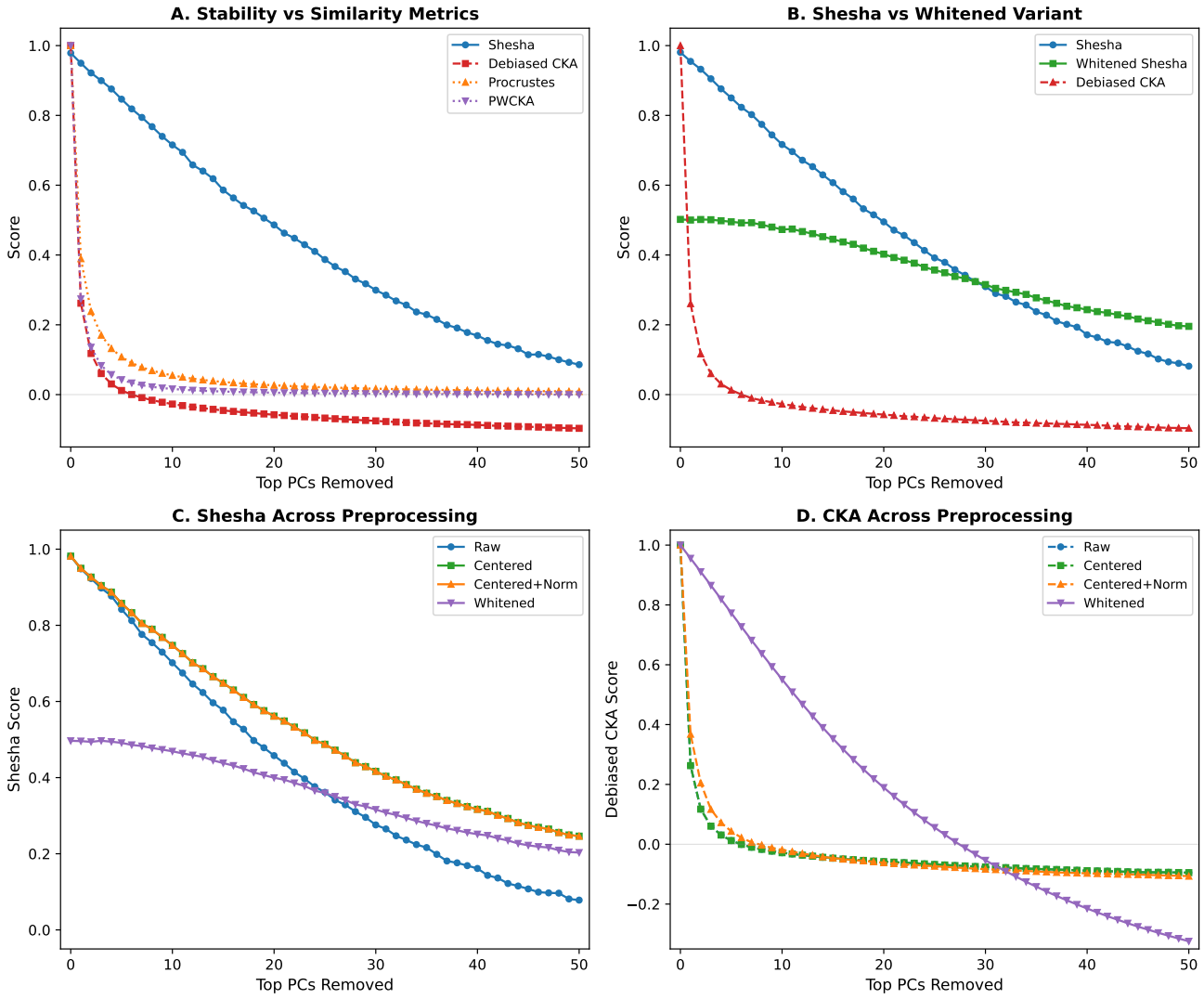

Figure 6: **Spectral Sensitivity Analysis.** We measure metric responses as the top k principal components are progressively removed from a power-law representation. **(A)** Shesha degrades gracefully while all similarity metrics (CKA, PWCKA, Procrustes) collapse after removing just 1 PC. **(B)** Comparison with whitened Shesha shows high correlation ($\rho = 0.999$), though whitening reduces baseline stability. **(C)** Shesha robustness across preprocessing conditions (raw, centered, normalized, whitened). **(D)** CKA behavior across preprocessing; notably, whitening causes CKA to recover sensitivity by equalizing the spectrum.

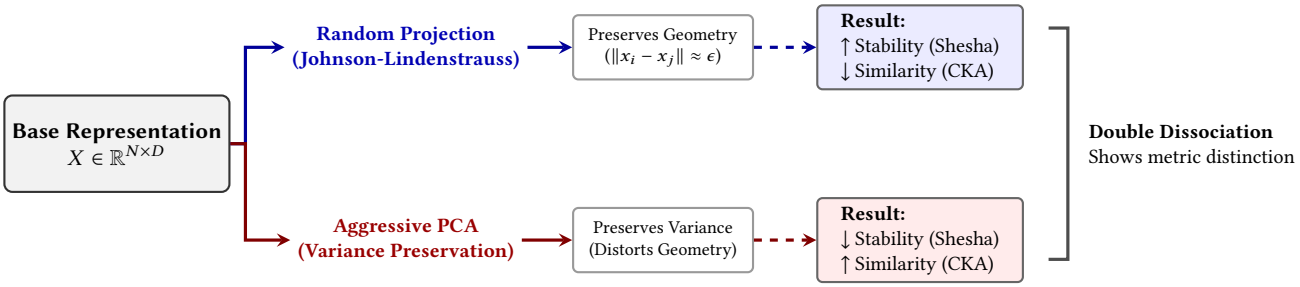

Figure 7: **Encoder-type breakdown.** Spearman correlation between Shesha and CKA stratified by transformation type (from Table 16 of the main Shesha manuscript). Geometry-preserving transforms show strong positive correlation (random features: $\rho = +0.92$; random projection: $\rho = +0.90$; top variance: $\rho = +0.64$; noise injection: $\rho = +0.58$). Natural encoders show weak positive correlation (normalization: $\rho = +0.34$; original: $\rho = +0.31$). PCA compression shows strong negative correlation ($\rho = -0.47$).

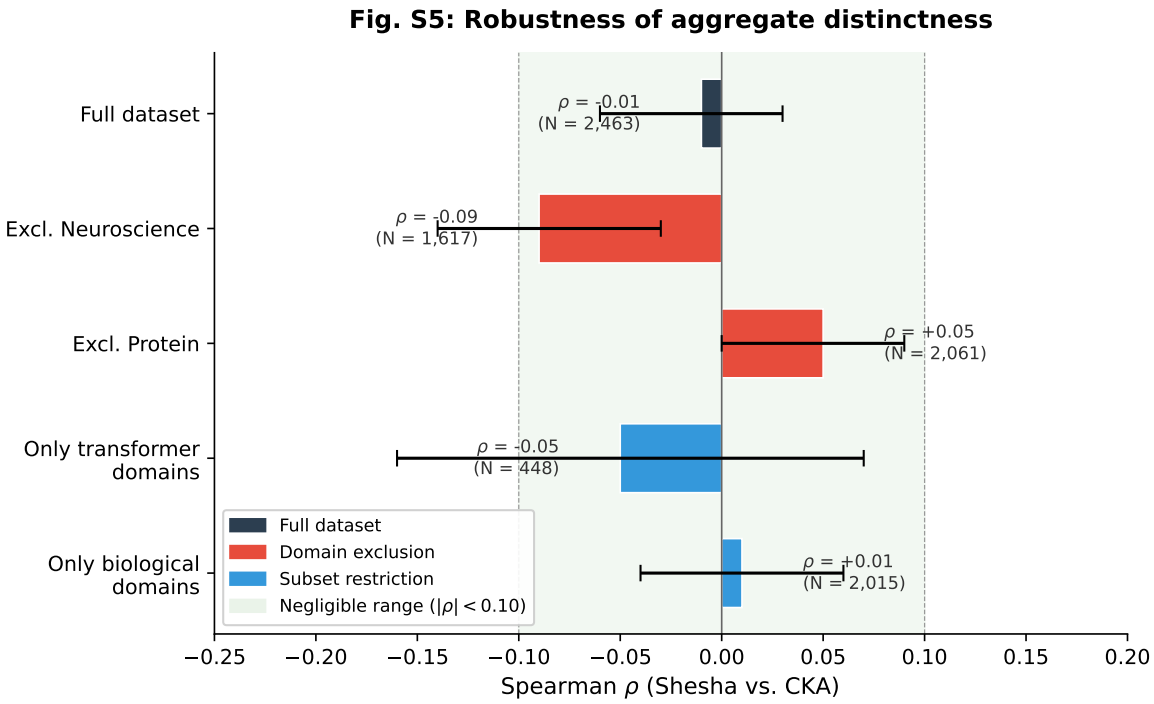

Figure 8: **Robustness: excluding domains.** Aggregate Shesha-CKA correlation after excluding each domain individually. All subsets maintain $|\rho| < 0.10$.

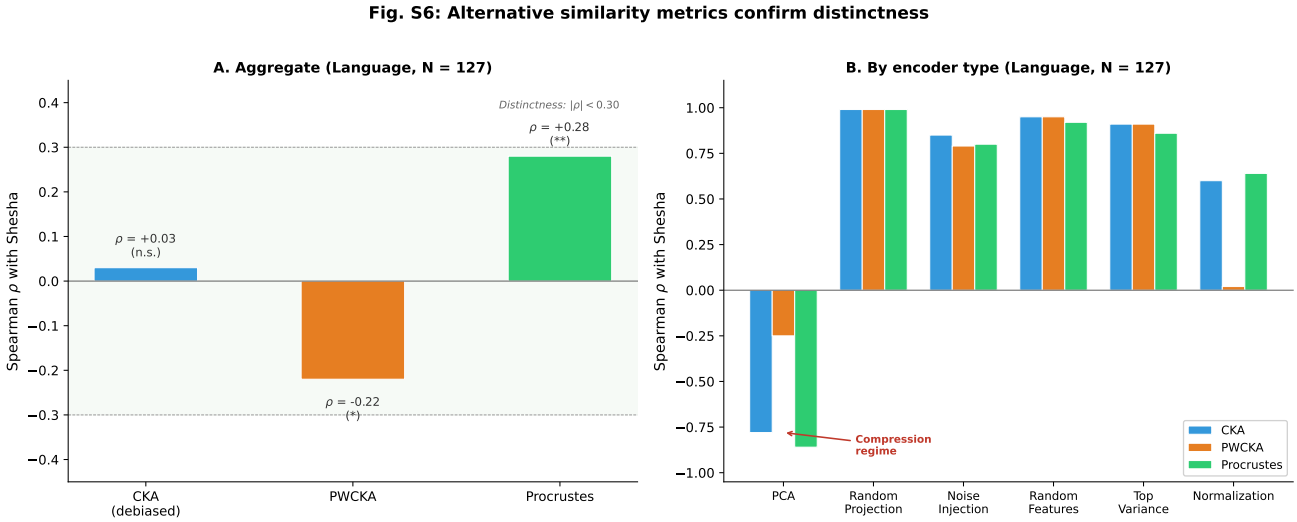

Figure 9: **Alternative similarity metrics.** Shesha vs. PWCKA and Shesha vs. Procrustes in the language domain (N=127). Both alternative metrics show weak correlations with Shesha ($|\rho| < 0.30$).

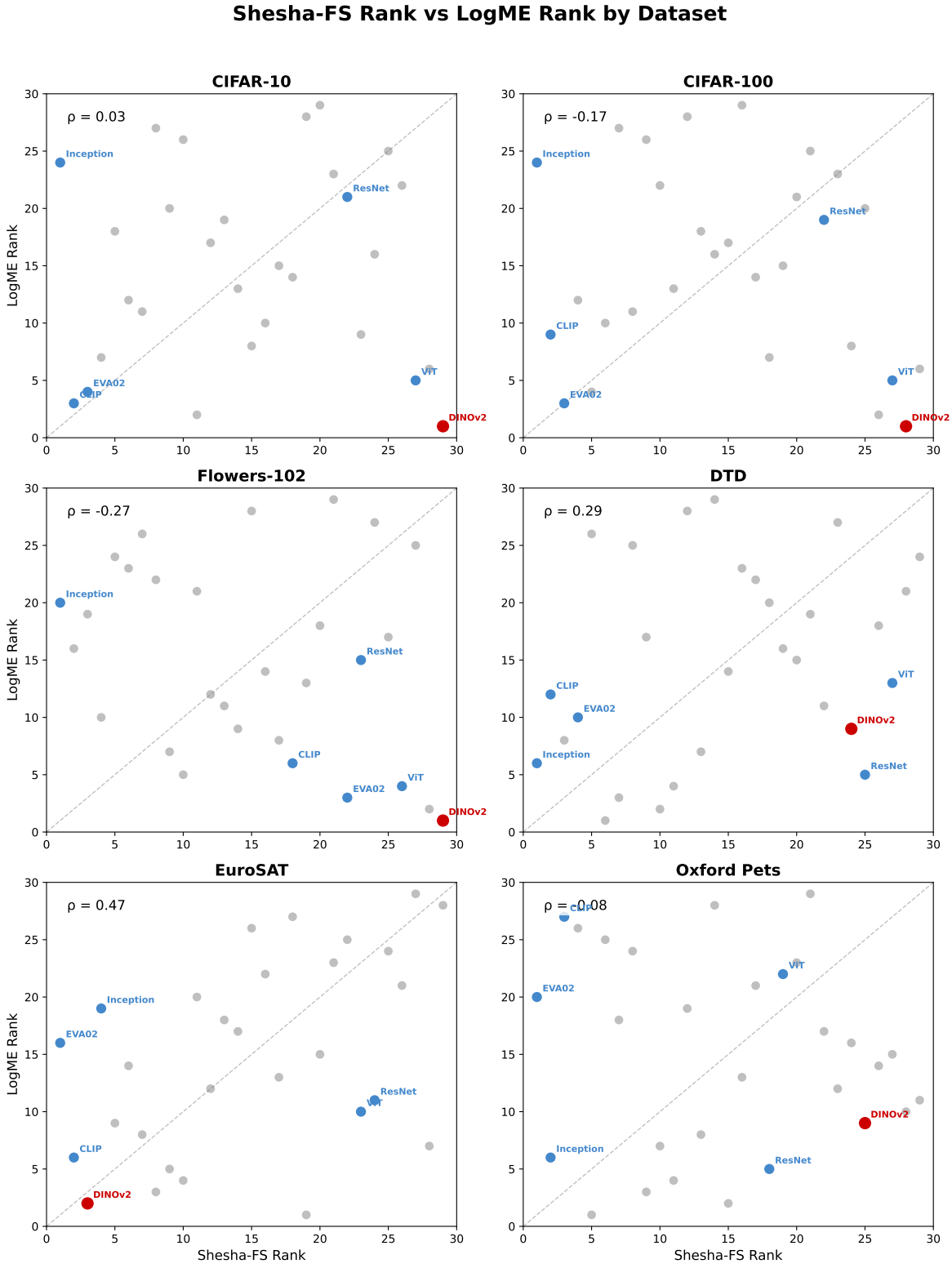

Figure 10: **Complete vision results (Fig. S7).** Shesha$_{FS}$ scores for all 94 models across 6 datasets, grouped by architectural family. Color encodes stability (darker = higher). DINOv2 family shows consistently low stability except on EuroSAT; CLIP and EVA families show consistently high stability.

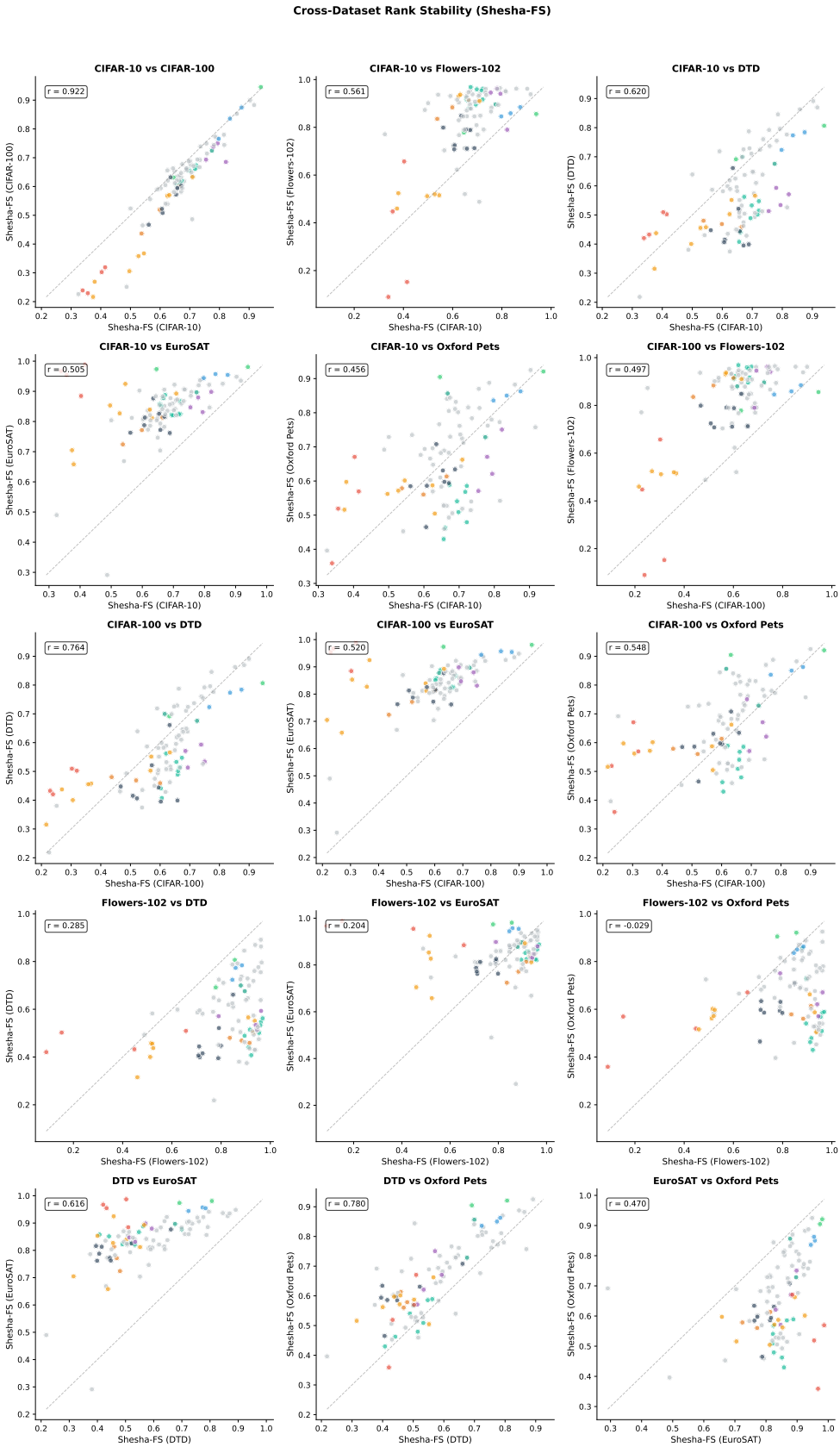

Figure 11: **Cross-dataset rank stability.** Pairwise scatter plots of $\text{Shesha}_{\text{FS}}$ family ranks across all 15 dataset pairs. CIFAR-10 vs. CIFAR-100: $\rho = 0.92$. Within-domain pairs (same image distribution, different task complexity) show high rank consistency; cross-domain pairs show weaker but consistently positive correlations.

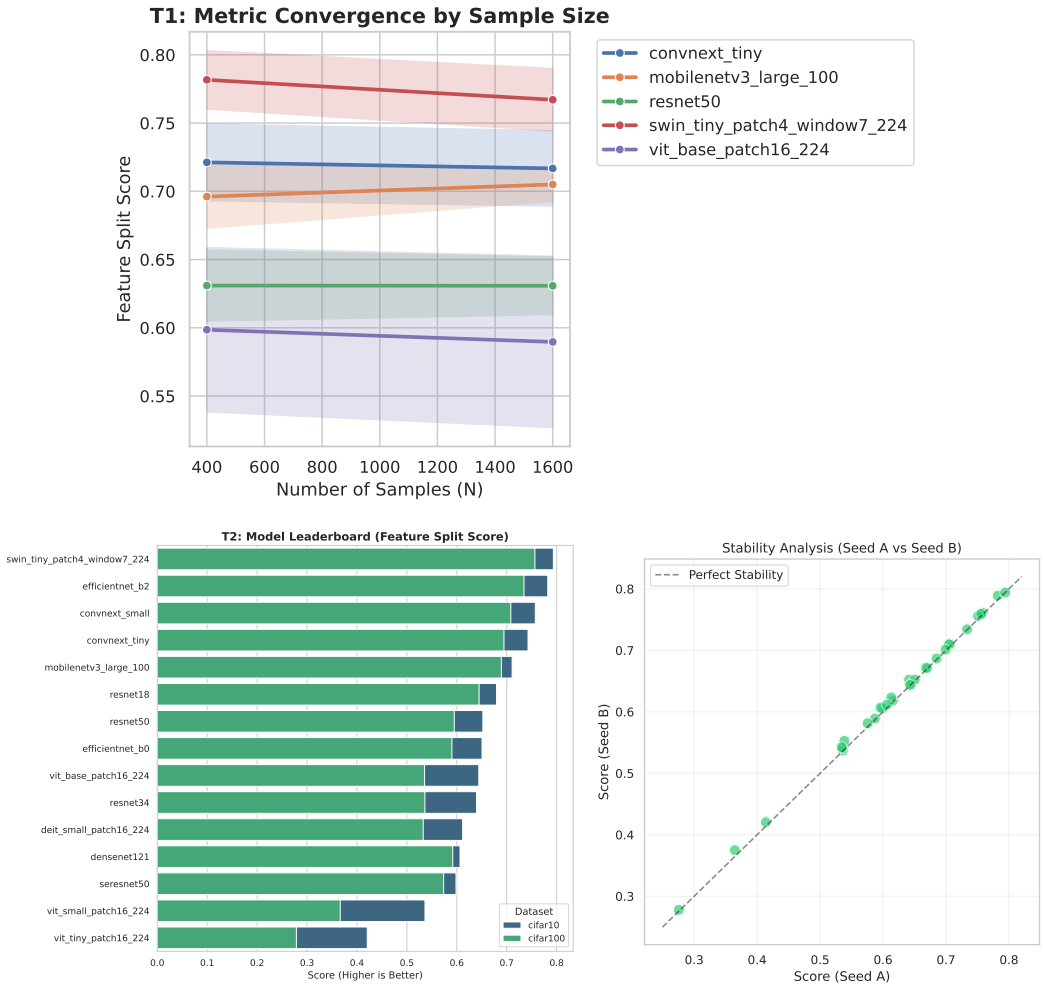

Figure 12: **Metric validation suite.** Summary panels for the ten validation tests in Table S6. (A) Shesha estimates remain stable as sample size increases from 400 to 1600 across representative architectures. The flat trajectories confirm rapid convergence and numerical reliability at modest sample sizes. (B) Ranking of 15 architectures by Shesha score. Bar segments show contributions from CIFAR-10 (teal) and CIFAR-100 (blue). Modern architectures with attention or dense connectivity achieve higher geometric stability. (C) Comparison of Shesha scores computed with two different random seeds (Seed A=100 vs. Seed B=200). Points align closely with the diagonal identity line, indicating high reproducibility across random initializations.

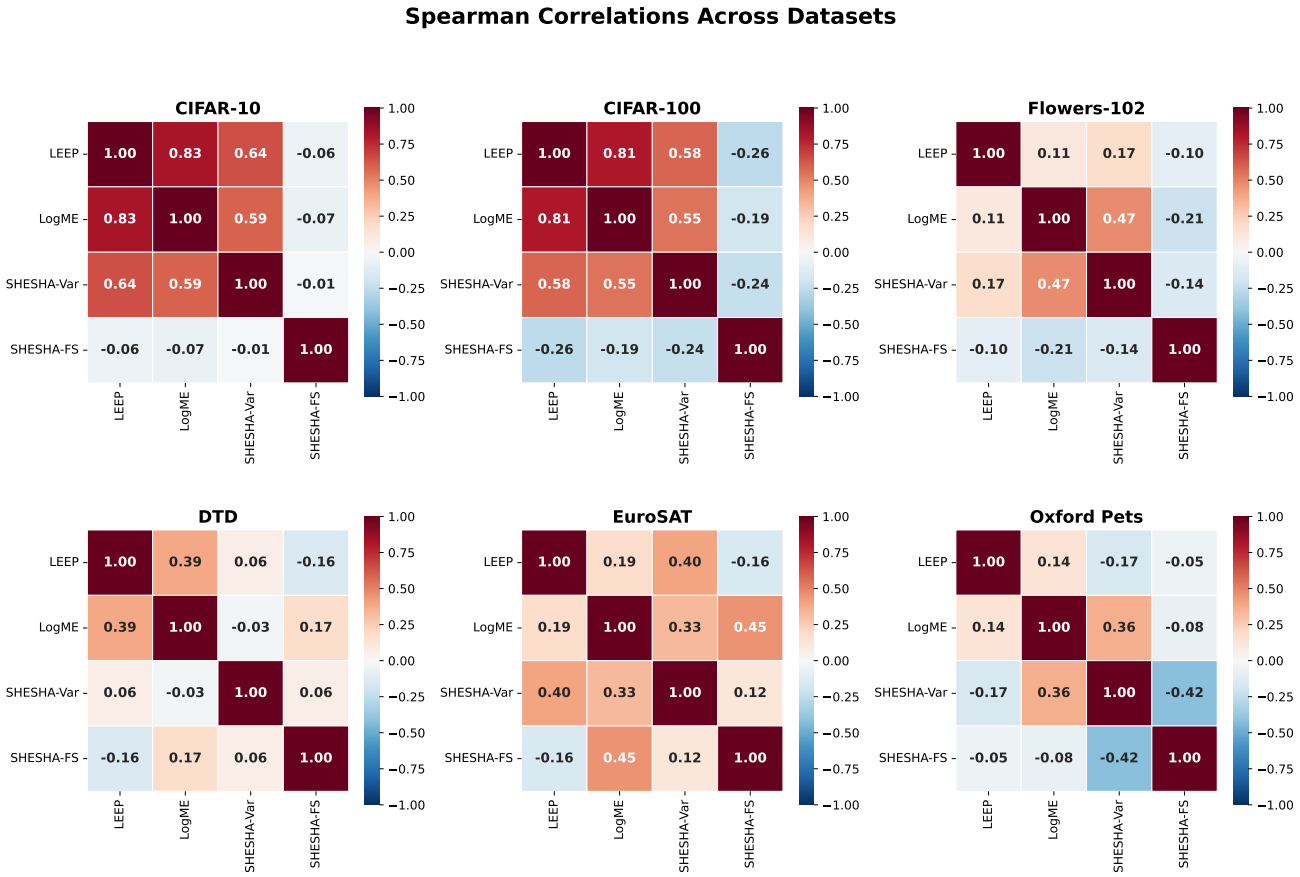

Figure 13: **Architecture heatmap.** $\text{Shesha}_{\text{FS}}$ by architectural family (rows) × dataset (columns). Moved from main paper to SI to make room for the conceptual eigenspectrum illustration (Fig. 3B). DINOv2 row shows low stability on all datasets except EuroSAT. CLIP, EVA, and Inception rows show consistently high stability. Swin and PVT rows show the hierarchical advantage on CIFAR and Flowers but not DTD or EuroSAT.

D SI Tables

Table 4: **Encoder transformation catalog.** All geometric interventions applied to base representations in each domain, following the protocol in the main Shesha manuscript.

Category	Variants	Description
PCA	$k \in \{5, 10, \ldots, 300\}$	Principal component projection to k dims
Random proj.	$k \in \{16, 32, \ldots, 256\}$	Gaussian random projection to k dims
Top variance	$k \in \{50, 100, \ldots, 800\}$	Selection of k highest-variance features
Random features	$k \in \{50, 100, 200\}$	Random subset of k features
Noise injection	$\sigma \in \{0.05, 0.1, \ldots, 1.0\}$	Additive Gaussian, $\sigma \cdot \text{std}(X)$
Normalization	Z-score, L2	Per-feature or per-sample normalization
Original	—	Unmodified base representation

Table 5: **The DINOv2 paradox at individual model level.** DINOv2-giant achieves highest LogME on 4/6 datasets while ranking in the bottom quartile for $\text{Shesha}_{\text{FS}}$, except on EuroSAT.

Dataset	LogME	LogME Rank	$\text{Shesha}_{\text{FS}}$	FS Rank
CIFAR-10	1.386	1/94	0.415	88/94
CIFAR-100	1.629	1/94	0.319	86/94
Flowers-102	3.521	1/93	0.152	92/93
DTD	0.952	19/93	0.502	62/93
EuroSAT	0.681	1/93	0.987	1/93
Oxford Pets	1.760	19/93	0.569	68/93

Table 6: **Contrastive vs. self-supervised stability.** Mann-Whitney U tests comparing CLIP (n=3) to self-supervised models (n=9) on $\text{Shesha}_{\text{FS}}$. $^{*}p < 0.05$.

Dataset	CLIP	SSL	Δ	p
CIFAR-10	0.84 ± 0.04	0.57 ± 0.20	+0.27	0.032*
CIFAR-100	0.83 ± 0.06	0.48 ± 0.24	+0.34	0.032*
Flowers-102	0.86 ± 0.02	0.54 ± 0.28	+0.32	0.032*
DTD	0.76 ± 0.03	0.55 ± 0.13	+0.21	0.032*
EuroSAT	0.95 ± 0.01	0.91 ± 0.08	+0.04	0.568
Oxford Pets	0.85 ± 0.01	0.68 ± 0.19	+0.17	0.141

Table 7: **Hierarchical vs. columnar transformer stability.** Mann-Whitney U tests on $\text{Shesha}_{\text{FS}}$. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Dataset	Hier. (n=18)	Col. (n=23)	Δ	p
CIFAR-10	0.70 ± 0.07	0.58 ± 0.18	+0.12	0.011^{*}
CIFAR-100	0.63 ± 0.07	0.48 ± 0.23	+0.15	0.007^{**}
Flowers-102	0.92 ± 0.04	0.66 ± 0.24	+0.26	$<0.001^{***}$
DTD	0.51 ± 0.06	0.52 ± 0.15	−0.01	0.197
EuroSAT	0.83 ± 0.04	0.83 ± 0.17	+0.00	0.844
Oxford Pets	0.59 ± 0.08	0.65 ± 0.16	−0.07	0.941

Table 8: **Robustness checks for aggregate distinctness.** All subsets maintain $|\rho| < 0.10$.

Analysis	N	ρ [95% CI]
Full dataset	2463	−0.01 [−0.06, +0.03]
Excluding Neuroscience	1617	−0.09 [−0.14, −0.03]
Excluding Protein	2061	+0.05 [−0.00, +0.09]
Only transformer domains	448	−0.05 [−0.16, +0.07]
Only biological domains	2015	+0.01 [−0.04, +0.06]

Table 9: **Metric validation suite summary.** Ten tests confirming Shesha's measurement properties.

Test	Property	Result
1	Convergence (n=400 vs. n=1600)	mean $\|\Delta\| = 0.0077$
2	Model leaderboard (cross-dataset rank)	$\rho = 0.93$
3	Determinism	90/90 bitwise identical
4	Numerical validity	0 NaN/Inf
5	Dimensionality sensitivity (PCA)	monotonic
6	Label noise sensitivity	91–93% reduction
7	Class imbalance robustness (20:1)	$\|\Delta\| < 0.03$
8	Input perturbation (σ=0.1)	$\|\Delta\| < 0.05$
9	Seed stability (15 seeds)	mean $\|\Delta\| = 0.0047$
10	Sanity baseline (random)	Shesha = 0.003

Table 10: **Alternative similarity metrics, language domain.** All metrics maintain $|\rho| < 0.30$ with Shesha, confirming distinctness generalizes beyond CKA.

Similarity metric	ρ with Shesha	p	Distinct?
CKA (debiased)	+0.03	0.74	Yes
PWCKA	−0.22	0.012	Yes
Procrustes	+0.28	0.001	Yes