Geometric Stability: The Missing Axis of Representations

Prashant C. Raju RAJUPRASHANT@GMAIL.COM

Abstract

Representational similarity analysis and related methods compare the internal geometries of neural networks, but they measure only alignment between spaces, leaving a blind spot—whether a representation's structure is reliably recoverable, not merely similar. We introduce geometric stability, a distinct axis, and *Shesha*, a metric that quantifies it from a single representation by correlating dissimilarity matrices built from complementary random halves of the feature dimensions. Unlike CKA and Procrustes distance, Shesha is provably non-invariant to orthogonal rotations of the feature basis. This is by design: the basis is privileged for learned models, since probes, patching, and steering act on coordinates, and a rotation-invariant metric cannot see whether the targeted structure survives them. A double dissociation isolates the mechanism—removing the top principal component collapses CKA while Shesha holds, whereas rotating a representation into its eigenbasis, which preserves the spectrum and CKA exactly, collapses Shesha. Across 2,463 encoder configurations in seven domains, the metrics are redundant under geometry-preserving transforms and anti-correlate under compression ($\rho = -0.47$). Across 170 vision models spanning 6 clean and 38 corruption-shifted datasets, DINOv2 ranks first or second in transferability on three of six clean datasets yet bottom-quartile in stability on five, an isolated dissociation rather than a trade-off.

Keywords: representational geometry, representational similarity analysis, representation learning, geometric stability, model evaluation, foundation models

1 Introduction

The representations learned by neural networks underlie their success, and characterizing these representations has become central to understanding both artificial and biological systems (Bengio et al., 2013). Characterizing the *geometry* of a high-dimensional representation, in particular, is a central problem in the analysis of neural networks and biological systems. The dominant framework addresses this through similarity: methods such as Representational Similarity Analysis (RSA, Kriegeskorte et al., 2008), Centered Kernel Alignment (CKA, Kornblith et al., 2019a), Procrustes distance (Schönemann, 1966; Rohlf and Slice, 1990; Masarotto et al., 2018; Dryden and Mardia, 1998), and their extensions (Raghu et al., 2017; Morcos et al., 2018; Lin and Kriegeskorte, 2024) quantify the alignment between two representational spaces, asking whether two systems encode comparable pairwise structure. This framework has proven productive: it has established correspondences between deep neural networks and ventral visual cortex (Yamins et al., 2014), linked recurrent network dynamics to motor cortical population activity (Sussillo et al., 2015), revealed hidden representational learning in mouse sensory cortex (Kumar et al., 2025) that parallels grokking in artificial networks (Power et al., 2022), uncovered structural differences between vision transformers and convolutional architectures (Raghu et al., 2021), and organized model families by their internal geometries (Kornblith et al., 2019a).

Yet similarity answers only one of two natural questions about a representation. The first question—do two systems encode similar structure?—is addressed by the methods above. The second question—does a single system's geometry hold reliably under perturbation of its feature basis?—is not. These two questions are distinct, and even the first is less settled than it appears: Davari et al. (2023) demonstrated that CKA values can be manipulated without altering functional behavior; Murphy et al. (2024) showed that biased CKA produces spuriously high scores for random matrices in the low-data high-dimensionality regime typical of neural recordings; and Cloos et al. (2025) showed that CKA prioritizes high-variance principal components to the point that critical task-relevant dimensions can be entirely missed while similarity scores remain high. Recent work has clarified what these invariant measures do capture: Harvey et al. (2024) show that CKA and CCA quantify the average alignment of optimal linear readouts across a distribution of decoding tasks, tying representational similarity directly to linear decodability. This makes precise the axis on which similarity metrics are informative, and, by the same token, the axis they cannot see: whether that geometry is reliably recoverable from subsets of the feature coordinates, a property their invariance to basis transformations renders invisible.

Two representations may be highly similar under CKA while one is geometrically fragile: its pairwise distance structure fractures when evaluated on complementary subsets of features, collapses when dominant principal components are removed, or shifts substantially under minor redistributions of geometric information across coordinate axes. The field's ongoing reliance on similarity metrics to validate these representations obscures a practical reality for mechanistic interpretability. Modern techniques, including linear probes (Alain and Bengio, 2017), activation patching (Wang et al., 2023; Meng et al., 2022), and steering vector interventions (Zou et al., 2023; Turner et al., 2023), all implicitly assume that the geometric structure they target is consistent across feature subsets. Practitioners rely on the assumption that probing half the residual stream, or intervening on a specific subset of neurons, recovers the same representational logic as probing the full representation. If a system's geometry is fragile, these interventions may be targeting latent geometric artifacts that lack cross-subset robustness.

The source of this blind spot is algebraic. CKA is a functional of the Gram matrix $XX^\top$: it is invariant to any transformation preserving inner products between samples, including orthogonal rotation of the feature space. RSA and Procrustes share this invariance structure. As a consequence, these metrics are insensitive to how geometric information is distributed across the coordinate axes, the basis in which a representation is actually read out. Two representations with identical Gram matrices, and therefore identical similarity scores, can distribute that geometry very differently across coordinates: one may encode its pairwise structure redundantly, so that any subset of features recovers it, while the other distributes it non-redundantly, so that random feature halves recover conflicting structure. The first is geometrically stable and the second is not, yet similarity metrics, invariant to the basis by construction, score them identically. Whether geometry is redundantly distributed across the coordinate basis, and not how it is distributed across the eigenspectrum, is what determines stability.

We introduce *geometric stability* as a formalization of this missing property, and present Shesha, a metric that quantifies it through split-half correlation of representational dissimilarity matrices (RDMs). RDMs were introduced by Kriegeskorte et al. (2008) as the

foundation of RSA and provide a stimulus-resolved, model-agnostic summary of representational geometry that is invariant to linear transformations of the feature space. This invariance makes them applicable without modification across domains where feature spaces are fundamentally incommensurable, from neural population vectors to protein sequence embeddings to transformer hidden states. For a representation matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ of n samples in d dimensions, the RDM $\mathbf{D} \in \mathbb{R}^{n \times n}$ captures pairwise dissimilarities between samples, so that each entry D_{ij} records how dissimilar the representations of samples i and j are (Sec. 2, Eq. 1). In RSA, both rows and columns index experimental conditions, and the RDM is compared across systems to assess whether two populations encode the same pairwise structure. Shesha constructs RDMs in the same way, but computes two such matrices from complementary halves of the d feature dimensions rather than from the full representation, asking whether the sample-level geometry recovered from one half of the feature basis agrees with that recovered from the other.

Shesha, named for the Hindu deity representing the invariant remainder of the cosmos (Vogel, 1926; Daniélou, 1964; Dimmitt and van Buitenen, 1978), quantifies this self-consistency as the average Spearman rank correlation between RDMs constructed from complementary random partitions of the feature dimensions, averaged over K independent splits (Sec. 2, Eq. 2). A key formal property distinguishes this metric from the similarity family: Shesha is *not* invariant to orthogonal transformations of the feature space (Sec. 2.3 and Appendix B). This non-invariance is not a limitation but a design property. A full RDM is itself rotation-invariant, but computing RDMs on complementary feature subsets forfeits that invariance by construction, since a subset of rotated coordinates is not the rotation of a subset. That is what makes the metric sensitive to the coordinate-basis distribution of geometric information that similarity metrics cannot see.

This positive framing also separates our contribution from recent critiques of the similarity axis. Where Cloos et al. (2025) analyze which dimensions CKA underweights when comparing two representations, Shesha measures a different quantity on a single representation: whether its pairwise geometry is redundantly encoded across the feature basis, so that independent subsets recover the same structure. The two are related, since both concern how a representation distributes variance rather than its overall similarity, but they answer different questions, and Shesha carries a predictive payoff that an analysis of CKA does not: it forecasts the reliability of the subset-based probes and interventions on which interpretability depends.

We validate geometric stability across three scales of analysis. First, at the level of controlled geometric interventions across 2,463 encoder configurations in seven domains spanning language models, vision systems, audio and video encoders, protein sequence representations, molecular profiles, and neural population recordings, the relationship between stability and similarity is governed by transformation regime: geometry-preserving transformations render the two metrics redundant ($\rho = +0.75$), while compression couples them negatively ($\rho = -0.47$), the regime in which CKA stays high while geometric stability collapses. Pooled across regimes the net correlation is near zero (Spearman $\rho = -0.01$, 95% CI $[-0.06, +0.03]$), but the net is uninformative: the signal is the regime split, not its average. Second, at the level of mechanism: two opposite spectral manipulations form a double dissociation. Removing the single top principal component collapses CKA to 0.27 while Shesha holds at 0.95; retaining only the top components recovers CKA while driving Shesha below

zero. CKA follows the leading components while Shesha responds to structure distributed across the basis, a controlled signature of the same basis-dependence that separates the two metrics. Third, at the level of practical consequence: applied to 170 pretrained vision models across six datasets, geometric stability exposes a dissociation that both similarity and accuracy miss. DINOv2, among the two most transferable models on three of six datasets, ranks in the bottom quartile of geometric stability on five of six (all but EuroSAT). This is not a general transferability-stability trade-off: across the 36 architectural families the two are not traded off (Theil-Sen $\rho = +0.21$, not significant), and contrastively aligned models such as CLIP reach high transfer and high stability together. DINOv2 is an isolated dissociation, and it does not arise from a concentrated eigenspectrum; it has the highest participation ratio in the benchmark. What sets it apart is how it distributes that variance across its coordinate basis, leaving its geometry poorly recoverable from random feature subsets.

This carries a concrete warning for interpretability. A widely used foundation model, DINOv2, has among the least recoverable geometry in the benchmark, so probes, patching, and steering applied to it operate on exactly the feature-level structure that is least reliable, even though its similarity scores and transfer performance give no hint of the problem. The risk is specific rather than universal: contrastively aligned foundation models such as CLIP are both transferable and geometrically stable. We find that contrastive alignment predicts higher stability across all six datasets, while a hierarchical, multi-scale architecture helps on one (Flowers-102), identifying the training objective as the dominant determinant of geometric stability. By quantifying whether learned structure is recoverable from feature subsets, geometric stability becomes a prerequisite for robust mechanistic interpretability and reliable model steering.

The convergence of this pattern across artificial and biological systems, from transformer language models to protein sequence encoders to neural population recordings, suggests that geometric stability, the redundancy of a representation's geometry across its coordinate basis, is a substrate-independent axis of representational structure, distinct from similarity and from transferability, with implications for model selection, representational analysis, the reliability of mechanistic interpretability interventions, and the design of training objectives that preserve it.

2 The Geometric Stability Framework

We introduce geometric stability, a property of a single representation that measures how reliably its pairwise distance geometry can be recovered from complementary halves of its feature dimensions. The subsections below define the $\text{Shesha}_{\text{FS}}$ estimator and characterize the invariances it does and does not inherit.

2.1 Formal Definition

Let $X \in \mathbb{R}^{n \times d}$ be a representation matrix of n samples in d dimensions. A random feature partition π_k divides the index set $\{1, \ldots, d\}$ into two complementary halves A_k and B_k of size $\lfloor d/2 \rfloor$ and $\lceil d/2 \rceil$ respectively. For each half, we construct an RDM using pairwise cosine

distances:

$$D_{ij}^{(k,s)} = 1 - \frac{x_i^{(s)} \cdot x_j^{(s)}}{\|x_i^{(s)}\| \, \|x_j^{(s)}\|}, \quad s \in \{A_k, B_k\}, \tag{1}$$

where $x_i^{(s)}$ and $x_j^{(s)}$ denote the subvectors of sample i and j restricted to the dimensions in half s. Geometric stability is then defined as the average Spearman rank correlation between the vectorized upper triangles of the two half-RDMs across K independent partitions:

$$\mathcal{S}(X) = \frac{1}{K} \sum_{k=1}^{K} \rho_s \Big(\mathrm{vec}(D^{(k,A_k)}),\ \mathrm{vec}(D^{(k,B_k)}) \Big). \tag{2}$$

$\mathcal{S}(X) \in [-1, 1]$, with $\mathcal{S}(X) \approx 1$ indicating that complementary feature subsets recover the same pairwise geometry, and $\mathcal{S}(X) \approx 0$ indicating that the distance structure is not consistently recoverable from partial observations of the feature basis. We use $K = 30$ throughout. The use of Spearman correlation makes $\mathcal{S}$ invariant to monotonic rescaling of individual distances, and the use of cosine distance makes it invariant to global scaling of the representation. The full invariance structure is summarized in Table 1 and established formally in Sec. 2.3.

2.2 Basis Interpretation

The feature-split procedure in Eq. (2) asks whether random halves of the coordinate axes recover the same pairwise geometry. What it measures is how geometric information is distributed across the coordinate basis of X, not a property of the eigenspectrum alone.

To see the distinction, write the covariance $\Sigma = \frac{1}{n} X^\top X$ as $\Sigma = V \Lambda V^\top$. Two factors govern split-half agreement: the eigenspectrum Λ, which fixes how many directions carry substantial variance, and the orientation V, which fixes how those directions map onto the d coordinates that a partition splits. Only their combination determines $\mathcal{S}$.

When variance is concentrated in a few coordinates, a random partition yields halves with asymmetric information content: one half projects onto high-variance directions and the other onto near-noise directions, the two RDMs disagree, and $\mathcal{S}$ is low. When variance is spread redundantly across coordinates so that each half recovers similar structure, $\mathcal{S}$ is high. Both statements concern the coordinate distribution, not the eigenvalues: an orthogonal rotation leaves Λ unchanged while redistributing variance across coordinates, and $\mathcal{S}$ moves with it (Section 2.3).

Consequently $\mathcal{S}$ is not a function of the eigenspectrum, and is not equivalent to spectral entropy or participation ratio, which are basis-invariant summaries of Λ. A representation can have a broad eigenspectrum yet a low $\mathcal{S}$ when its variance is spread non-redundantly, so that no half recovers the whole. Where $\mathcal{S}$ does track spectral structure is in the breadth of its response: unlike CKA, which is dominated by the leading components and collapses once they are removed, $\mathcal{S}$ stays sensitive to structure throughout the spectrum (Table S2).

2.3 Formal Dissociation from Similarity Metrics

The invariance structure of $\mathcal{S}$ differs fundamentally from that of existing similarity metrics, and this difference is the algebraic source of their empirical independence. We establish this through four transformations illustrated in Fig. 1; full proofs are given in Appendix B.

Table 1: Invariance properties of geometric 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 is dominated by the top eigenvalues of $XX^\top$, which PCA preserves. [b]CKA depends only on $XX^\top$. [c]Procrustes explicitly optimizes over the set of orthogonal matrices.

	Global Scaling	Feature Permutation	PCA Compression	Orthogonal Rotation	Monotonic Distance	Isotropic Scaling
$\text{Shesha}_{\text{FS}}$	✓	✓	×	×	✓	✓
Linear CKA	✓	✓	✓[a]	✓[b]	×	✓
Procrustes	✓	✓	✓	✓[c]	×	✓
CCA	×	×	×	×	×	✓
PWCCA	×	×	✓	×	×	✓

2.3.1 Global Scaling

Cosine distance normalizes sample magnitudes, so $\mathcal{S}(\alpha X) = \mathcal{S}(X)$ for any $\alpha > 0$. CKA shares this invariance. See Fig. 1A.

2.3.2 Feature Permutation

Relabeling coordinate indices does not change representational content. Since partitions are drawn uniformly at random over coordinate indices, the distribution over partitions is exchangeable under permutation, and $\mathcal{S}$ is invariant. CKA is invariant for the same reason as orthogonal rotation: permutation matrices are orthogonal, so $XX^\top$ is preserved. See Fig. 1B.

2.3.3 PCA Compression

Projecting X onto its top $r < d$ principal components preserves dominant variance and leaves $XX^\top$ approximately unchanged for large r, so CKA is approximately invariant. $\mathcal{S}$ falls sharply: the projected representation concentrates all geometric information into r coordinates, and any partition that places those coordinates asymmetrically across its two halves produces maximally disagreeing RDMs. Concentrating variance into a coordinate subset is one basis configuration that lowers $\mathcal{S}$ while leaving CKA fixed. It is not the configuration behind the dissociation in Section 4: DINOv2's low stability instead reflects variance spread non-redundantly across many coordinates (the highest participation ratio in the benchmark), a different basis route to the same loss of recoverability. See Fig. 1C.

2.3.4 Orthogonal Rotation

Let $Y = XQ$ for $Q \in \mathcal{O}(d)$. Then $YY^\top = XQQ^\top X^\top = XX^\top$, so the Gram matrix is preserved and linear $\text{CKA}(X, Y) = 1$ for any orthogonal Q. $\text{Shesha}_{\text{FS}}$, by contrast, is not invariant: Q redistributes geometric information across the coordinate axes, so a random feature partition of Y captures different directions than the same partition of X, and $\mathcal{S}(Y) \neq \mathcal{S}(X)$ in general. This is the key formal dissociation: CKA is provably blind to redistributions of geometric information across the feature basis, while $\text{Shesha}_{\text{FS}}$ provably is

not, since an orthogonal Q can change $\mathcal{S}$ while holding CKA fixed. As a direct demonstration, rotating a stable representation into its own eigenbasis, an orthogonal transformation that preserves the eigenspectrum and rank exactly, collapses $\text{Shesha}_{\text{FS}}$ from 0.903 to near zero while leaving CKA at 1.000. The constructive counterexample and this numerical demonstration are given in Appendix B.6. See Fig. 1D.

2.3.5 Monotonic Distance Invariance

Spearman rank correlation depends only on the relative ordering of pairwise distances, not their magnitudes. Any strictly monotone transformation of the distance values preserves all rank orderings within each half-RDM, leaving $\mathcal{S}$ unchanged. CKA, which operates on inner products rather than ranks, is not invariant.

2.3.6 Isotropic Scaling

Isotropic scaling $X \mapsto \alpha X$ is a special case of global scaling; the argument is identical. CKA shares this invariance.

2.4 Connection to the RSA Noise Ceiling

The RSA noise ceiling (Nili et al., 2014) estimates the maximum Spearman correlation a model RDM can achieve with the data RDM, given measurement noise, by applying split-half logic across *observations*: odd and even trials, or subsets of subjects. A low noise ceiling indicates that the data RDM is unreliable due to measurement variability.

Shesha applies the same split-half logic across *features* rather than observations. A low $\mathcal{S}$ indicates that the representation's pairwise distance structure is not consistently recoverable from partial observations of the feature basis, which is a property of the representational architecture rather than of measurement quality. The two diagnostics are thus complementary: the noise ceiling audits data reliability; Shesha audits geometric reliability. Both are special cases of a general principle in which a self-consistency estimator is applied along one axis of the data matrix to characterize the structure along the other (see Appendix C for more details).

3 Distinctness of Stability and Similarity

A geometric stability measure earns its place only if it captures something representational similarity metrics miss. This section establishes that $\text{Shesha}_{\text{FS}}$ is distinct from CKA, RSA, and Procrustes distance both formally, through its non-invariance to basis rotations, and empirically, through controlled dissociations across thousands of encoder configurations.

3.1 Construct Validation

We first establish that $\mathcal{S}$ recovers known ground truth and that stability and similarity are separable by construction. Synthetic representations with parametrically controlled stability $\alpha \in [0, 1]$ (signal-to-noise mixing; Appendix E) confirm that Shesha recovers ground truth with near-perfect fidelity ($\rho = 0.997$, $p < 10^{-86}$). Balanced sampling across all four quadrants of the stability–similarity space, including adversarial cases where CKA > 0.97

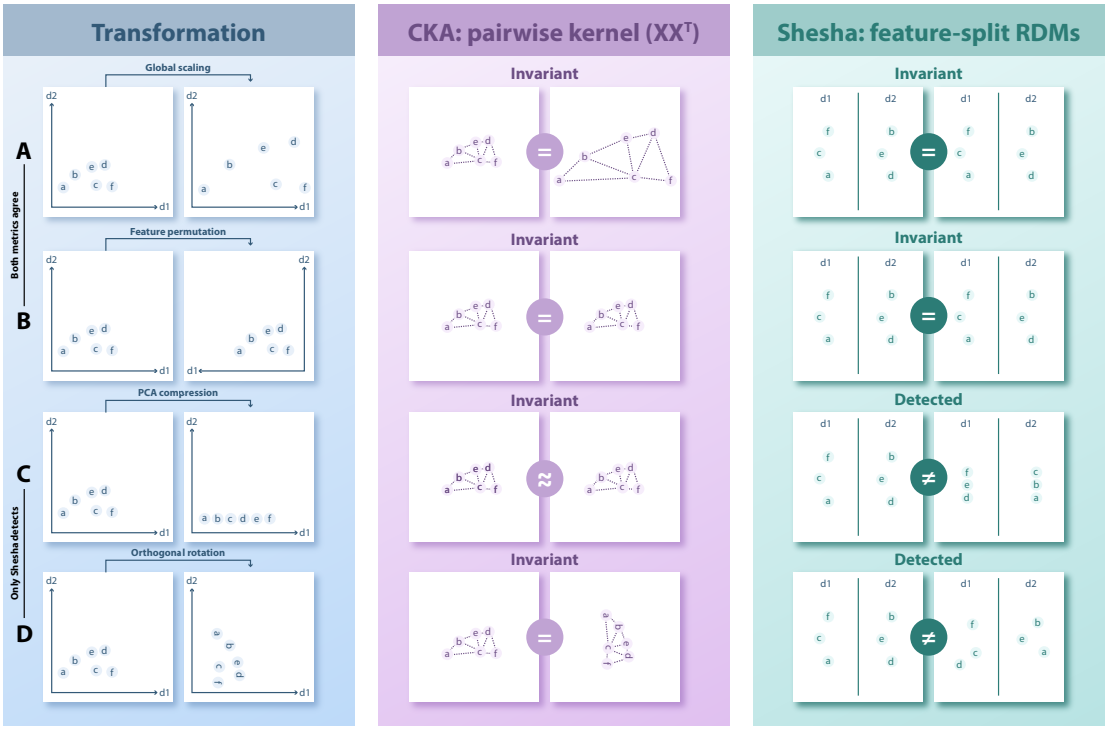

Figure 1: CKA and Shesha have complementary blind spots under geometric transformations. Each row applies a transformation to the same six-point representation, then shows how CKA (center) and Shesha (right) respond. CKA computes pairwise kernel alignment from the Gram matrix $XX^\top$; Shesha splits 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 C and D 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). Concentration into a coordinate subset is one basis route to low stability. 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.

despite near-zero $\mathcal{S}$, confirms that the two properties are separable: high similarity does not imply high stability, and vice versa (see Appendix E.9 for details).

3.2 CKA Tracks Dominant Variance & Shesha Tracks Full-Manifold Geometry

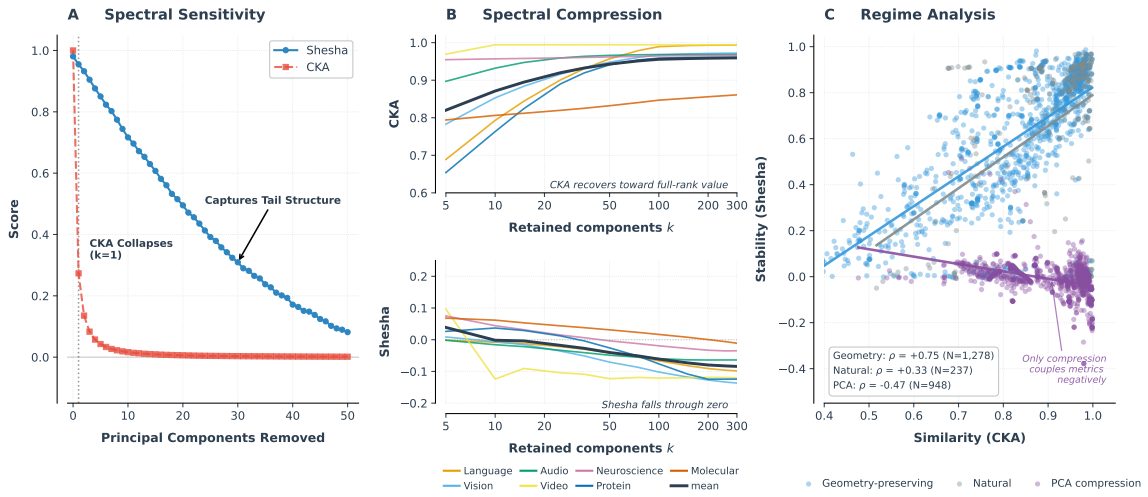

Figure 2: CKA tracks dominant variance; Shesha measures full-manifold geometry. Panels A and B apply opposite spectral manipulations and expose a double dissociation in which CKA follows the leading components and Shesha the distributed tail. A. Spectral sensitivity: removing the single top principal component collapses CKA (red) to 0.27, while Shesha (blue) remains at 0.95 and decays only gradually as further components are stripped. B. Spectral compression, the mirror: as retained components k increase, CKA recovers toward its full-rank value while Shesha falls from near zero into negative values, in all seven domains (curves are per-domain means). C. Regime analysis across 2,463 configurations in seven domains. The metrics agree for geometry-preserving transforms ($\rho = +0.75$, $N = 1{,}278$), couple weakly for natural encoders ($\rho = +0.33$, $N = 237$), and invert under PCA compression ($\rho = -0.47$, $N = 948$), configurations with higher CKA tend to have lower $\text{Shesha}_{\text{FS}}$, the across-configuration signature of CKA remaining high while stability degrades. A correlation pooled over the full suite is a composition-weighted average and is not interpreted here; the per-regime split is the result. Panel A is replicated with debiased CKA, Procrustes, and PWCKA in SI Fig. S3.

Having established the formal dissociation in Sec. 2.3, we now show it has an exact mechanistic basis and that the mechanism operates empirically at scale. CKA depends on the Gram matrix $XX^\top$ and is therefore determined by the dominant directions of variance; it is provably invariant to any transformation that preserves them (SI Appendix B). Shesha depends on whether independent halves of the feature set recover the same pairwise geometry, so it is sensitive to how that geometry is distributed across coordinates. Two complementary manipulations of the eigenspectrum expose the gap between what the two metrics can see, and together they form a double dissociation: in both, CKA follows the leading components while Shesha follows the distributed tail.

Removing leading components isolates CKA's dependence on the head of the spectrum. Using synthetic representations with a power-law eigenspectrum ($\lambda_i \propto i^{-1}$, mimicking trained networks; Appendix E.4), we progressively remove the top k principal components. Removing the single leading component collapses CKA from 1.0 to 0.27, while Shesha remains at 0.95 and decays only gradually as further components are stripped (Fig. 2A); at $k = 26$ removed, Shesha still carries roughly 92× the signal of CKA. Procrustes and PWCCA collapse identically to CKA (SI Fig. S3), so the blind spot is a property of the similarity-metric family rather than of one estimator. The divergence is robust across preprocessing, with the single exception of whitening, which equalizes the spectrum and partially restores CKA's sensitivity, exactly as the mechanism predicts (Appendix E.6). Independent analysis by Cloos et al. (2025) derives that CKA's sensitivity to a principal component scales with its variance, which is why CKA stays high under compression that discards low-variance but potentially informative directions. A complementary ablation confirms the specificity of this sensitivity: injecting scaled Gaussian noise into tail components carrying under 1% of total variance leaves $\text{Shesha}_{\text{FS}}$ above 0.95 even at 5× amplification, matching CKA's robustness (Appendix E.5).

PCA compression is the mirror manipulation, and it isolates the mechanism with a single knob. As the number of retained components k increases, CKA climbs monotonically back toward its full-rank value while Shesha moves in the opposite direction, falling from near zero into negative values (Fig. 2B). In vision representations, retaining $k = 300$ components restores CKA to its uncompressed value (0.972) while Shesha falls from +0.732 to −0.136; in language, CKA recovers to 0.993 while Shesha falls from +0.761 to −0.098. The opposite-direction monotonicity holds in every domain: within PCA, the Spearman correlation between retained dimension and CKA is positive (vision +0.84, language +0.91, audio +0.79) while the correlation between retained dimension and Shesha is negative (vision −0.86, language −0.94, audio −0.88; negative in all seven domains). This was not engineered to cancel; varying one knob inside one transform on one representation drives the metrics apart. CKA recovers because retained variance recovers. Shesha declines because PCA components are decorrelated and variance-ranked by construction, so each random feature-split half samples disjoint variance scales and the split-half geometry fragments, and adding components sharpens this fragmentation rather than repairing it.

Across the full panel of transformations, the two metrics coincide when pairwise distances are preserved and diverge when variance is concentrated (Fig. 2C). Geometry-preserving transforms (random projection, random feature subsets, feature selection, noise injection) couple the metrics positively ($\rho = +0.75$, $N = 1{,}278$): for random projection the Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2002) guarantees approximate distance preservation, and the remaining transforms preserve pairwise distances for analogous reasons, making the metrics redundant in this regime (per-transform values in Appendix Table S4). Natural encoders couple weakly ($\rho = +0.33$, $N = 237$), with Shesha contributing roughly 90% unique variance beyond CKA. PCA compression is the sole regime of negative coupling ($\rho = -0.47$, $N = 948$): concentrating variance into a low-dimensional, axis-aligned subspace holds CKA high while Shesha collapses. This is the controlled form of the dissociation we examine in trained models (Sec. 3), though there the gap arises from how variance is distributed across the coordinate basis rather than from low-rank compression.

This mechanism operates at scale. Across 2,463 encoder configurations in seven domains spanning machine learning (vision, language, audio, video) and biology (neuroscience, proteins, molecular), computed with linear CKA (Kornblith et al., 2019a) over 15 random seeds per configuration, the dissociation is reproduced domain by domain (Appendix Table S6). A mixed-effects model controlling for base-model identity attributes under 10% of stability variance to encoder identity (ICC = 0.10), ruling it out as a confound. The three domains with moderate correlations are negative, with sign and magnitude following from the regime split above; with sign and magnitude following from the regime split above. We do not report a pooled correlation as evidence of distinctness: because the suite mixes regimes that couple the metrics positively and negatively, any aggregate is a weighted average whose value is set by the composition of the suite rather than by a property of the metrics. The distinctness claim rests instead on the formal non-invariance (Sec. 2.3), the single-transform double dissociation above, and the per-regime correlations; the pooled and per-domain values are tabulated for completeness in Appendix Table S6.

3.3 Geometric Stability Extends to Biological Representations

The mechanism is substrate-independent. In protein sequence encoders, stability and similarity show moderate negative correlation ($\rho = -0.36$, 95% CI $[-0.45, -0.28]$), driven by PCA compression of low-dimensional encoders (20–500 dims): the controlled compression regime that anti-correlates the metrics elsewhere operates here too when dimensionality is reduced. In molecular profiles from single-cell RNA sequencing (pbmc3k), the correlation is negligible ($\rho = +0.06$), consistent with the natural encoder regime. In neural population recordings from 26 electrophysiology sessions spanning 68 brain regions (Steinmetz et al., 2019), 846 configurations yield $\rho = +0.01$ (95% CI $[-0.06, +0.09]$), among the tightest intervals in the dataset and the closest to zero, placing the representational geometry of sensory and motor cortices squarely in the natural encoder regime. That this pattern converges across systems trained by gradient descent, evolution, development, or biological learning suggests that the dissociation between geometric stability and similarity is a property of learned representations generally, not an artifact of deep learning optimization.

4 Geometric Stability in Pretrained Vision Models

A natural prediction is that high transferability is bought at the cost of geometric stability, since a representation optimized for downstream discriminability need not distribute that information redundantly across its coordinates. We test this across 170 pretrained vision models organized into 36 architectural families, evaluated on six datasets spanning four visual domains: natural images (CIFAR-10 and 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). Transferability is estimated via LogME (You et al., 2021, 2022), a label-efficient proxy for linear probing performance. To confirm these rankings are not artifacts of a single feature-partition seed, we recomputed the full CIFAR-10 sweep under three seeds (9, 320, and 1991) for all 170 models; $\text{Shesha}_{\text{FS}}$ is highly reproducible (Spearman $\rho \geq 0.993$, median per-model CV 0.75%; Appendix G.5).

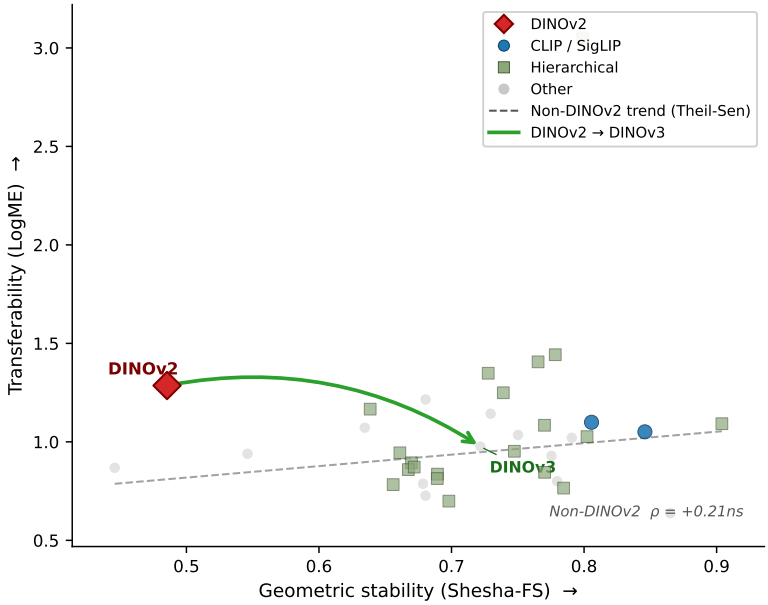

Figure 3: DINOv2 dissociates geometric stability from transferability; the population at large does not. Family-mean transferability (LogME) against geometric stability ($\text{Shesha}_{\text{FS}}$) for 36 architectural families, averaged across the six datasets. DINOv2 (red) is the lone outlier, sitting far to the low-stability side of the population while remaining highly transferable. Across the other 35 families the two quantities are not traded off: the robust Theil-Sen trend is flat to weakly positive ($\rho = +0.21$, not significant), and CLIP and SigLIP combine high stability with competitive transferability, so low stability is not a general cost of transfer. The arrow marks the generational change from DINOv2 to DINOv3, which under Gram anchoring returns from the outlier position onto the population trend. The per-dataset breakdown, including the EuroSAT exception where DINOv2 is itself highly stable, is given in Appendix S8.

4.1 The DINOv2 Paradox

Per dataset, the family rankings make the dissociation concrete (Table 2; Appendix Fig. S8). DINOv2 ranks first or second in transferability on three of six datasets (LogME rank 1/36 on Flowers-102, 2/36 on CIFAR-10 and CIFAR-100) while ranking last or next-to-last in geometric stability on the same three datasets (36/36, 35/36, and 36/36 respectively) and in the bottom quartile on Oxford Pets (33/36) and DTD (29/36; Table 2). The sole exception is EuroSAT, where DINOv2 achieves both high transfer and high stability ($\text{Shesha}_{\text{FS}} = 0.950$, rank 4/36). This ordering is seed-invariant: DINOv2 holds the lowest family-mean $\text{Shesha}_{\text{FS}}$ under all three CIFAR-10 seeds (Appendix G.5). On EuroSAT DINOv2's representation is also its most spectrally concentrated (top-eigenvalue share 0.206 and participation ratio 16.1, against 0.045 and 99.0 on CIFAR-10), so the exception is consistent with the relationship in Section 4.4, where greater concentration accompanies higher, not lower, stability. DINOv2 is

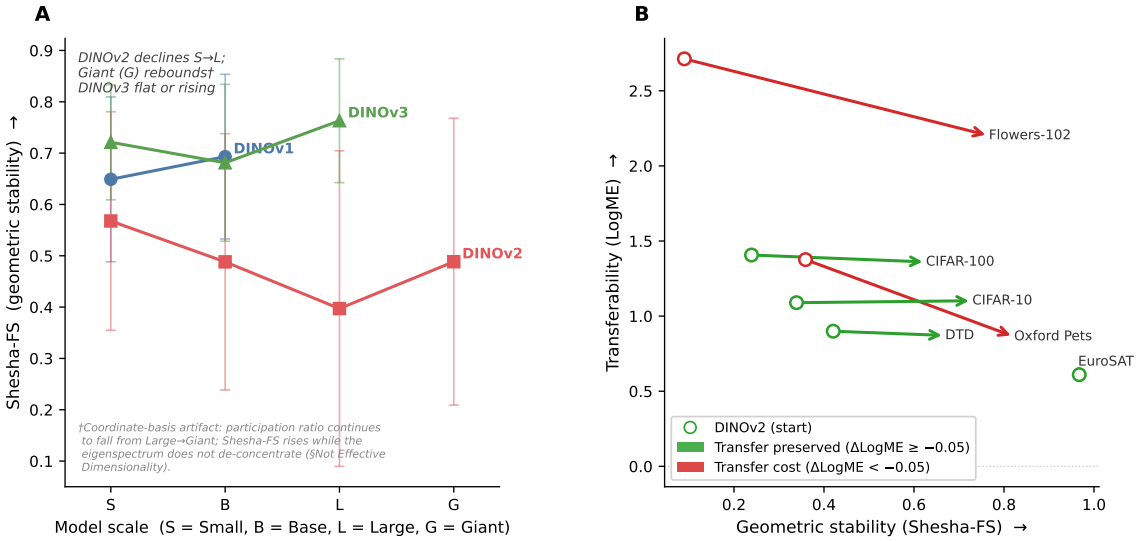

Figure 4: Gram anchoring closes the DINOv2 dissociation. A. Geometric stability ($\text{Shesha}_{\text{FS}}$) against model scale, mean ± SD across the six datasets. DINOv2 stability falls from small to large; the apparent rebound at giant is a coordinate-basis artifact, since the participation ratio keeps falling from large to giant and the eigenspectrum does not de-concentrate (Section 4.4). DINOv3 stays flat or rises, and DINOv1 (small and base only) sits above DINOv2 throughout, so the dissociation tracks the training objective rather than model scale. B. Matched at the large scale, the change from DINOv2 to DINOv3 on the stability-transfer plane, one arrow per dataset. On four of six datasets DINOv3 gains stability with transfer preserved ($\Delta\text{LogME} \geq -0.05$, green); on Flowers-102 and Oxford Pets the stability gain carries a transfer cost (red). EuroSAT shows no arrow because both generations are already highly stable there.

the extreme case of a broader stability ordering: self-supervised models trained with masked image modeling or self-distillation are less geometrically stable than contrastively aligned ones (Section 4.3). What singles out DINOv2 is that it pairs that instability with top-tier transferability; the other low-stability families do not transfer as well, so they stay on the population trend rather than off it.

This dissociation is invisible to CKA, which depends on the Gram matrix, dominated by the top eigenvalues regardless of how the remaining variance is distributed across coordinates (Section 3). What sets DINOv2 apart is not a concentrated eigenspectrum: it has the highest participation ratio in the benchmark (Section 4.4). It is instead how that variance is distributed across the learned coordinate basis, since a representation can be high-rank yet recover poorly from random coordinate subsets. The principle that coordinate-basis distribution, rather than eigenvalue concentration, governs stability is established in Section 4.4 and demonstrated under controlled conditions by the PCA-compression analysis (Appendix B.5) and the optimizer ablation (Section 4.6).

Table 2: DINO generations across six datasets ($N = 36$ families per dataset). For each dataset and generation, family-mean LogME and $\text{Shesha}_{\text{FS}}$ with rank among 36 families. DINOv2 attains top transferability ranks while ranking last or near-last in stability, except on EuroSAT; DINOv3 recovers stability rank under Gram anchoring, at some cost in transferability.

Dataset	DINO Generation	LogME	LogME Rank	$\text{Shesha}_{\text{FS}}$	FS Rank
CIFAR-10	v1	0.425	24/36	0.642	27/36
	v2	1.013	2/36	0.369	36/36
	v3	0.819	9/36	0.688	19/36
CIFAR-100	v1	1.045	28/36	0.581	29/36
	v2	1.373	2/36	0.266	35/36
	v3	1.206	9/36	0.583	28/36
Flowers-102	v1	1.260	21/36	0.852	19/36
	v2	2.586	1/36	0.320	36/36
	v3	1.709	9/36	0.765	27/36
DTD	v1	0.692	25/36	0.448	32/36
	v2	0.876	13/36	0.472	29/36
	v3	0.814	15/36	0.598	17/36
EuroSAT	v1	0.547	10/36	0.775	35/36
	v2	0.568	6/36	0.950	4/36
	v3	0.588	4/36	0.922	8/36
Oxford Pets	v1	0.755	31/36	0.773	10/36
	v2	1.301	12/36	0.535	33/36
	v3	0.717	33/36	0.774	9/36

4.2 DINOv3 Closes the Dissociation

DINOv3 (Siméoni et al., 2025) provides a natural test of whether DINOv2's low stability is an inherent property of self-distillation or a correctable design choice. DINOv3 introduces Gram anchoring, a training modification designed to prevent dense feature degradation during long training schedules. Viewed through the lens of geometric stability, Gram anchoring acts as an implicit regularizer that preserves coordinate-basis redundancy. Averaged over the six datasets, DINOv3 reaches substantially higher $\text{Shesha}_{\text{FS}}$ than DINOv2 (0.722 vs. 0.485) at a lower mean transferability (LogME 0.976 vs. 1.286; Table 2), and improves on DINOv1 in both. This is not simply transfer traded for stability. The scaling behavior reverses: DINOv2's stability collapses as parameter count increases (0.571 at small to 0.402 at large), whereas DINOv3 remains stable across scales (0.721 at small to 0.763 at large). Matched at the large scale, the stability gain comes at little or no transfer cost on four of six datasets: on CIFAR-10 DINOv3 reaches $\text{Shesha}_{\text{FS}}$ 0.730 against DINOv2's 0.339 at equal transferability (LogME 1.102 vs. 1.089), with the same pattern on CIFAR-100, DTD, and EuroSAT; only on Flowers-102 and Oxford Pets does the gain still involve a transfer reduction. Matching the scale roughly halves the apparent transfer gap (large-scale mean LogME 1.349 for DINOv2 vs. 1.183 for DINOv3), so the family-average gap in Table 2 is inflated by DINOv2's giant variant, which DINOv3 does not include. Register variants of DINOv2, designed to address patch artifacts, consistently though modestly reduce stability at every scale (for example

0.402 vs. 0.392 at large). Together, these generations indicate that the dissociation can be closed by explicitly anchoring the structural redundancy of the feature space.

4.3 Architectural and Training Determinants

Geometric stability varies systematically with architecture and training objective. Contrastive alignment predicts high stability: CLIP-family models outperform self-supervised models on all six datasets (Mann-Whitney $p < 0.05$ on every dataset), and EVA-02, which reconstructs CLIP features rather than raw pixels, ranks among the most stable models on most benchmarks. The alignment target, not the training mechanism, determines geometric stability. Hierarchical architecture provides a complementary but dataset-dependent route: Swin, PVT, and CoAtNet significantly exceed isotropic ViT and DeiT on Flowers-102 ($p < 0.001$), though this advantage does not reach significance on the other five datasets, indicating that the benefit of multi-scale processing is contingent on the visual domain. Cross-dataset rank consistency ($\rho = 0.95$ between CIFAR-10 and CIFAR-100) confirms that geometric stability is an intrinsic architectural property rather than a dataset-specific artifact.

4.4 Geometric Stability is Not Effective Dimensionality

A natural objection is that $\text{Shesha}_{\text{FS}}$ merely restates the effective dimensionality of a representation: one that spreads variance over many dimensions might be expected to divide into two informative feature halves, while a low-dimensional one would not. The data rejects this. Across all 170 models and six datasets, $\text{Shesha}_{\text{FS}}$ is *negatively* correlated with the participation ratio (mean Spearman $\rho = -0.36$; all six datasets significant at $p < 0.01$, from -0.29 on CIFAR-10 to -0.47 on EuroSAT) and positively correlated with the top-eigenvalue variance share (mean $\rho = +0.39$). Representations that use *more* effective dimensions are, if anything, *less* recoverable from random coordinate subsets.

The DINOv2 family makes the dissociation concrete. On the natural-image datasets it records the lowest $\text{Shesha}_{\text{FS}}$ in the benchmark (rank 36/36 on CIFAR-10) yet the highest participation ratio of all 36 families there (98.98 on CIFAR-10 against a benchmark mean of 51.64; 227.85 on CIFAR-100). A representation can therefore occupy nearly twice the effective dimensionality of the typical model while remaining the least geometrically stable. The scale ladder supplies a within-family divergence in the opposite direction: across all six datasets the giant variant has a *lower* participation ratio than the large (87.9 versus 117.7 on CIFAR-10) while its $\text{Shesha}_{\text{FS}}$ is *higher*. Effective dimensionality and geometric stability move in opposite directions here.

The reason is that the participation ratio is a rotation-invariant function of the eigenvalue distribution, whereas $\text{Shesha}_{\text{FS}}$ is a basis-dependent function of how variance is distributed across the model's learned coordinates. The two coincide only when the eigenbasis is aligned with the coordinate axes. PCA compression is exactly that aligned limit, which is why projecting onto leading principal components drives $\text{Shesha}_{\text{FS}}$ down (Section 3.2): it concentrates variance and rotates it onto coordinate axes at once. Learned representations need not behave this way. A model can spread variance across many eigendirections, raising its participation ratio, while distributing that variance non-redundantly across coordinates,

lowering $\text{Shesha}_{\text{FS}}$. Geometric stability and effective dimensionality are thus distinct, and in this benchmark anti-correlated, properties.

4.5 Geometric Stability is Distinct From Corruption Robustness

We next asked whether a representation's clean geometric stability anticipates its robustness to distribution shift. For each model we computed ΔLogME, the drop in LogME from the clean to the corrupted evaluation (CIFAR-10-C and CIFAR-100-C (Hendrycks and Dietterich, 2019), severity 5, 19 corruption types), and related it to clean $\text{Shesha}_{\text{FS}}$ by partial Spearman correlation, controlling for clean LogME so that the relationship is not driven by high-transfer models simply having more to lose ($N = 170$ models per dataset; bootstrap 95% confidence intervals, 10,000 resamples).

The predicted relationship does not appear. If geometric stability conferred robustness, more stable models would degrade less and the correlation would be negative. On CIFAR-100 there is no relationship in either direction (partial $\rho = -0.03$, 95% CI $[-0.20, +0.27]$ across all 19 corruptions). On CIFAR-10 the aggregate correlation is weakly positive and significant (partial $\rho = +0.26$, 95% CI $[+0.08, +0.43]$), the opposite of the predicted sign: more geometrically stable models degrade slightly more, not less. This effect is small and does not replicate on CIFAR-100, so we do not read it as evidence that stability harms robustness; we read the pair of results as the absence of any consistent link between the two.

This places $\text{Shesha}_{\text{FS}}$ precisely. Together with the subset-reliability result of Section 4.7, it shows that $\text{Shesha}_{\text{FS}}$ predicts whether a representation's geometry is recoverable from a random subset of its coordinates, a redundancy property internal to the clean representation, but does not predict how that representation fares under input distribution shift. Geometric stability is therefore distinct from corruption robustness, as it is from accuracy (Section 4.6) and from CKA (Section 3). $\text{Shesha}_{\text{FS}}$ is a diagnostic of representational redundancy, not a general proxy for representation quality.

4.6 Optimizer Geometry Modulates Stability

To test whether geometric stability reflects properties of the optimization landscape rather than learned features alone, we trained ResNet-18 models on CIFAR-10 and CIFAR-100 under identical conditions, varying only the Sharpness-Aware Minimization (SAM; Foret et al. 2021) perturbation radius $\rho \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2\}$, where $\rho = 0$ recovers standard SGD. Each configuration was run over 15 random seeds, and we report the mean and standard deviation of each metric across seeds (Table 3).

Test accuracy remained nearly constant across the sweep (94.91–95.56% on CIFAR-10; 76.96–78.05% on CIFAR-100), yet CKA between each SAM model and its SGD baseline fell steadily as ρ increased, to 0.925 on CIFAR-10 and 0.765 on CIFAR-100. $\text{Shesha}_{\text{FS}}$ did not track this decline: on CIFAR-10 it rose from 0.806 (SGD) to a peak of 0.872 at $\rho = 0.05$, and on CIFAR-100 it rose from 0.805 to 0.822 at $\rho = 0.2$. The effect is consistent across seeds: at the peak radius, $\text{Shesha}_{\text{FS}}$ exceeds the SGD baseline on all 15 seeds for both datasets (CIFAR-10, $\rho = 0.05$: $+0.067$, paired $t_{14} = 25.3$, $p < 10^{-12}$; CIFAR-100, $\rho = 0.2$: $+0.018$, $t_{14} = 16.6$, $p < 10^{-9}$; two-sided). SAM therefore moves the representation away from the SGD baseline in similarity terms while leaving it at least as geometrically stable, and over the relevant range more so.

This is controlled evidence that $\text{Shesha}_{\text{FS}}$ measures a property of representational geometry orthogonal to both accuracy and similarity: a training intervention can hold accuracy fixed and drive steady, monotonic declines in CKA, while triggering non-monotonic or thresholded shifts in $\text{Shesha}_{\text{FS}}$; on CIFAR-10 this takes the form of an interior optimum (an optimization sweet spot near $\rho = 0.05$–0.1), whereas on CIFAR-100 stability rises across the sweep. The supervised Shesha variants, which move opposite to $\text{Shesha}_{\text{FS}}$, and the full per-seed results are reported in Appendix G.6.

Table 3: SAM perturbation radius ablation. ResNet-18 trained on CIFAR-10 and CIFAR-100 with identical hyperparameters, varying only the SAM perturbation radius ρ ($\rho = 0$ is standard SGD). Values are mean $\pm$ SD over 15 seeds. Test accuracy stays approximately constant and CKA falls steadily, while $\text{Shesha}_{\text{FS}}$ does not track the CKA decline.

	CIFAR-10		CIFAR-100	
ρ	CKA vs. SGD	$\text{Shesha}_{\text{FS}}$	CKA vs. SGD	$\text{Shesha}_{\text{FS}}$
0.00 (SGD)	1.000	0.806 ± 0.008	1.000	0.805 ± 0.003
0.01	0.949	0.831 ± 0.008	0.772	0.805 ± 0.003
0.02	0.946	0.851 ± 0.005	0.773	0.805 ± 0.003
0.05	0.938	0.872 ± 0.007	0.777	0.806 ± 0.003
0.10	0.933	0.872 ± 0.008	0.775	0.814 ± 0.004
0.20	0.925	0.851 ± 0.020	0.765	0.822 ± 0.003

4.7 Geometric Stability Predicts Probe Subset-Sensitivity

Interpretability methods that operate on a subset of a representation's features, such as linear probes trained on part of the residual stream, implicitly assume that the probed subset recovers the same structure as the full representation. Geometric stability is a direct measure of whether this assumption holds. We tested this prediction directly.

For each of 170 vision models, we extracted clean CIFAR-10 representations and trained logistic-regression probes on 20 random halves of the feature dimensions, holding the train and test sample split fixed so that variability reflects feature choice alone. We then measured the standard deviation of probe test accuracy across the 20 subsets. If geometric stability governs subset reliability, low $\text{Shesha}_{\text{FS}}$ should predict high probe-accuracy variability.

Across the 170 models, $\text{Shesha}_{\text{FS}}$ correlates negatively with probe-accuracy standard deviation ($\rho = -0.30$, $p < 10^{-4}$). The relationship is not an artifact of probe accuracy itself: the partial correlation controlling for mean probe accuracy is stronger than the raw correlation ($\rho_{\text{partial}} = -0.38$, $p < 10^{-6}$), and the effect survives normalization by the coefficient of variation ($\rho = -0.28$, $p < 10^{-3}$). Representations with lower geometric stability yield probes whose measured accuracy depends materially on which feature subset is used, establishing that $\text{Shesha}_{\text{FS}}$ captures a property directly relevant to the reliability of subset-based interpretability analysis.

5 Discussion

Geometric stability is an axis of representational analysis that similarity metrics leave unmeasured, governed by how a representation distributes variance across its coordinate basis rather than by its distance geometry alone. We discuss what this distinction reveals for interpretability and model selection, the isolated transfer-stability dissociation it brings into view, and where the measure does and does not apply.

5.1 Two Axes of Representational Geometry

Representational analysis has, until now, operated along a single axis: similarity, the alignment between two representational spaces. The results presented here establish that a second axis exists, geometric stability, distinct from the first by formal proof: CKA and its relatives are invariant to orthogonal rotation of the feature basis, and therefore to how geometric information is distributed across coordinate axes, the very property that determines stability, so they are blind to it by construction. Empirically the two axes are not reducible to one another across 2,463 configurations in seven domains, where their relationship is governed by transformation regime rather than by any single correlation.

The trajectory of the field is instructive here. RSA (Kriegeskorte et al., 2008) abstracted from individual neural responses to pairwise dissimilarity matrices, enabling comparison across systems with different numbers of units (Nili et al., 2014; Walther et al., 2016; Diedrichsen and Kriegeskorte, 2017). Statistical inference methods for representational geometries followed (Schütt et al., 2023; Schütt, 2025). CKA (Kornblith et al., 2019a) provided a normalized kernel alignment measure invariant to orthogonal rotation, enabling systematic comparison across architectures (Kornblith et al., 2019b; Nguyen et al., 2021) and training regimes (Mehrer et al., 2020; Zhuang et al., 2021). Subspace alignment methods (SVCCA, Raghu et al. 2017; PWCCA, Morcos et al. 2018) and Procrustes analysis (Schönemann, 1966; Rohlf and Slice, 1990; Masarotto et al., 2018; Dryden and Mardia, 1998) enriched the toolkit further. Topological RSA (Lin and Kriegeskorte, 2024) then abstracted from geometry to topological features, demonstrating that geotopological summary statistics provide more robust signatures of computational function across brain regions and deep network layers. A recent synthesis (Sucholutsky et al., 2025) surveyed this landscape across cognitive science, neuroscience, and machine learning, proposing a unifying framework for representational alignment.

Geometric stability moves along a different direction entirely: rather than further abstracting the content of representations, it asks whether that content is structurally reliable. These are complementary additions to the same toolkit, not competing replacements. Concurrently, Cayco-Gajic and Pellegrino (2026) introduced metric similarity analysis on Riemannian manifolds, demonstrating that existing similarity metrics fail to capture intrinsic manifold geometry; geometric stability addresses a distinct blind spot: not the intrinsic versus extrinsic distinction, but whether the coordinate basis reliably encodes the geometry at all. Concurrent work continues to fragment the similarity axis into distinct sub-properties: Dhimoïla et al. (2026) show that concept alignment is multi-objective, with translation and concept consistency failing to imply one another, so that a single alignment score conflates properties that must be measured separately. Geometric stability is orthogonal to this de-

composition as well: it concerns not how two systems' concepts correspond, but whether a single system's geometry is reliably encoded across its feature basis.

Williams (2024) recently demonstrated that RSA and CKA are largely equivalent once mean-centering is incorporated into the RSA computation, unifying two frameworks the community had treated as distinct. This equivalence reinforces the point that the similarity axis is well understood: whether one computes centered kernel matrices or centered dissimilarity matrices, the resulting scores capture the same geometric relationship between two representations. Shesha operates along a different axis entirely. It is not a similarity measure comparing two representations but a stability diagnostic applied to a single representation, correlating RDMs computed on complementary feature partitions for self consistency rather than on distinct neural systems. The RSA–CKA equivalence therefore does not extend to Shesha: the split-half feature partitioning, the use of Spearman rank correlation (which is not equivalent to linear CKA), and the single-representation setting all place Shesha outside the scope of Williams' unification.

Moreover, the reliability of the similarity axis itself has been questioned: Davari et al. (2023) demonstrated that CKA values can be directly manipulated without altering models' functional behavior, calling for caution when interpreting alignment metrics. Geometric stability sidesteps this concern entirely: it assesses a single representation's internal consistency rather than comparing two representations via a potentially manipulable score.

Recent work independently underscores that global alignment measures leave important structure uncharacterized. Conwell et al. (2024) found that architecturally diverse models achieve near-equivalent brain predictivity despite clear variation in their underlying representations, suggesting that standard alignment methods may be too flexible to distinguish meaningfully different computational strategies. Feather et al. (2023) showed that models matching brain representations can nonetheless learn fundamentally divergent invariances, a failure mode invisible to standard alignment benchmarks. Avitan and Golan (2025) demonstrated that even with millions of behavioral trials, linear alignment fails to recover the data-generating model, with misidentification driven by shifts in representational geometry and effective dimensionality that flexible metrics cannot resolve. Muttenthaler et al. (2025) demonstrate that vision models fail to capture human-like hierarchical abstraction despite high overall alignment scores, while 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 distinction between content and reliability recapitulates the classical separation of validity and reliability in psychometrics (Cohen, 1988): a test may measure the right construct yet produce inconsistent scores across administrations. A parallel principle appears in data science, where Yu and Kumbier (2020) establish stability alongside predictability and computability as a foundational requirement for veridical inference. The noise ceiling (Nili et al., 2014) formalized a related idea for neural data: split-half correlation across observations bounds how well any model can account for an empirical RDM given measurement noise. Shesha applies the same split-half logic across features rather than observations, diagnosing representational architecture rather than data quality.

This relationship suggests a practical protocol: geometric stability should be assessed before any similarity analysis is conducted. A representation with low $\mathcal{S}$ has a pairwise geometry that is not reliably recoverable from independent subsets of its feature basis. A similarity comparison involving such a representation is comparing a potentially unrepresentative snapshot of the geometry, not the geometry itself. Just as the noise ceiling (Nili et al., 2014) bounds how well any model can account for an empirical RDM given measurement noise, $\mathcal{S}$ bounds how much of the representational geometry is accessible from any single observation of the feature basis. Reporting $\mathcal{S}$ alongside RSA or CKA scores would allow the field to distinguish cases in which two representations genuinely differ from cases in which one or both representations are too geometrically fragile for the comparison to be meaningful.

The regime analysis clarifies when each axis adds unique information. Under geometry-preserving transformations, stability and similarity are redundant: either suffices, as the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2002) guarantees approximate pairwise distance preservation. Under compression, the controlled regime in which variance is forced into a coordinate subspace, they anti-correlate: similarity remains high while stability falls. Stability provides diagnostic value precisely where it diverges from similarity, and that divergence is exactly what a similarity score cannot reveal for a deployed model whose geometry may not be recoverable from feature subsets.

A natural objection is that a metric sensitive to orthogonal rotation measures an arbitrary basis rather than intrinsic geometry. This is the right concern and the wrong conclusion for the representations we study. Rotation invariance is the correct property only when the basis is itself arbitrary, but for learned representations the feature basis is privileged: it is the basis in which the model computes and the basis on which interpretability and deployment act. Linear probes read coordinate subsets, steering directions are applied along specific axes, and pruning and dropout delete specific units. A rotation-invariant metric is by construction blind to whether the structure these methods target survives the coordinate-level perturbations they impose; geometric stability measures precisely that. Its basis dependence is therefore appropriate rather than incidental, because the reliability it predicts is itself basis-relative.

One boundary condition deserves emphasis: whitening the representation equalizes the eigenspectrum by construction, restoring CKA's sensitivity to coordinate-level structure (Fig. S3D). In the whitened setting, stability and similarity become partially redundant. The practical distinctness of Shesha therefore applies specifically to the unwhitened representations that practitioners overwhelmingly use in transfer learning and zero-shot deployment.

5.2 Geometric Stability as a Distinct Selection Axis

The DINOv2 dissociation is not an anomaly, but neither is it an instance of a general law. Across the 36 architectural families, transferability and geometric stability are not traded off (Theil-Sen $\rho = +0.21$, not significant): contrastively aligned models such as CLIP and SigLIP reach high transfer and high stability together, and most families sit on a flat-to-weakly-positive trend. DINOv2 is the lone family that combines top-tier transfer with

bottom-quartile stability, and it does so without a concentrated eigenspectrum, since it has the highest participation ratio in the benchmark. The dissociation reflects how DINOv2 distributes variance across the coordinate basis, leaving the geometry poorly recoverable from feature subsets, rather than any concentration of the spectrum that a similarity metric or an effective-dimensionality measure would register.

This dissociation is not in tension with the known benefits of high-dimensional geometry. Sorscher et al. (2022) show that high-dimensional concept manifolds improve few-shot learning of novel concepts, letting new categories be acquired from fewer examples. Geometric stability is a distinct axis: spreading variance across many dimensions aids this separability, but says nothing about whether that variance is distributed redundantly enough for the pairwise geometry to survive coordinate subsetting. DINOv2 sits at exactly this corner, high participation ratio, high transfer, low $\text{Shesha}_{\text{FS}}$, which is why dimensionality-based accounts of representational quality and geometric stability must be measured separately.

Because stability is a separate axis rather than a fixed cost of transfer, it belongs in model selection as its own criterion. Current benchmarks, including LogME, LEEP (Nguyen et al., 2020), visual task adaptation suites (Zhai et al., 2019), and holistic evaluation frameworks (Liang et al., 2023), score a single axis and cannot surface a dissociation like DINOv2's. A practitioner training a linear head on a known downstream task should optimize for transferability; a practitioner deploying a model whose representation will be probed, steered, or pruned, that is, subjected to feature-level interventions, should also check stability, because a high-transfer model can still have geometry that fractures under exactly those coordinate-level perturbations. The relevant failure mode is at the level of these interventions, not input distribution shift, which stability does not predict (Section 4.5). Neyshabur et al. (2020) showed that successful transfer depends on both feature reuse and convergence to a shared basin; geometric stability adds that the transferable information may be distributed non-redundantly across coordinates and so be vulnerable to perturbation, a property invisible to existing transfer metrics.

The dissociation is also correctable. DINOv3, whose Gram anchoring acts as an implicit regularizer on coordinate-basis redundancy, recovers stability at little or no transfer cost on four of six datasets and Pareto-dominates DINOv1 (Section 4.2); because this is a targeted training modification rather than a change of model family, it is the closest evidence we have that coordinate-basis redundancy is the manipulable lever and that the deficit is not inherent to high-transfer self-supervised training. We name this correctable, objective-linked penalty the geometric tax: a stability cost that surfaces under some training objectives, is absent under others, and that a change of objective can repeal. The term labels this phenomenon, not the general trade-off the family-level data rules out, and the link to any single objective is associational rather than established by intervention (Section 5.5). Contrastive alignment shows the same association: CLIP-family models are more stable than single-modality self-supervised models on all six datasets, and EVA-02, which reconstructs CLIP features rather than raw pixels, is among the most stable models in the benchmark, so stability tracks the alignment target rather than the training mechanism. Whether a regularizer that directly targets coordinate-basis redundancy can close the remaining gap on the datasets where DINOv3 still pays a transfer cost is an open and precisely posed question.

5.3 Relation to Mechanistic Interpretability

Mechanistic interpretability methods, such as linear probes (Alain and Bengio, 2017), causal tracing and activation patching (Meng et al., 2022), and steering vector interventions (Zou et al., 2023; Turner et al., 2023), share a common implicit assumption: that the geometric structure they identify in a representation is consistent enough to support the intervention being applied. A linear probe trained on a subset of residual stream dimensions implicitly assumes that the probed subspace carries the same information as the full representation, an assumption of orthogonal invariance that, as demonstrated in Section 2, is routinely violated in practice. A steering vector applied along a direction found by difference-in-means assumes that direction is robustly encoded across the feature basis, rather than concentrated in a fragile subspace (such as those arising from superposition (Elhage et al., 2022)) that a small perturbation could destroy. A parallel concern applies to explanation and visualization methods. Fel et al. (2022) showed that saliency-based explanations require their own stability guarantees, and Geirhos et al. (2024) demonstrated that feature visualizations can be manipulated to display arbitrary patterns disconnected from a network's actual behavior, proving that the class of functions reliably explained by feature visualization is vanishingly small. These findings establish that interpretability tools rest on implicit reliability assumptions; geometric stability provides a representation-level diagnostic for when those assumptions are likely to hold.

Shesha makes these assumptions explicit and testable. A low stability score is a direct warning that the geometric structure a probe or steering vector targets may be an artifact of which features happen to be measured, rather than a robust, linear property of the global representation (Park et al., 2023). This is not merely a conceptual concern: across 170 vision models, $\text{Shesha}_{\text{FS}}$ predicts the degree to which linear probe accuracy depends on the feature subset used (Section G.7), confirming that geometric stability predicts the reliability of subset-based linear probing. Activation patching and steering act on the same object, structure localized to particular coordinate directions, so the diagnostic extends to them by the same mechanism; we do not test those interventions directly, and doing so is a natural next step. Conversely, high stability provides positive evidence that an identified circuit or direction generalizes beyond the specific measurement context in which it was found. The DINOv2 finding adds a further caution: some foundation models commonly used as test beds in mechanistic interpretability (Bommasani et al., 2021) have low geometric stability, so the implicit assumption underlying subset-based methods can be violated in exactly the models the field studies most. The risk is model-specific rather than universal, since contrastively aligned foundation models are geometrically stable, which makes stability a useful screen for choosing reliable test beds. This concern is not hypothetical: Zimmermann et al. (2023) found that larger, more accurate vision models are no more mechanistically interpretable than a decade-old GoogLeNet, with the most modern models appearing even less interpretable, sacrificing interpretability for accuracy. Geometric stability offers a candidate explanation: models like DINOv2 distribute representational geometry non-redundantly across the coordinate basis, lowering the cross-subset consistency that interpretability methods assume. This pattern tracks the training objective rather than transfer performance, though we do not isolate the objective's causal role here.

One natural concern is whether low $\text{Shesha}_{\text{FS}}$ in such models reflects genuine geometric fragility or merely polysemantic feature coding. Liu et al. (2026) demonstrated that standard alignment metrics can conflate representational content with encoding format when models operate under superposition, raising the question of whether $\text{Shesha}_{\text{FS}}$ faces the same conflation. This concern is addressed by construction: $\text{Shesha}_{\text{FS}}$ is formally non-invariant to orthogonal transformations (Table 1; Appendix B). Because superposition redistributes information across the coordinate basis via orthogonal rotation, it fundamentally alters the basis-dependent redundancy of the manifold. $\text{Shesha}_{\text{FS}}$ is specifically designed to detect this lack of redundancy. A representation whose geometry is not redundantly encoded across its coordinate dimensions will produce asymmetric split-half RDMs, regardless of whether that non-redundancy arises from eigenspectral collapse, polysemantic encoding, or any other mechanism. Low $\text{Shesha}_{\text{FS}}$ therefore does not distinguish between these underlying causes, nor does it need to: in all cases, the representation's geometric structure is not redundantly encoded across its feature dimensions, leaving the manifold geometrically vulnerable to coordinate-level perturbation (e.g., pruning or dropout). The PCA compression proof (Appendix B) formalizes one such route, proving that any transformation concentrating variance into $r \ll d$ dimensions strictly reduces $\text{Shesha}_{\text{FS}}$ while leaving basis-independent metrics like CKA approximately invariant.

Recent work on sparse autoencoder (SAE) stability independently corroborates this concern. Paulo and Belrose (2025) showed that SAEs trained on the same model with different random seeds learn substantially different feature sets, and Leask et al. (2025) argued that SAE latents are not canonical units of analysis. Bhalla et al. (2026) provide a complementary geometric account, showing that multidimensional concepts can admit multiple valid SAE bases, making seed-dependent decompositions expected. Geometric stability offers a quantitative framing for this instability: a representation with low $\text{Shesha}_{\text{FS}}$ encodes its geometry non-redundantly across coordinates, so that no coordinate basis is privileged for recovering the pairwise structure from feature subsets. The non-uniqueness that SAE researchers observe empirically is a direct consequence of the coordinate-basis fragility that $\text{Shesha}_{\text{FS}}$ measures formally.

5.4 Geometric Stability Across Substrates

The extension to protein sequences (Bateman et al., 2022), molecular profiles (Zheng et al., 2017), and neural population recordings (Steinmetz et al., 2019) is not incidental. It establishes that geometric stability—the redundancy of a representation's geometry across its coordinate basis—is a substrate-independent axis of representational structure, measurable wherever a representation matrix can be formed. The geometric perspective now pervades fields beyond neuroscience: in computational biology, protein foundation model embeddings encode geometry that predicts structure and function (Jumper et al., 2021; Lin et al., 2023), and genomic foundation models learn sequence-level representations whose geometric organization reflects regulatory structure (Schiff et al., 2024; Avsec et al., 2026; Brixi et al., 2026); in single-cell genomics, transcriptomic profiles define points in gene expression space whose pairwise distances reflect cell type identity (Luecken and Theis, 2019), developmental trajectory (Trapnell et al., 2014), and perturbation response (Butler et al., 2018); in systems neuroscience, population activity vectors (Pandarinath et al., 2018; Saxena and Cunning-

ham, 2019) encode sensory stimuli (Nogueira et al., 2023; Ding et al., 2023), decisions (Gold and Shadlen, 2007; Mante et al., 2013), motor plans (Churchland et al., 2012; Kaufman et al., 2014), and abstract task variables (Bernardi et al., 2020; Tafazoli et al., 2025). In each domain, the analytical strategy abstracts from specific feature identity to population-level geometry. Independent evidence from Wu et al. (2026) supports this perspective: geometry-preserving metrics recover more meaningful structure in both artificial and neural data than metrics that discard geometric information, suggesting that the coordinate-level properties $\text{Shesha}_{\text{FS}}$ measures are functionally relevant rather than incidental. The universality of this strategy is what makes a blind spot in stability assessment consequential across all of them.

In protein encoders, PCA compression induces the same negative stability-similarity correlation observed in the compression regime across all domains. In neural recordings, the natural encoder regime produces the same negligible correlation observed in language and vision encoders trained without explicit compression. The biological systems do not "know" about gradient descent, but they produce the same geometric signatures because the underlying constraint is physical rather than computational. Any system that must represent high-dimensional structure in a limited-capacity basis faces the same question of whether that structure is redundantly distributed across coordinates (Barlow, 1961).

This substrate-independence has a practical implication for neuroscience. Geometric stability complements existing RSA reliability measures (Nili et al., 2014; Walther et al., 2016) by assessing a different failure mode. A low noise ceiling indicates that the data are too noisy to support reliable RDM estimation. A low $\mathcal{S}$ indicates that the representational geometry itself is fragile, regardless of data quality. The pairwise distance structure fractures under independent feature observations even when individual measurements are reliable. These are distinguishable conditions that call for different interventions, and existing tools conflate them.

5.5 Limitations

Shesha is a global metric: it characterizes the full representational geometry of a given layer or region as a single scalar and does not resolve localized instabilities within subsets of the representation. A representation could show high aggregate $\mathcal{S}$ while specific submanifolds corresponding to rare categories or low-frequency stimuli are geometrically fragile. Token-level and region-level variants are natural extensions but are not developed here.

Feature extraction for the vision benchmark uses a single seed per dataset, except for CIFAR-10, which we recomputed across three seeds for all 170 models with near-identical rankings (Spearman $\rho \geq 0.993$, median per-model CV 0.75%; Appendix G.5). The remaining five datasets should be read as point estimates. Cross-domain stability estimates average over 15 seeds and are not subject to this caveat.

The encoder transformation framework provides controlled evidence for the three-regime analysis but does not exhaust the space of transformations a deployed model encounters in practice. A more complete characterization of the stability-similarity relationship under realistic distribution shifts (Kumar et al., 2022) and post-training interventions (Aghajanyan et al., 2020; Li et al., 2025) remains to be done.

The biological domain results establish that geometric stability is measurable and interpretable in protein, molecular, and neural representations, but sample sizes and domain coverage differ substantially from the machine learning analysis. The neuroscience result in particular ($N = 846$ configurations from 26 sessions) reflects a specific recording paradigm and may not generalize across modalities or behavioral contexts. Geometric stability extends naturally to single-cell perturbation screens and to neural population recordings under behavioral tasks, which we develop separately.

Our attribution of low stability to the self-distillation objective is associational: it compares DINOv2 with contrastive families that differ in more than their objective. The causal evidence we have is narrower and concerns the lever rather than the objective. Gram anchoring, which directly targets coordinate-basis redundancy, raises stability at matched scale (DINOv3), and the optimizer and supervision ablations (Section 4.6) move stability as predicted under controlled changes. A controlled objective swap on a fixed backbone would isolate the objective's contribution and remains future work.

5.6 Outlook

Geometric stability should become a standard reporting metric for learned representations alongside accuracy, transferability, and robustness. The `shesha-geometry` PyPI package (Raju, 2026c) provides a single-function interface that requires only a representation matrix as input and returns $\mathcal{S}$ with bootstrap confidence intervals, imposing no requirement for labels, repeated measurements, or a reference representation. The metric is applicable wherever a representation matrix can be extracted: pretrained encoders, fine-tuned models, biological population vectors (Edelman, 1998; Kriegeskorte and Kievit, 2013; Sorscher et al., 2022), or any high-dimensional embedding of structured data.

The DINOv3 result makes a concrete prediction: training objectives that regularize coordinate-basis redundancy, as Gram anchoring does implicitly, will produce more geometrically stable models at little cost to transferability. DINOv3 already shifts the dissociation rather than merely navigating it; whether a regularizer targeting basis redundancy directly can remove the residual transfer cost on the datasets where DINOv3 still pays one is the central open question that follows from this work.

Appendix A. Shesha Variants

The main text presents Feature-Split Shesha ($\text{Shesha}_{\text{FS}}$), the primary 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 $\text{Shesha}_{\text{FS}}$.

A.1 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 geomet-

ric structure is redundantly distributed across the feature basis and requires no labels or repeated measurements.

A.2 Sample-Split Shesha (Shesha$_{SS}$)

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).

Appendix B. Invariance Proofs and Counterexample

We prove the invariance properties listed in Table 1 of the main text. Throughout, $X \in \mathbb{R}^{n \times d}$ is a representation matrix, $\pi_k = (A_k, B_k)$ denotes a random feature partition, and $D^{(k,s)}$ the cosine-distance RDM on half s.

B.1 Global Scaling Invariance

Proof Let $Y = \alpha X$ for $\alpha > 0$. The cosine distance between rows i and j of Y is

$$1 - \frac{(\alpha x_i^{(s)}) \cdot (\alpha x_j^{(s)})}{\|\alpha x_i^{(s)}\| \, \|\alpha x_j^{(s)}\|} = 1 - \frac{x_i^{(s)} \cdot x_j^{(s)}}{\|x_i^{(s)}\| \, \|x_j^{(s)}\|},$$

so $D^{(k,s)}(Y) = D^{(k,s)}(X)$ for every partition and both halves. Hence $\mathcal{S}(Y) = \mathcal{S}(X)$. ■

B.2 Isotropic Scaling Invariance

Proof Follows identically from global scaling, since isotropic scaling $X \mapsto \alpha X$ does not change cosine distances. ■

B.3 Feature Permutation Invariance

Proof Let $Y = XP$ for a permutation matrix $P \in \{0, 1\}^{d \times d}$. The partition π_k is drawn uniformly at random from all $\binom{d}{\lfloor d/2 \rfloor}$ ways to assign coordinate indices to two halves. Because P merely relabels coordinate indices, the distribution of partitions over the relabeled indices is identical to the distribution over the original indices. Formally, for any realization $\pi_k = (A_k, B_k)$ of the original partition, the partition $(P^{-1}A_k, P^{-1}B_k)$ is an equally probable realization of the permuted partition, and the corresponding RDMs satisfy $D^{(k,s)}(XP) = D^{(k,P^{-1}s)}(X)$. Averaging over K independent draws therefore gives

$\mathcal{S}(XP) = \mathcal{S}(X)$. ■

B.4 Monotonic Distance Invariance

Proof Spearman rank correlation ρ_s depends only on the relative ordering of pairwise distances, not their values. Let g be strictly monotone increasing. For any two pairs (i, j) and (k, l),

$$D_{ij} < D_{kl} \iff g(D_{ij}) < g(D_{kl}),$$

so the rank vectors of $\mathrm{vec}(D^{(k,A_k)})$ and $\mathrm{vec}(D^{(k,B_k)})$ are unchanged under g, and ρ_s is invariant. ■

B.5 Non-Invariance to PCA Compression

Proof Let $X \in \mathbb{R}^{n\times d}$ have geometric information distributed across all d coordinates, so that $\mathcal{S}(X) \approx 1$. Let Y be the rank-r PCA approximation of X with $r \ll d$. After compression, only r columns of Y carry nonzero variance; the remaining $d-r$ columns are identically zero. For any random equipartition (A_k, B_k) that places all r informative columns in the same half, the other half-RDM is degenerate (all pairwise cosine distances undefined), yielding $\rho_s\big(\mathrm{vec}(D^{(k,A_k)}), \mathrm{vec}(D^{(k,B_k)})\big) = 0$. Such splits occur with positive probability when $r \le \lfloor d/2 \rfloor$, so $\mathcal{S}(Y) < \mathcal{S}(X)$. Meanwhile, $YY^\top$ retains the dominant eigenvalues of $XX^\top$, so $\mathrm{CKA}(X, Y) \approx 1$ for spectra concentrated in the top components. ■

B.6 Non-Invariance to Orthogonal Transformations: Constructive Counterexample

Proof We exhibit X and $Q \in \mathcal{O}(d)$ such that $\mathcal{S}(XQ) \neq \mathcal{S}(X)$, while $\mathrm{CKA}(X, XQ) = 1$.

Let $d = 4$ and

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

Let Q be the orthogonal matrix

$$Q = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix},$$

whose rows are orthonormal, so $QQ^\top = I$. Setting $Y = XQ$ gives

$$Y = \begin{pmatrix} 0 & 0 & \sqrt{2} & \sqrt{2} \\ 0 & 0 & \sqrt{2} & -\sqrt{2} \\ 0 & 0 & -\sqrt{2} & \sqrt{2} \end{pmatrix}.$$

Because Q is orthogonal, $YY^\top = XQQ^\top X^\top = XX^\top$, so linear $\mathrm{CKA}(X, Y) = 1$ exactly.

We average the split-half correlation over the three equipartitions of the four coordinates into halves of size two. For X, columns 3 and 4 duplicate columns 1 and 2, so the splits $\{1,2\} \mid \{3,4\}$ and $\{1,4\} \mid \{2,3\}$ build each half from the same pair of column vectors and give $\rho_s = 1$, while the split $\{1,3\} \mid \{2,4\}$ separates the duplicates and gives $\rho_s = -\frac{1}{2}$. Hence $\mathcal{S}(X) = \frac{1}{3}\big(1 + 1 - \frac{1}{2}\big) = \frac{1}{2}$.

The rotation moves all row variance of Y into columns 3 and 4, leaving columns 1 and 2 identically zero. The split $\{1,2\} \mid \{3,4\}$ now has a degenerate half on $\{1,2\}$: every row restricted to those coordinates is the zero vector, its pairwise cosine distances are undefined, and we follow the convention $\rho_s = 0$ for a degenerate half. The other two splits give $\rho_s = -\frac{1}{2}$ each, so $\mathcal{S}(Y) = \frac{1}{3}\big(0 - \frac{1}{2} - \frac{1}{2}\big) = -\frac{1}{3}$.

Therefore $\mathcal{S}(Y) = -\frac{1}{3} < \frac{1}{2} = \mathcal{S}(X)$ while $\mathrm{CKA}(X, Y) = 1$, so $\mathcal{S}$ is not invariant under orthogonal transformations. ■

The counterexample generalizes: any Q that concentrates the column energy of X into a strict subset of coordinates will reduce $\mathcal{S}$ while leaving $XX^\top$ unchanged. The degree of reduction depends on how severely Q breaks the distributional uniformity of geometric information across coordinate axes.

B.6.1 Numerical Confirmation at Fixed Spectrum

The constructive example uses $\mathcal{S}(X) = \frac{1}{2}$ for hand-verifiability; for genuinely stable representations the dissociation is far sharper. We generated $X \in \mathbb{R}^{200\times 64}$ by projecting a five-dimensional latent ($Z \in \mathbb{R}^{200\times 5}$, standard normal) through a dense random map ($W \in \mathbb{R}^{5\times 64}$) with small additive noise (seed 320), so that every coordinate carries the full latent geometry and the representation is highly recoverable from random feature halves ($\mathcal{S}(X) = 0.903$). Rotating X into its own eigenbasis, $Y = XQ$ with Q the matrix of right singular vectors of X, is an orthogonal transformation that leaves $XX^\top$, and hence the entire singular spectrum, the rank, and linear CKA, unchanged ($\mathrm{CKA}(X, Y) = 1.000$). Yet $\mathcal{S}(Y) = -0.008$: the rotation concentrates the variance onto the leading coordinates (coordinate j of Y has norm s_j), so balanced feature splits that isolate the trailing near-zero coordinates yield degenerate half-RDMs and the split-half agreement collapses. Because the rotation changes nothing about the eigenvalues, the collapse is attributable to the coordinate basis alone, confirming that $\mathrm{Shesha}_{\mathrm{FS}}$ measures basis-dependent redundancy rather than a property of the spectrum.

Appendix C. 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:

1. *Axis of splitting*
   - Noise ceiling: observations (trials, subjects)
   - Shesha: features (neurons, embedding dimensions)
2. *Diagnostic target*
   - Noise ceiling: data quality (is the measurement reliable?)
   - Shesha: representational architecture (is the geometry redundantly encoded?)
3. *Requirements*
   - Noise ceiling: requires repeated measurements of the same conditions
   - Shesha: requires only a single matrix $X \in \mathbb{R}^{n \times d}$, enabling assessment of pre-trained embeddings, single-cell profiles, and other systems where observation-level replication is unavailable.
4. *Partition scheme*
   - Noise ceiling: leave-one-out across subjects (number of partitions fixed by sample size)
   - Shesha: K independent random equipartitions of the feature index set, averaged to reduce partition noise

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 established for the noise ceiling apply directly to Shesha.

Appendix D. 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 using Eq. 1. Spearman rank correlation between the vectorized upper triangles of the two RDMs was then computed. This procedure was repeated for K=30 independent random partitions and the results 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 across 15 models on CIFAR-10 and CIFAR-100 confirmed that estimates at $n = 400$ deviate from those at $n = 1{,}600$ by a mean absolute difference of 0.0077 (Sec. E.1), supporting the use of $n_{\max} = 1{,}600$ as a conservative ceiling.

All computations used fixed random seed 320 for reproducibility, unless otherwise noted. 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 of approximately 0.4 for independent random matrices present in standard linear CKA (Kornblith et al., 2019a).

Appendix E. Ground Truth Validation

This section validates $\text{Shesha}_{\text{FS}}$ on synthetic and controlled data where the ground-truth answer is known. The governing question is construct validity: does the metric measure geometric stability, and is that measurement distinct from representational similarity? A valid measure must satisfy two requirements. It must respond to genuine variation in stability (sensitivity), and it must not reduce to a re-description of similarity (discriminant validity).

We establish sensitivity with a signal-to-noise sweep over representations of known stability (Section E.3, Fig. S2), where $\text{Shesha}_{\text{FS}}$ recovers the ground-truth ordering almost exactly ($\rho = 0.997$). We establish discriminant validity with a balanced four-quadrant design (Section E.9, Fig. S5) that decouples stability from similarity by construction: across the balanced sample the Spearman correlation between $\text{Shesha}_{\text{FS}}$ and debiased CKA falls to $\rho = 0.204$, showing that the two indices track largely independent properties of a representation. The encoder-transformation analysis of the main text (Section 3.1) supplies a complementary sanity check in the sense of Kornblith et al. (2019a) and Ding et al. (2021): geometry-preserving operations (random projection, isotropic noise) move $\text{Shesha}_{\text{FS}}$ and CKA in parallel, whereas geometry-altering operations (aggressive PCA, feature selection) dissociate them. This pattern follows from the basis-dependence of $\text{Shesha}_{\text{FS}}$ (Appendix B): operations that redistribute variance across the coordinate basis change how recoverably the distance geometry can be reconstructed from feature subsets, whereas operations that preserve the pairwise distance geometry keep the two indices in agreement.

The remaining subsections confirm that these measurements are numerically reliable, converging at modest sample sizes (Section E.1) and reproducing across independent feature splits, and that $\text{Shesha}_{\text{FS}}$ responds to spectral structure exactly as the basis-dependence account predicts.

E.1 Convergence Over K and Subsampling

To assess whether Shesha estimates converge reliably as sample size varies, we measured $\mathcal{S}$ at two sample sizes ($n \in \{400, 1600\}$) across 15 models on both CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). For each model-dataset combination, we randomly sampled n examples without replacement using a fixed random generator and measured the drift $\Delta = \mathcal{S}_{n=400} - \mathcal{S}_{n=1600}$. Stability was defined as $|\Delta| < 0.05$.

Shesha estimates demonstrated excellent convergence across all architectures (Fig. S1). The mean absolute drift across all 30 model-dataset combinations was $|\bar{\Delta}| = 0.0115$, well below the stability threshold. When averaged per model across both datasets, drifts ranged from 0.0002 (ResNet-50, most stable) to 0.0176 (ViT-Tiny, least stable), with mean 0.0077. All 15 models achieved stable estimates at $n = 400$, confirming that reliable measurements

can be obtained at modest sample sizes. We use $n_{\max} = 1{,}600$ throughout as a conservative ceiling and $K = 30$ splits as the estimation protocol.

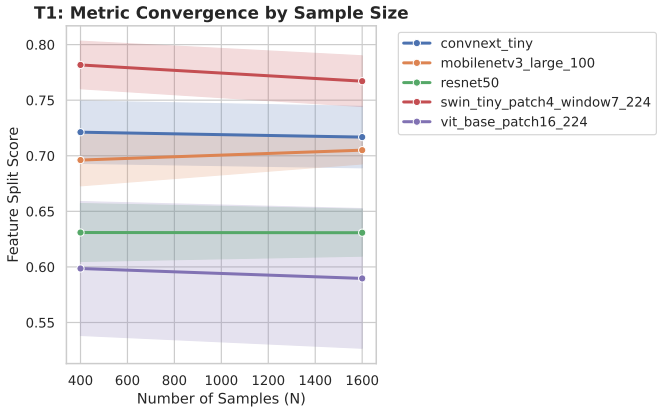

Figure S1: Metric convergence: Shesha estimates remain stable as sample size increases from 400 to 1600 across representative architectures. The flat trajectories confirm rapid convergence and numerical reliability at modest sample sizes.

E.2 Dimensionality Sensitivity

To assess how $\text{Shesha}_{\text{FS}}$ behaves under dimensionality reduction, which is common in visualization and computational-efficiency contexts, we applied Principal Component Analysis (PCA) to reduce embeddings from their native dimensionality (512–2048, depending on the architecture) to 64 dimensions. For each of the 30 model-dataset combinations, we extracted 400 samples, fit PCA with `n_components=64` and `random_state=320`, transformed the embeddings, and recomputed $\text{Shesha}_{\text{FS}}$ on the reduced representations.

Reduction to 64 components drove $\text{Shesha}_{\text{FS}}$ from a full-dimensional mean of $+0.620$ to a negative mean of -0.112 (range $[-0.204, -0.055]$ across the 30 conditions). This is the empirical signature of the compression regime characterized in Appendix B: projecting onto the leading principal components concentrates variance into a low-dimensional coordinate subset, so the pairwise distance structure is no longer redundantly recoverable from arbitrary coordinate halves. A random split then divides a non-redundant code, the two half-RDMs carry overlapping rather than complementary structure, and their rank correlation falls to zero or slightly below. This confirms empirically what the PCA compression proof (Appendix B) establishes formally: any projection that concentrates variance into $r \ll d$ coordinates strictly reduces $\text{Shesha}_{\text{FS}}$, while basis-independent metrics such as CKA remain approximately unchanged. $\text{Shesha}_{\text{FS}}$ measurements should therefore be computed on full-dimensional embeddings.

E.3 Sensitivity to Known Stability Levels

We generated representations with parametrically controlled stability by mixing a low-rank signal component with isotropic noise:

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

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. We tested 21 levels from $\alpha = 0$ to $\alpha = 1$ in increments of 0.05, using seeds $\mathcal{S}[i \mod 15] \times 100 + i$ for each level $i \in \{0, \ldots, 20\}$. Shesha showed near-perfect rank correlation with ground truth stability ($\rho = 0.997$), confirming it accurately measures internal representational consistency.

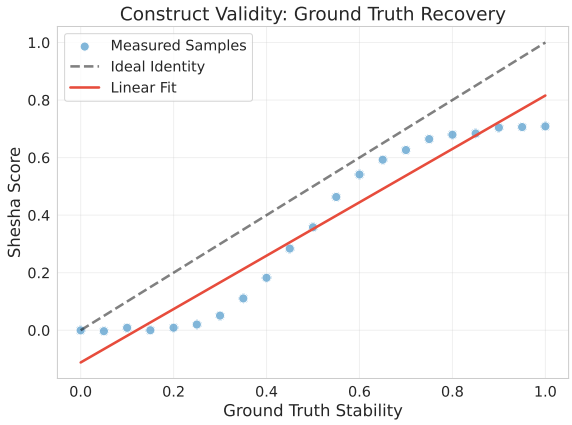

Figure S2: Construct validity, ground-truth recovery: $\text{Shesha}_{\text{FS}}$ scores plotted against parametrically controlled stability levels (signal-to-noise ratio α) in synthetic representations. The metric shows a near-perfect monotonic response ($\rho = 0.997$) to the underlying ground truth, confirming high sensitivity to geometric consistency.

E.4 Spectral Deletion

The spectral interpretation of Sec. 2.2 predicts that CKA should collapse after removing leading principal components while $\mathcal{S}$ retains sensitivity across the eigenspectrum. We tested this directly by constructing representations with a power-law eigenspectrum ($S_{ii} = 100/(i + 1)$) and progressively removing the top k principal components. All similarity metrics collapse to near-zero after removing just 1–2 dominant components (CKA, PWCKA, and Procrustes all fall below 0.5 at $k = 1$), while Shesha retains meaningful signal until $k = 20$ and remains above zero at $k = 50$ (Table S2). The divergence is robust across preprocessing conditions: in this controlled setting CKA collapses because it is dominated by the leading components, while $\text{Shesha}_{\text{FS}}$ responds across the eigenspectrum.

E.5 Tail Noise Ablation

The spectral deletion experiment (Table S2) demonstrates that $\text{Shesha}_{\text{FS}}$ retains sensitivity when signal-carrying principal components are removed. A natural complementary question is whether this sensitivity produces false alarms: does $\text{Shesha}_{\text{FS}}$ react to pure noise injected into the spectral tail, where no functional representational content resides?

E.5.1 Protocol

We generated synthetic representations $X \in \mathbb{R}^{2000\times512}$ with a power-law eigenspectrum ($\lambda_i \propto i^{-1.5}$, matching the spectral profile of trained deep network penultimate layers; seed 320). We decomposed X via SVD and identified the top $k = 34$ principal components explaining 90% of total variance as the signal subspace. We then injected isotropic Gaussian noise exclusively into the remaining 478 tail components at 14 scale levels ($\sigma \in [0.001, 50]$), reconstructed the perturbed representation $\tilde{X}$ in the original coordinate space, and measured three diagnostics: linear CKA between X and $\tilde{X}$, Shesha RDM similarity between X and $\tilde{X}$ (Spearman correlation of pairwise distance vectors), and internal $\text{Shesha}_{\text{FS}}$ of $\tilde{X}$. Each condition was repeated with 3 independent noise draws.

E.5.2 Results

$\text{Shesha}_{\text{FS}}$ does not false-alarm on non-functional tail noise. At low noise scales ($\sigma \leq 0.01$), both CKA and Shesha RDM similarity remain above 0.999, and internal $\text{Shesha}_{\text{FS}}$ holds at its baseline value of 0.971. As noise increases, the two cross-comparison metrics degrade at similar rates: CKA drops below 0.95 at $\sigma = 0.20$, while Shesha RDM similarity crosses the same threshold slightly earlier at $\sigma = 0.10$. At moderate noise ($\sigma = 0.05$), both metrics remain above 0.99 and 0.99 respectively. At high noise scales ($\sigma \geq 1.0$), where injected tail energy overwhelms the original tail variance, both metrics collapse to near zero.

Critically, the degradation profiles of CKA and Shesha track in parallel across the full noise range. There is no regime in which Shesha detects a change that CKA does not, confirming that Shesha's sensitivity to spectral tail structure (Table S2) is specific to genuine structural changes (removal of signal-carrying components) rather than energetic perturbations of non-functional dimensions.

E.5.3 Interpretation

This result resolves the apparent tension between two findings: the spectral deletion experiment shows $\text{Shesha}_{\text{FS}}$ detects when tail structure is removed (a genuine geometric change that alters pairwise distance rankings), while this ablation shows it does not react when tail noise is merely amplified (a perturbation that preserves pairwise distance rankings because the signal subspace dominates). The Spearman rank-order correlation that underlies $\text{Shesha}_{\text{FS}}$ is the mechanism: monotone transformations of pairwise distances, including additive noise that does not alter rank order, leave $\mathcal{S}$ unchanged (Table 1, monotonic distance invariance). Only when noise is large enough to scramble the rank ordering of pairwise distances does $\text{Shesha}_{\text{FS}}$ degrade, and at that point CKA degrades equally.

E.6 Preprocessing Ablation

Following (Walther et al., 2016), we tested robustness across preprocessing conditions: raw, centered, centered with L2 normalization, and whitened (ZCA with shrinkage $\lambda = 0.1$). The Shesha-CKA divergence persists across raw, centered, and normalized conditions (Table S1).

Table S1: Shesha and CKA values at $k = 30$ PCs removed under different preprocessing. The divergence is robust except under whitening, which equalizes the spectrum.

Preprocessing	Shesha	Debiased CKA	Difference
Raw	0.276	−0.076	0.352
Centered	0.417	−0.076	0.493
Centered + Normalized	0.417	−0.083	0.500
Whitened	0.316	−0.054	0.370

E.6.1 Mechanistic Interpretation of Whitening

Under whitening, CKA remains negative at $k = 30$ (-0.054), though less so than under raw preprocessing (-0.076). The whitened Shesha baseline drops from 0.98 to 0.50 at $k = 0$, reflecting noise amplification from spectral equalization.

E.7 Comparison with RSA Reliability Methods

We additionally compared Shesha to whitened RDM stability (Walther et al., 2016; Diedrichsen and Kriegeskorte, 2017) and noise ceiling estimation procedures (Nili et al., 2014). Standard Shesha correlates almost perfectly with whitened Shesha ($\rho = 1.000$, $p < 10^{-70}$), confirming methodological consistency with established RSA reliability practices. The key distinction is that Shesha operates on raw representations without requiring whitening, avoiding the numerical instability and noise amplification associated with ZCA on high-dimensional neural activations.

These results demonstrate that Shesha captures geometric structure distributed across the eigenspectrum, whereas similarity metrics are dominated by the top principal components. The divergence is robust across preprocessing choices. Its underlying cause is basis-dependence rather than the spectrum alone: an orthogonal rotation that concentrates variance into a coordinate subset lowers $\text{Shesha}_{\text{FS}}$ while leaving CKA unchanged (Appendix B).

Table S2: Metric values after removing top k principal components. All similarity metrics collapse immediately while Shesha degrades gracefully, retaining sensitivity to spectral tail structure. k = PCs Removed. [a]Shesha at $k = 0$ reflects split-half reliability rather than trivial self-similarity.

k	Shesha	CKA	Debiased CKA	PWCKA	Procrustes
0	0.981[a]	1.000	1.000	1.000	1.000
1	0.955	0.273	0.262	0.274	0.389
2	0.932	0.136	0.118	0.136	0.238
3	0.905	0.083	0.060	0.083	0.170
4	0.876	0.057	0.031	0.057	0.132
5	0.850	0.043	0.012	0.043	0.108
6	0.823	0.033	0.000	0.033	0.091
7	0.802	0.027	−0.009	0.027	0.078
8	0.774	0.022	−0.016	0.022	0.069
9	0.744	0.019	−0.022	0.019	0.061
10	0.717	0.016	−0.027	0.016	0.055
⋮					
15	0.607	0.009	−0.045	0.009	0.036
20	0.495	0.006	−0.058	0.006	0.027
25	0.392	0.004	−0.067	0.004	0.021
30	0.309	0.003	−0.075	0.002	0.017
35	0.238	0.003	−0.082	0.002	0.014
40	0.171	0.002	−0.087	0.001	0.012
45	0.124	0.002	−0.092	0.000	0.010
50	0.086	0.001	−0.097	0.000	0.009

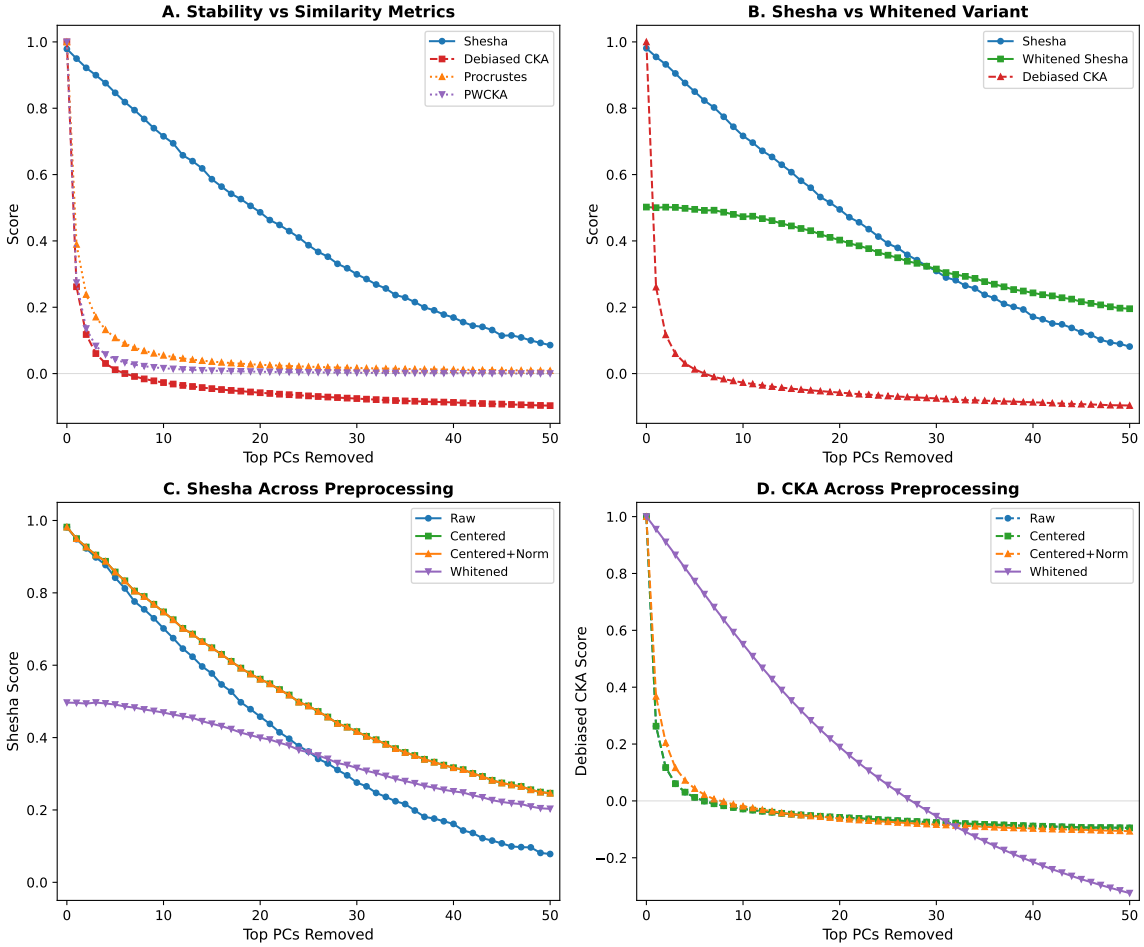

Figure S3: Spectral sensitivity analysis: Metric responses as the top k principal components are progressively removed from a power-law representation. (A) Shesha degrades gracefully while all similarity metrics (CKA, PWCKA, Procrustes) collapse after removing just 1 PC. (B) Comparison with whitened 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; whitening causes CKA to recover sensitivity by equalizing the spectrum.

E.8 Seed Stability

To verify that the stochastic feature-splitting procedure produces consistent estimates across random initializations, we computed $\mathcal{S}$ twice for each of 15 models on both CIFAR-10 and CIFAR-100, using `seed=100` and `seed=200` respectively. Each seed generates a different sequence of $K = 30$ random feature partitions. Sensitivity was measured as $|\mathcal{S}_{\text{seed=100}} - \mathcal{S}_{\text{seed=200}}|$.

The metric demonstrated excellent seed stability across all architectures and datasets (Fig. S4). The mean sensitivity across all 30 model-dataset combinations was 0.0047, with a maximum of 0.0142 (ResNet-34 on CIFAR-100) and a minimum of 0.00015 (ResNet-50 on CIFAR-10). All 30 combinations fell well below the 0.05 stability threshold, with 25/30 below 0.01. These results confirm that averaging over $K = 30$ random splits provides sufficient variance reduction to yield highly reproducible estimates, with typical seed-to-seed variation below 1% of the score magnitude.

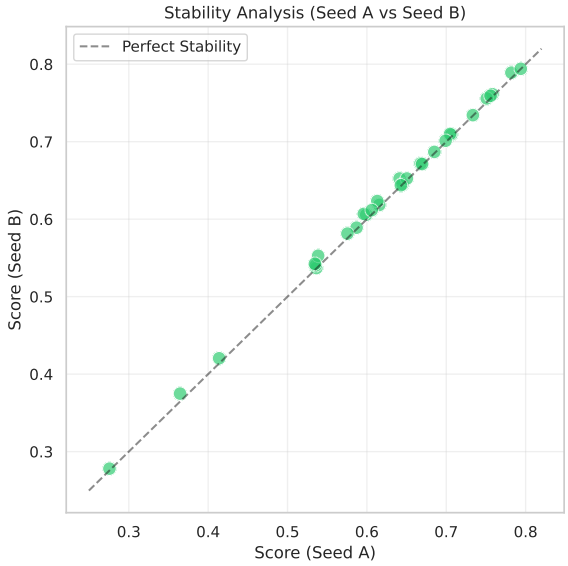

Figure S4: Seed stability: Shesha scores computed with `seed=100` vs. `seed=200` across 15 architectures on CIFAR-10 and CIFAR-100. Points align closely with the diagonal, confirming high reproducibility across random initializations. Mean sensitivity = 0.0047; maximum = 0.0142.

E.9 Dissociation with Balanced Quadrant Sampling

Naïve random sampling of stability levels induces spurious correlation between Shesha and CKA because high-stability representations (strong signal) tend to show low between-representation similarity (independent signals), while low-stability representations (noise) show elevated CKA due to finite-sample effects. To break this coupling, we explicitly sampled from four quadrants (15 pairs each, using seeds derived from $\mathcal{S}$):

1. **High stability, high similarity** (Q1): Representations derived from the same latent structure ($\alpha = 0.9$) with small additive noise ($\sigma = 0.1$). Seeds: $\mathcal{S}[i] \times 1000 + 1$ for $i \in \{1, \ldots, 15\}$. Results: Shesha $= 0.701 \pm 0.003$, CKA $= 0.998 \pm 0.000$.

2. **High stability, low similarity** (Q2): Independent high-signal representations ($\alpha = 0.9$) with different latent draws. Seeds: $\mathcal{S}[i] \times 1000 + 2$ and $\mathcal{S}[i] \times 1000 + 3$ for each pair. Results: Shesha $= 0.701 \pm 0.004$, CKA $= 0.001 \pm 0.010$.

3. **Low stability, low similarity** (Q3): Independent noise representations ($\alpha = 0.1$). Seeds: $\mathcal{S}[i] \times 1000 + 4$ and $\mathcal{S}[i] \times 1000 + 5$ for each pair. Results: Shesha $= 0.001 \pm 0.003$, CKA $= -0.001 \pm 0.010$.

4. **Low stability, high similarity** (Q4): Adversarial quadrant constructed via rejection sampling. We generated pairs where $X \sim \mathcal{N}(0, I)^{200 \times 256}$ and $Y = X + \mathcal{N}(0, 0.15^2 I)$, accepting only samples where Shesha < 0.4 and CKA > 0.4. This creates representations with aligned sample geometry (high CKA) but inconsistent feature-split structure (low Shesha). Acceptance rate: 100% (15/15). Results: Shesha $= -0.001 \pm 0.005$, CKA $= 0.978 \pm 0.000$.

The Spearman correlation of $\rho = 0.204$ between Shesha and debiased CKA using equal numbers of samples from each of the four quadrants shows that these two metrics assess largely different attributes of the data, as shown in Figure S5.

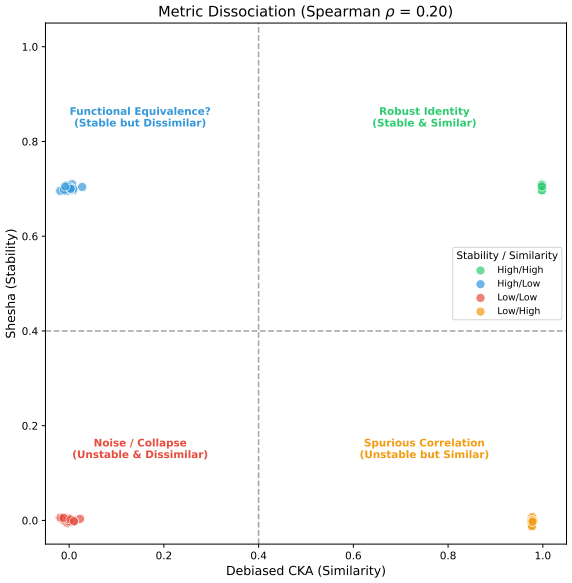

Figure S5: Four-quadrant dissociation: Shesha vs. debiased CKA for 60 representation pairs sampled equally from four quadrants of the stability × similarity space. Q1 (high/high): Shesha $= 0.701 \pm 0.003$, CKA $= 0.998 \pm 0.000$. Q2 (high/low): Shesha $= 0.701 \pm 0.004$, CKA $= 0.001 \pm 0.010$. Q3 (low/low): Shesha $= 0.001 \pm 0.003$, CKA $= -0.001 \pm 0.010$. Q4 (low/high, adversarial): Shesha $= -0.001 \pm 0.005$, CKA $= 0.978 \pm 0.000$. Balanced Spearman $\rho = 0.20$.

Appendix F. Encoders

The distinctness of stability and similarity (Section 3.1) is established on a large, heterogeneous corpus of representations rather than on any single model or dataset. This appendix specifies that corpus: 2,463 configurations spanning seven data domains, with full data sources, encoders, and preprocessing.

F.1 Cross-Domain Validation: Data Sources and Preprocessing

To rule out modality-specific artifacts, the corpus spans seven domains ranging from natural language to neural population recordings. Within each domain we fix a stimulus set and enumerate configurations by varying the encoder or encoding scheme and its preprocessing; the resulting counts N and full protocols are given below.

F.1.1 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`.

F.1.2 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-patch16-224`, and ResNet50 (ImageNet-V2 weights).

F.1.3 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`.

F.1.4 Video (N=128)

Action clips were drawn from UCF-101 (Soomro et al., 2012): 100 videos were sampled uniformly at random (seed-controlled) from the on-disk corpus, with 16 frames per clip selected by uniform temporal indexing and resized to 224×224 with ImageNet normalization. Base models comprised two temporal transformers (`facebook/timesformer-base-finetuned-k400` (8 frames), `MCG-NJU/videomae-base` (16 frames)) and two frame-level encoders: ViT-B/16 (`google/vit-base-patch16-224`) applied to the temporal mean frame, and CLIP ViT-B/32 (`openai/clip-vit-base-patch32`) with embeddings from four uniformly spaced frames averaged per clip.

A preliminary analysis using 100 clips uniformly sampled from a single-source Jellyfish video (Allyn, 2016) with the same base models and preprocessing yielded nearly identical results ($\rho = -0.24$ vs. $\rho = -0.27$), suggesting the stability–similarity relationship is robust to video source diversity.

F.1.5 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).

F.1.6 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).

F.1.7 Neural Population Recordings (N=846)

Neuropixels recordings from Steinmetz et al. (2019), 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.

F.2 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 $(3, 7, 9, 11, 12, 18, 103, 108, 320, 411, 724, 1754, 1991, 2222, 7258)$.

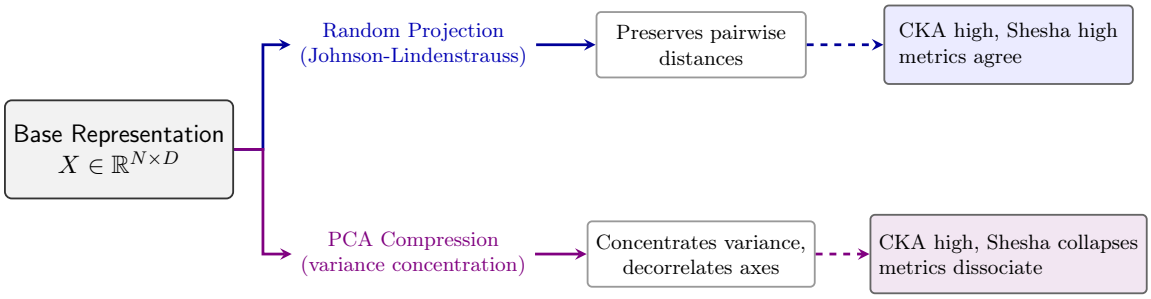

Figure S6: Two coupling regimes. Distance-preserving transforms such as random projection keep both metrics high, so CKA and Shesha agree. Variance concentration through PCA compression preserves dominant variance, keeping CKA high, while decorrelating the retained axes, which collapses Shesha, so the metrics dissociate. Only the second regime produces negative coupling, which is why the pooled correlation is near zero while the regime-level correlations are not.

F.2.1 PCA

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

F.2.2 Random Projection

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

F.2.3 Top-Variance Feature Selection

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

F.2.4 Random Feature Subsets

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

F.2.5 Gaussian Noise Injection

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

F.2.6 Normalization

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

F.3 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) and are reported in Table S3.

F.3.1 Effective-Rank Projection-Weighted CKA (PWCKA)

This variant projects both representations to a shared dimensionality determined by the effective rank before computing CKA. Given the centered representations $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$, we compute their singular value decompositions:

$$\mathbf{X} = \mathbf{U}_X \mathbf{S}_X \mathbf{V}_X^\top, \quad \mathbf{Y} = \mathbf{U}_Y \mathbf{S}_Y \mathbf{V}_Y^\top$$

The effective rank k is the minimum number of components explaining 99% of variance in either representation:

$$k = \min\left(k_X^{(0.99)}, k_Y^{(0.99)}\right), \quad \text{where } k_Z^{(\tau)} = \min\left\{ j : \frac{\sum_{i=1}^{j} s_z^{(i)2}}{\sum_i s_z^{(i)2}} \geq \tau \right\}$$

CKA is then computed on the truncated projections:

$$\mathbf{X}' = \mathbf{U}_X^{(1:k)} \mathbf{S}_X^{(1:k)}, \quad \mathbf{Y}' = \mathbf{U}_Y^{(1:k)} \mathbf{S}_Y^{(1:k)}$$

$$\mathrm{PWCKA}(\mathbf{X}, \mathbf{Y}) = \mathrm{CKA}(\mathbf{X}', \mathbf{Y}')$$

F.3.2 Procrustes Similarity

Procrustes analysis finds the optimal orthogonal transformation that aligns two representations. Given centered representations $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$, we first normalize them to a unit Frobenius norm:

$$\tilde{\mathbf{X}} = \frac{\mathbf{X}}{\|\mathbf{X}\|_F}, \quad \tilde{\mathbf{Y}} = \frac{\mathbf{Y}}{\|\mathbf{Y}\|_F}$$

The optimal orthogonal matrix $\mathbf{R}^* = \underset{\mathbf{R}^\top \mathbf{R} = \mathbf{I}}{\arg\min} \|\tilde{\mathbf{X}} - \tilde{\mathbf{Y}}\mathbf{R}\|_F^2$ is obtained via the SVD of the cross-covariance matrix:

$$\tilde{\mathbf{Y}}^\top \tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \implies \mathbf{R}^* = \mathbf{U}\mathbf{V}^\top$$

Procrustes similarity is defined as follows:

$$\text{Procrustes}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\|\tilde{\mathbf{X}} - \tilde{\mathbf{Y}}\mathbf{R}^*\|_F^2}{\|\tilde{\mathbf{X}}\|_F^2 + \|\tilde{\mathbf{Y}}\mathbf{R}^*\|_F^2}$$

Table S3: 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	+0.03	0.74	Yes
PWCKA	−0.22	0.012	Yes
Procrustes	+0.28	0.001	Yes

F.4 Statistical Methods

This subsection specifies the inferential procedures behind the distinctness analysis (Section 3.1) and the vision benchmark. We describe the resampling scheme used for confidence intervals, the mixed-effects control for base-model identity, the group comparisons, and our handling of multiple comparisons.

F.4.1 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.

F.4.2 Mixed-Effects Models

To rule out base model identity as a confound, we fit a mixed-effects model with $\mathcal{S}$ as outcome, debiased CKA as fixed effect, and base model as random intercept. The intraclass correlation coefficient (ICC) for base model was 0.10, indicating that base model identity explains less than 10% of the variance in stability. The fixed-effect slope of CKA on $\mathcal{S}$ was -0.03 (95% CI $[-0.08, +0.02]$), consistent with the aggregate near-zero correlation reported in the main text.

F.4.3 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.

F.4.4 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.

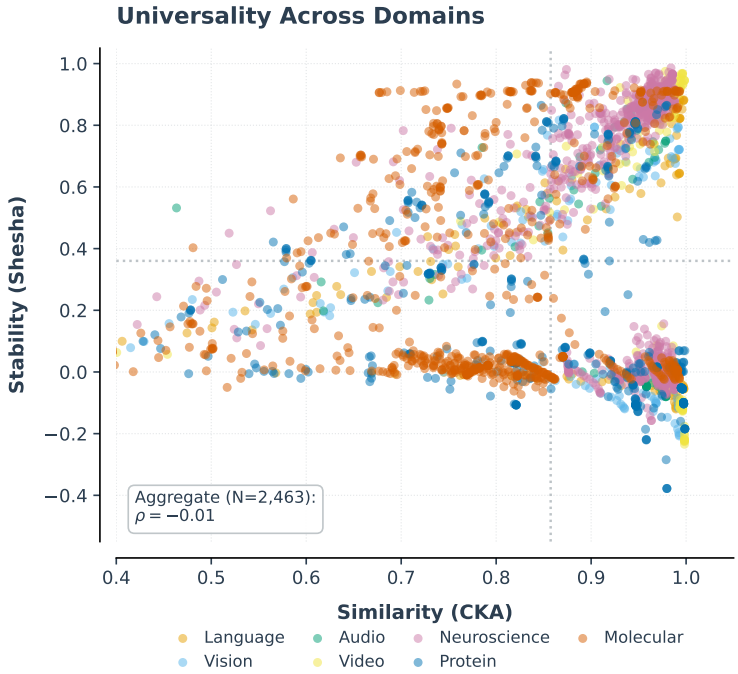

Figure S7: Universality: Across 2,463 encoder configurations spanning seven domains, Shesha and CKA show negligible net correlation ($\rho = -0.01$, 95% CI $[-0.06, +0.03]$), confirming they capture distinct geometric properties.

Table S4: Correlation between Shesha and CKA by encoder type.

Encoder Type	N	ρ [95% CI]
Random Features	201	+0.92 [+0.89, +0.94]
Random Projection	395	+0.89 [+0.86, +0.92]
Noise Injection	395	+0.58 [+0.49, +0.66]
Top Variance	287	+0.62 [+0.54, +0.71]
Normalization	158	+0.34 [+0.17, +0.50]
Original	79	+0.31 [+0.04, +0.55]
PCA	948	−0.47 [−0.52, −0.42]

Table S5: 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, -0.04]
Excluding Protein	2061	$+0.04$ [-0.00, $+0.09$]
Only transformer domains	448	-0.05 [-0.15, $+0.07$]
Only biological domains	2015	$+0.01$ [-0.04, $+0.06$]

Table S6: 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).

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	-0.27	$[-0.47, -0.05]$	0.002
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

Appendix G. Vision Benchmark: Extended Results

This appendix specifies the vision benchmark behind Section 4: how the 170 models were selected, and how their features and transferability scores were computed. The subsections that follow document these choices and the extended analyses that support the main-text claims.

G.1 Model Selection

170 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 36 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").

G.2 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).

G.3 Transferability Metrics

LogME (You et al., 2022, 2021) was computed on the same features using the authors' implementation. LEEP (Nguyen et al., 2020) was computed for models with classification heads.

G.4 Extended Results

Table S7: The DINOv2 paradox at the individual-model level. DINOv2-giant ranks in the bottom quartile for $\text{Shesha}_{\text{FS}}$ on every dataset except EuroSAT, while attaining top-six transferability on CIFAR-10, CIFAR-100, and Flowers-102. On EuroSAT it is instead among the most stable models (4/170).

Dataset	LogME	LogME Rank	$\text{Shesha}_{\text{FS}}$	FS Rank
CIFAR-10	1.384	6/170	0.414	160/170
CIFAR-100	1.629	3/170	0.319	158/170
Flowers-102	3.521	3/170	0.152	168/170
DTD	0.952	30/170	0.502	129/170
EuroSAT	0.681	16/170	0.987	4/170
Oxford Pets	1.760	30/170	0.569	141/170

Table S8: Contrastive vs. self-supervised stability: Mann-Whitney U tests comparing contrastive models (CLIP, SigLIP, ViTamin; n=29) to self-supervised models (n=41) on $\text{Shesha}_{\text{FS}}$. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Contrastive models are more stable on all six datasets.

Dataset	Contrastive (n=29)	SSL (n=41)	Δ	p
CIFAR-10	0.80 ± 0.05	0.67 ± 0.17	+0.13	$<0.001^{***}$
CIFAR-100	0.76 ± 0.08	0.61 ± 0.21	+0.15	$<0.001^{***}$
Flowers-102	0.83 ± 0.03	0.69 ± 0.24	+0.14	0.012^{*}
DTD	0.68 ± 0.08	0.58 ± 0.14	+0.09	$<0.001^{***}$
EuroSAT	0.95 ± 0.01	0.89 ± 0.07	+0.06	$<0.001^{***}$
Oxford Pets	0.86 ± 0.03	0.69 ± 0.12	+0.17	$<0.001^{***}$

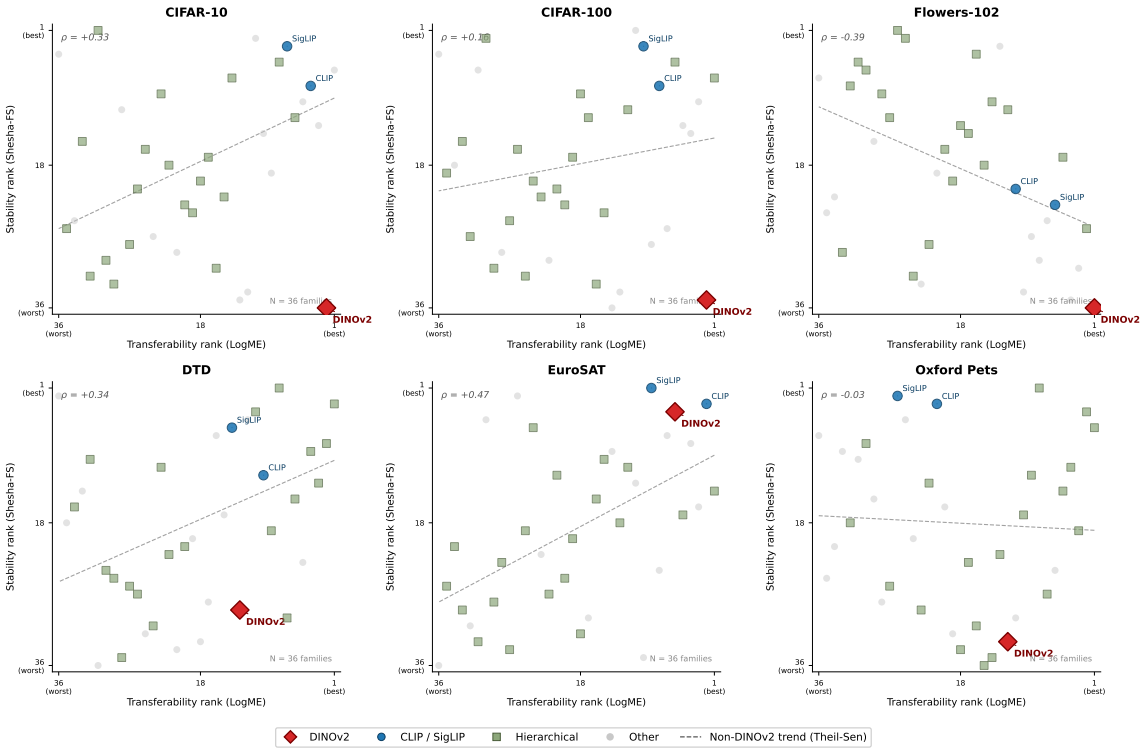

Figure S8: Per-dataset family rankings underlying Figure 3. Each panel plots family-mean transferability rank (LogME) against geometric stability rank ($\text{Shesha}_{\text{FS}}$) for the 36 families, both axes oriented worst to best. DINOv2 occupies the high-transfer, low-stability corner on five datasets, but not on EuroSAT, where it ranks among the most stable families. This is the within-family confirmation of the concentration-stability relationship of Section 4.4: EuroSAT is the dataset on which DINOv2's spectrum is most concentrated. The robust Theil-Sen cross-family trend is positive or null on five datasets (CIFAR-10 +0.33, CIFAR-100 +0.16, DTD +0.34, EuroSAT +0.47, Oxford Pets −0.03) and negative only on Flowers-102 (−0.39), so there is no consistent transferability-stability trade-off across families. Flowers-102, where DINOv2's own dissociation is sharpest, is the sole dataset that shows one.

G.5 Seed Stability of Vision Models

The $\text{Shesha}_{\text{FS}}$ estimator averages over $K = 30$ random feature partitions, and the partition sequence is fixed by a random seed. To confirm that the benchmark rankings are not artifacts of a particular partition sequence, we recomputed the entire CIFAR-10 sweep under three independent seeds (9, 320, and 1991), each generating a different set of $K = 30$ partitions, for all 170 models. For each model we summarized the three per-seed scores by their mean, standard deviation, and coefficient of variation (CV, standard deviation divided by mean).

Table S9: Hierarchical vs. columnar stability: Mann-Whitney U tests on $\text{Shesha}_{\text{FS}}$ comparing hierarchical, multi-scale architectures (e.g. Swin, PVT, CoAtNet, and convolutional networks; n=81) to columnar, single-scale ones (e.g. ViT, DeiT; n=89). $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. The advantage is significant only on Flowers-102.

Dataset	Hier. (n=81)	Col. (n=89)	Δ	p
CIFAR-10	0.71 ± 0.09	0.70 ± 0.16	+0.01	0.891
CIFAR-100	0.67 ± 0.10	0.64 ± 0.20	+0.03	0.646
Flowers-102	0.86 ± 0.11	0.75 ± 0.19	+0.11	$<0.001^{***}$
DTD	0.61 ± 0.14	0.61 ± 0.14	+0.00	0.582
EuroSAT	0.86 ± 0.06	0.90 ± 0.11	−0.04	1.000
Oxford Pets	0.67 ± 0.13	0.74 ± 0.14	−0.07	1.000

Rankings are nearly invariant to the seed. Between any pair of seeds, $\text{Shesha}_{\text{FS}}$ rankings agree at Kendall $\tau \in [0.944, 0.945]$ and raw values at Spearman $\rho \in [0.993, 0.995]$; LogME is even more stable ($\tau \in [0.988, 0.991]$, $\rho \geq 0.9996$). Per-model variation is small: the median $\text{Shesha}_{\text{FS}}$ CV is 0.75% (LogME 1.08%), and only 5 of 170 models (2.9%) exceed a 5% CV (3 of 170, 1.8%, for LogME). The five high-variability models are mid-to-low-stability architectures (deit3_base, vit_tiny, vit_base.mae, wide_resnet50_2, densenet201) whose small absolute fluctuations inflate the ratio; in every case the absolute $\text{Shesha}_{\text{FS}}$ standard deviation across seeds is below 0.06.

The DINOv2 paradox is itself seed-invariant. The least geometrically stable model in the benchmark, the register-augmented large DINOv2 variant, holds rank 170 of 170 under all three seeds ($\text{Shesha}_{\text{FS}} = 0.291, 0.287, 0.297$), and the DINOv2 family records the lowest family-mean $\text{Shesha}_{\text{FS}}$ under every seed.

A Friedman test across the 170 paired models does detect a small systematic difference between seeds for $\text{Shesha}_{\text{FS}}$ ($\chi^2 = 8.48$, $p = 0.014$; for LogME $\chi^2 = 175.78$, $p < 10^{-6}$). This reflects the sensitivity of a paired test at $n = 170$ to sub-percent shifts rather than any practical instability: the effect is negligible in magnitude (median CV 0.75%) and leaves the rankings essentially unchanged. We therefore report seed 320 throughout the main text and treat the metric as reproducible across partition seeds.

G.6 SAM vs. SGD Ablation: Full Protocol and Extended Results

Sharpness-Aware Minimization (SAM; Foret et al. 2021) penalizes loss-landscape curvature by maximizing loss within an ℓ_2-ball of radius ρ before each gradient step. If $\text{Shesha}_{\text{FS}}$ responds to optimization geometry, it should change with the flatness penalty even when accuracy and learned features remain approximately constant.

G.6.1 Protocol

We trained ResNet-18 models on CIFAR-10 and CIFAR-100 using identical hyperparameters: SGD base optimizer with learning rate 0.05, momentum 0.9, weight decay 5×10^{-4}, batch size 128, cosine annealing over 100 epochs. The only variable was the SAM pertur-

Table S10: CIFAR-10 seed stability across three feature-partition seeds (9, 320, 1991) for all 170 models. Median $\text{Shesha}_{\text{FS}}$ and LogME per seed, and Spearman ρ between seeds on raw values. A Friedman test detects a small systematic seed effect for $\text{Shesha}_{\text{FS}}$ ($\chi^2 = 8.48$, $p = 0.014$); its magnitude is negligible (median per-model CV 0.75%) and rank agreement is near-perfect ($\rho \geq 0.993$, Kendall $\tau \geq 0.944$).

Metric	Seed 9	Seed 320	Seed 1991
Median LogME	0.5486	0.5615	0.5566
Median $\text{Shesha}_{\text{FS}}$	0.7167	0.7137	0.7161
Spearman ρ (LogME)	9 vs 320: 0.9997	9 vs 1991: 0.9996	320 vs 1991: 0.9996
Spearman ρ ($\text{Shesha}_{\text{FS}}$)	9 vs 320: 0.9944	9 vs 1991: 0.9930	320 vs 1991: 0.9948

bation radius $\rho \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2\}$, where $\rho = 0$ recovers standard SGD. Each configuration was trained over 15 random seeds. The ResNet-18 architecture was adapted for CIFAR with a 3×3 initial convolution (stride 1, padding 1) and identity max-pooling. After training, we extracted 512-dimensional penultimate-layer representations (post-average-pooling) for 2,000 test images, and computed test accuracy, debiased linear CKA against the SGD baseline, $\text{Shesha}_{\text{FS}}$ ($K = 30$ splits), and three supervised Shesha variants (variance ratio, supervised alignment, class separation ratio), which we introduced in another work (Raju, 2026b). The label-aware variants are: the variance ratio (between-class to total variance), a supervised alignment score, and a class-separation ratio; all three increase with the concentration of variance along class-discriminative directions and serve here only as a contrast to the unsupervised $\text{Shesha}_{\text{FS}}$. All metrics are reported as mean $\pm$ SD over the 15 seeds $(3, 7, 9, 11, 12, 18, 103, 108, 320, 411, 724, 1754, 1991, 2222, 7258)$.

G.6.2 Results

Tables S11 and S12 report all metrics. Three patterns hold.

First, the intervention dissociates $\text{Shesha}_{\text{FS}}$ from CKA. As ρ increases, CKA against the SGD baseline falls steadily, from 1.000 to 0.925 on CIFAR-10 and, more sharply, from 1.000 to 0.772 by $\rho = 0.01$ on CIFAR-100, where it then stays near 0.77. $\text{Shesha}_{\text{FS}}$ does not follow this decline. At the peak radius it exceeds the SGD baseline on every one of the 15 seeds on both datasets (CIFAR-10, $\rho = 0.05$: +0.067, paired $t_{14} = 25.3$, $p = 4 \times 10^{-13}$; CIFAR-100, $\rho = 0.2$: +0.018, $t_{14} = 16.6$, $p = 1 \times 10^{-10}$; two-sided), so a single training knob holds accuracy fixed, moves the representation in CKA, and moves $\text{Shesha}_{\text{FS}}$ in the opposite direction.

Second, the optimum is interior and dataset-dependent rather than monotone. On CIFAR-10 $\text{Shesha}_{\text{FS}}$ peaks at $\rho = 0.05$–0.1 (0.872) and declines at $\rho = 0.2$ (0.851), consistent with the over-regularization at large perturbation radii documented by Andriushchenko and Flammarion (2022); the variance ratio shows the matching reversal, bottoming at $\rho = 0.1$ (0.786) and rising at $\rho = 0.2$ (0.790), which indicates the $\text{Shesha}_{\text{FS}}$ drop reflects a real change in how variance is distributed across coordinates rather than seed noise. On CIFAR-100 the rise is shallower and shifts to the high end of the sweep, peaking at $\rho = 0.2$ (0.822). We report both shapes rather than averaging across datasets.

Third, the label-aware variants move opposite to $\text{Shesha}_{\text{FS}}$. On CIFAR-10 the variance ratio declines from 0.849 to 0.790 and the class-separation ratio from 3.08 to 2.41 across the sweep, while the supervised alignment stays flat (≈ 0.51); CIFAR-100 shows the same pattern. This confirms that $\text{Shesha}_{\text{FS}}$ captures coordinate-basis redundancy rather than classification geometry: SAM distributes features more uniformly, raising split-half consistency while reducing the concentration of variance along class-discriminative directions.

Table S11: SAM ablation, CIFAR-10 (full results), mean ± SD over 15 seeds.

ρ	Test Acc. (%)	CKA vs. SGD	$\text{Shesha}_{\text{FS}}$	Var. Ratio	Sup. Align.	Class Sep.
0.00 (SGD)	94.91 ± 0.15	1.000	0.806 ± 0.008	0.849 ± 0.004	0.511 ± 0.004	3.08 ± 0.05
0.01	95.07 ± 0.13	0.949 ± 0.001	0.831 ± 0.008	0.830 ± 0.004	0.511 ± 0.004	2.81 ± 0.04
0.02	95.29 ± 0.19	0.946 ± 0.001	0.851 ± 0.005	0.817 ± 0.003	0.511 ± 0.004	2.65 ± 0.03
0.05	95.46 ± 0.08	0.938 ± 0.002	0.872 ± 0.007	0.796 ± 0.004	0.511 ± 0.004	2.46 ± 0.03
0.10	95.56 ± 0.14	0.933 ± 0.002	0.872 ± 0.008	0.786 ± 0.003	0.511 ± 0.003	2.38 ± 0.02
0.20	95.50 ± 0.12	0.925 ± 0.003	0.851 ± 0.020	0.790 ± 0.003	0.511 ± 0.004	2.41 ± 0.02

Table S12: SAM ablation, CIFAR-100 (full results), mean ± SD over 15 seeds.

ρ	Test Acc. (%)	CKA vs. SGD	$\text{Shesha}_{\text{FS}}$	Var. Ratio	Sup. Align.	Class Sep.
0.00 (SGD)	76.96 ± 0.21	1.000	0.805 ± 0.003	0.562 ± 0.003	0.159 ± 0.004	1.505 ± 0.005
0.01	77.11 ± 0.25	0.772 ± 0.003	0.805 ± 0.003	0.542 ± 0.003	0.158 ± 0.005	1.468 ± 0.006
0.02	77.32 ± 0.22	0.773 ± 0.002	0.805 ± 0.003	0.533 ± 0.002	0.158 ± 0.004	1.452 ± 0.004
0.05	77.48 ± 0.20	0.777 ± 0.002	0.806 ± 0.003	0.520 ± 0.003	0.158 ± 0.004	1.430 ± 0.005
0.10	77.82 ± 0.24	0.775 ± 0.002	0.814 ± 0.004	0.504 ± 0.002	0.157 ± 0.005	1.406 ± 0.004
0.20	78.05 ± 0.20	0.765 ± 0.002	0.822 ± 0.003	0.485 ± 0.002	0.156 ± 0.005	1.381 ± 0.004

G.7 Probe Subset-Sensitivity: Full Protocol and Stratified Analysis

This subsection tests whether geometric stability has practical diagnostic value. If a representation scores low on $\text{Shesha}_{\text{FS}}$, its distance geometry is not recoverable from feature subsets, so linear probes trained on different halves of the features should disagree in accuracy. We measure this across 170 vision models, control for the probe-accuracy ceiling, and identify where in the model distribution the relationship holds and where it attenuates.

G.7.1 Protocol

For each of 170 vision models we extracted 512- to 1536-dimensional penultimate-layer representations on a fixed 5,000-image CIFAR-10 subset (seed 320). We partitioned the samples once into a 60% train and 40% test split, stratified by class, and held this split fixed across all subsequent probes so that performance variability would reflect feature-subset choice rather than sample choice. For each model we drew 20 random halves of the feature dimensions (subset fraction 0.5). On each half we standardized features using training-split statistics, trained a logistic-regression probe, and recorded test accuracy. We summarized each model by the mean, standard deviation, and range of probe accuracy across the 20 subsets, and by the coefficient of variation (standard deviation divided by mean). $\text{Shesha}_{\text{FS}}$ was computed on the full clean representation with $K = 30$ splits.

G.7.2 Headline Result

Across the 170 models, $\text{Shesha}_{\text{FS}}$ predicts probe-accuracy standard deviation ($\rho = -0.302$, $p = 6.4 \times 10^{-5}$) and range ($\rho = -0.260$, $p = 6.1 \times 10^{-4}$). Because a probe near ceiling accuracy has limited room to vary, we ran two controls. The coefficient-of-variation correlation ($\rho = -0.280$, $p = 2.2 \times 10^{-4}$) confirms the effect is not a ceiling artifact, and the partial correlation controlling for mean probe accuracy is in fact stronger than the raw correlation ($\rho_{\text{partial}} = -0.382$, $p = 3.0 \times 10^{-7}$), indicating that probe accuracy was suppressing rather than inflating the relationship.

G.7.3 Stratified analysis

Splitting the models into accuracy terciles shows that the relationship is strongest in the middle of the distribution (mid-accuracy tercile: $\rho = -0.473$, $p = 2.3 \times 10^{-4}$; high-accuracy: $\rho = -0.261$, $p = 0.05$; low-accuracy: $\rho = -0.072$, $p = 0.59$). The weak relationship in the low-accuracy tercile reflects a boundary effect: the most extreme low-stability models, such as the DINOv2 family ($\text{Shesha}_{\text{FS}} \approx 0.29$ to 0.37), produce probes that are consistently mediocre across subsets rather than highly variable. These models distribute their geometry non-redundantly across many coordinates rather than concentrating it; DINOv2 has the highest participation ratio in the benchmark (Section 4.4). When no random half recovers the full discriminative structure, every probe is limited in the same way, so probe accuracy is uniformly low rather than variable. The predictive relationship between $\text{Shesha}_{\text{FS}}$ and probe variability therefore holds across the bulk of the model distribution but attenuates at the extreme low-stability tail, where non-redundant coding caps every subset at similar accuracy.

Appendix H. Code Availability

All custom code is available on GitHub (`https://github.com/prashantcraju/geometric-stability`, Raju (2026a)). We have also released an open source Python library through PyPI (`https://pypi.org/project/shesha-geometry`; Raju (2026c)).

References

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.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop Track Proceedings*, 2017.

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

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, 2022.

Itamar Avitan and Tal Golan. Model-behavior alignment under flexible evaluation: When the best-fitting model isn't the right one. In *Advances in Neural Information Processing Systems*, 2025.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099), 2026. doi: 10.1038/s41586-025-10014-0.

Horace Barlow. Possible principles underlying the transformations of sensory messages. *Sensory Communication*, 1:217–234, January 1961. doi: 10.7551/mitpress/9780262518420.003.0013.

Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Shadab Ahmad, Emanuele Alpi, Emily H Bowler-Barnett, Ramona Britto, Hema Bye-A-Jee, Austra Cukura, Paul Denny, Tunca Dogan, ThankGod Ebenezer, Jun Fan, Penelope Garmiri, Leonardo Jose da Costa Gonzales, Emma Hatton-Ellis, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Marija Lugaric, Jie Luo, Yvonne Lussi, Alistair MacDougall, Fabio Madeira, Mahdi Mahmoudy, Alok Mishra, Katie Moulang, Andrew Nightingale, Sangya Pundir, Guoying Qi, Shriya Raj, Pedro Raposo, Daniel L Rice, Rabie Saidi, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Edward Turner, Nidhi Tyagi, Preethi Vasudev, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Alan J Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H Auchincloss, Kristian B Axelsen, Parit Bansal, Delphine Baratin, Teresa M Batista Neto, Marie-Claude Blatter, Jerven T Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard de Castro, Anne Estreicher, Maria L Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Venkatesh Muthukrishnan, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J A Sigrist, Karin Sonesson, Shyamala Sundaram, Cathy H Wu, Cecilia N Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Qinghua Wang, Yuqi Wang, and Jian Zhang. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2022. ISSN 1362-4962. doi: 10.1093/nar/gkac1052.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967.e21, November 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.09.031. URL http://dx.doi.org/10.1016/j.cell.2020.09.031.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Vasudev Shyam, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. Do sparse autoencoders capture concept manifolds? *arXiv*, 2026.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Koulako Bala Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv*, 2021.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Jonathan C. Schmok, Ali Taghibakhshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Michael T. Pearce, Elana Simon, Etowah Adams, Zachary J. Amador, Euan A. Ashley, Stephen A. Baccus, Haoyu Dai, Steven Dillmann, Stefano Ermon, Daniel Guo, Michael H. Herschl, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, John St. John, Jeremy Sullivan, Joseph Tey, Ben Viggiano, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Sudarshan Pinglay, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modelling and design across all domains of life with Evo 2. *Nature*, March 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10176-5. URL http://dx.doi.org/10.1038/s41586-026-10176-5.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, April 2018. ISSN 1546-1696. doi: 10.1038/nbt.4096. URL http://dx.doi.org/10.1038/nbt.4096.

N Alex Cayco-Gajic and Arthur Pellegrino. Geometry-aware similarity metrics for neural representations on riemannian and statistical manifolds. *arXiv*, 2026.

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. *Nature*, 487(7405):51–56, June 2012. ISSN 1476-4687. doi: 10.1038/nature11129. URL http://dx.doi.org/10.1038/nature11129.

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.

Nathan Cloos, Moufan Li, Markus Siegel, Scott L Brincat, Earl K Miller, Guangyu Robert Yang, and Christopher J Cueva. Differentiable optimization of similarity scores between models and brains. In *International Conference on Learning Representations*, 2025.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge Member of the Taylor and Francis Group, 2nd edition, August 1988. ISBN 978-0805802832.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1), October 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53147-y. URL http://dx.doi.org/10.1038/s41467-024-53147-y.

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

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2002. ISSN 1098-2418. doi: 10.1002/rsa.10073.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations*, 2023.

Grégoire Dhimoïla, Victor Boutin, Agustin Martin Picard, Thomas Fel, and Thomas Serre. A unifying framework for concept-based representational similarity. *arXiv*, 2026.

Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508, 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005508.

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

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, 2021.

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.

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

Shimon Edelman. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(4):449–467, August 1998. ISSN 1469-1825. doi: 10.1017/s0140525x98001253. URL `http://dx.doi.org/10.1017/s0140525x98001253`.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Jenelle Feather, Guillaume Leclerc, Aleksander Mądry, and Josh H. McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, October 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01442-0. URL `http://dx.doi.org/10.1038/s41593-023-01442-0`.

Thomas Fel, David Vigouroux, Rémi Cadène, and Thomas Serre. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 720–730, January 2022.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Robert Geirhos, Roland S. Zimmermann, Blair Bilodeau, Wieland Brendel, and Been Kim. Don't trust your eyes: on the (un)reliability of feature visualizations. In *International Conference on Machine Learning*, 2024.

Joshua I. Gold and Michael N. Shadlen. The neural basis of decision making. *Annual Review of Neuroscience*, 30(1):535–574, July 2007. ISSN 1545-4126. doi: 10.1146/annurev.neuro.29.051605.113038. URL `http://dx.doi.org/10.1146/annurev.neuro.29.051605.113038`.

Sarah E Harvey, David Lipshutz, and Alex H Williams. What representational similarity measures imply about decodable information. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

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.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Conference on Modern Analysis and Probability*, page 189–206, 1984. ISSN 0271-4132. doi: 10.1090/conm/026/737400.

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. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL `http://dx.doi.org/10.1038/s41586-021-03819-2`.

Matthew T Kaufman, Mark M Churchland, Stephen I Ryu, and Krishna V Shenoy. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience*, 17(3):440–448, February 2014. ISSN 1546-1726. doi: 10.1038/nn.3643. URL `http://dx.doi.org/10.1038/nn.3643`.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019a.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019b.

Nikolaus Kriegeskorte and Rogier A. Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, August 2013. ISSN 1364-6613. doi: 10.1016/j.tics.2013.06.007. URL `http://dx.doi.org/10.1016/j.tics.2013.06.007`.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, 2009.

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.

Tanishq Kumar, Blake Bordelon, Cengiz Pehlevan, Venkatesh N Murthy, and Samuel J. Gershman. Do mice grok? Glimpses of hidden progress in sensory cortex. In *International Conference on Learning Representations*, 2025.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, 2025.

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.

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. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42), October 2024. ISSN 1091-6490. doi: 10.1073/pnas.2317881121. URL `http://dx.doi.org/10.1073/pnas.2317881121`.

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. *Science*, 379(6637), 2023. doi: 10.1126/science.ade2574.

Sunny Liu, Habon Issa, André Longon, Liv Gorton, Meenakshi Khosla, and David Klindt. Measuring the representational alignment of neural systems in superposition. *arXiv*, 2026.

Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6), June 2019. ISSN 1744-4292. doi: 10.15252/msb.20188746. URL `http://dx.doi.org/10.15252/msb.20188746`.

Florian P. Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N. Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature*

Machine Intelligence, 7(6):848–859, June 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01041-7. URL http://dx.doi.org/10.1038/s42256-025-01041-7.

Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, November 2013. ISSN 1476-4687. doi: 10.1038/nature12742. URL http://dx.doi.org/10.1038/nature12742.

Valentina Masarotto, Victor M. Panaretos, and Yoav Zemel. Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhya A*, 81(1):172–213, 2018. ISSN 0976-8378. doi: 10.1007/s13171-018-0130-1.

Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. Individual differences among deep neural network models. *Nature Communications*, 11 (1), November 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19632-w. URL http://dx.doi.org/10.1038/s41467-020-19632-w.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

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.

Alex Graeme Murphy, Joel Zylberberg, and Alona Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

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. *Nature*, 647 (8089):349–355, November 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09631-6. URL http://dx.doi.org/10.1038/s41586-025-09631-6.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, 2020.

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.

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.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553, 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003553.

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.

Ramon Nogueira, Chris C. Rodgers, Randy M. Bruno, and Stefano Fusi. The geometry of cortical representations of touch in rodents. *Nature Neuroscience*, 26(2):239–250, January 2023. ISSN 1546-1726. doi: 10.1038/s41593-022-01237-9. URL http://dx.doi.org/10.1038/s41593-022-01237-9.

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*, pages 5206–5210. IEEE, 2015.

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. *Nature Methods*, 15(10):805–815, September 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0109-9. URL http://dx.doi.org/10.1038/s41592-018-0109-9.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS*, 2023.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025.

Alethea Power, Yuri Burda, Harrison Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv*, 2022.

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.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, 2021.

Prashant C. Raju. Github repository. https://github.com/prashantcraju/geometric-stability, 2026a.

Prashant C. Raju. The geometric canary: Predicting steerability and detecting drift via representational stability. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026b.

Prashant C. Raju. Shesha: Self-consistency metrics for representational stability. Zenodo, 2026c. URL https://doi.org/10.5281/zenodo.18227453.

F. James Rohlf and Dennis Slice. Extensions of the procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, 39(1):40, 1990. ISSN 0039-7989. doi: 10.2307/2992207.

Shreya Saxena and John P Cunningham. Towards the neural population doctrine. *Current Opinion in Neurobiology*, 55:103–111, April 2019. ISSN 0959-4388. doi: 10.1016/j.conb.2019.02.002. URL http://dx.doi.org/10.1016/j.conb.2019.02.002.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *International Conference on Machine Learning*, 2024.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. ISSN 1860-0980. doi: 10.1007/bf02289451.

Heiko H Schütt. Bayesian comparisons between representations. In *Conference on Cognitive Computational Neuroscience*, 2025.

Heiko H Schütt, Alexander D Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. Statistical inference on representational geometries. *eLife*, 12, 2023. ISSN 2050-084X. doi: 10.7554/elife.82566.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv*, 2025.

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.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47): 1393–1434, 2012.

Khurram Soomro, Amir Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, abs/1212.0402, 2012. URL https://api.semanticscholar.org/CorpusID:7197134.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43), October 2022. ISSN 1091-6490. doi: 10.1073/pnas.2200800119. URL http://dx.doi.org/10.1073/pnas.2200800119.

Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576 (7786):266–273, 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1787-x.

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. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

David Sussillo, Mark M Churchland, Matthew T Kaufman, and Krishna V Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, 18(7):1025–1033, June 2015. ISSN 1546-1726. doi: 10.1038/nn.4042. URL `http://dx.doi.org/10.1038/nn.4042`.

Sina Tafazoli, Flora M. Bouchacourt, Adel Ardalan, Nikola T. Markov, Motoaki Uchimura, Marcelo G. Mattar, Nathaniel D. Daw, and Timothy J. Buschman. Building compositional tasks with shared neural subspaces. *Nature*, 650(8100):164–172, November 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09805-2. URL `http://dx.doi.org/10.1038/s41586-025-09805-2`.

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. *Nature Biotechnology*, 32(4):381–386, March 2014. ISSN 1546-1696. doi: 10.1038/nbt.2859. URL `http://dx.doi.org/10.1038/nbt.2859`.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv*, 2023.

Jean Philippe Vogel. *Indian Serpent-Lore: Or, The Nāgas in Hindu Legend and Art*. Arthur Probsthain, London, 1926.

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

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Alex H. Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 2018. doi: 10.1186/s13059-017-1382-0.

Jialin Wu, Shreya Saha, Yiqing Bo, and Meenakshi Khosla. Comparing and integrating different notions of representational correspondence in neural systems. *arXiv*, 2026.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, May 2014. ISSN 1091-6490. doi: 10.1073/pnas.1403112111. URL http://dx.doi.org/10.1073/pnas.1403112111.

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.

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. *Journal of Machine Learning Research*, 23(1), 2022. ISSN 1532-4435.

Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, February 2020. ISSN 1091-6490. doi: 10.1073/pnas.1901326117. URL http://dx.doi.org/10.1073/pnas.1901326117.

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. *arXiv*, 2019.

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. *Nature Communications*, 8(1), 2017. doi: 10.1038/ncomms14049.

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. *Proceedings of the National Academy of Sciences*, 118(3), January 2021. ISSN 1091-6490. doi: 10.1073/pnas.2014196118. URL http://dx.doi.org/10.1073/pnas.2014196118.

Roland S. Zimmermann, Thomas Klein, and Wieland Brendel. Scale alone does not improve mechanistic interpretability in vision models. In *Advances in Neural Information Processing Systems*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv*, 2023.