Geometric Stability: The Missing Axis of Representations

Prashant C. Raju
rajuprashant@gmail.com

Abstract

Analysis of learned representations has a blind spot: it focuses on *similarity*, measuring how closely embeddings align with external references, but similarity reveals only what is represented, not whether that structure is robust. We introduce *geometric stability*, a distinct dimension that quantifies how reliably representational geometry holds under perturbation, and present *Shesha*, a framework for measuring it. Across 2,463 configurations in seven domains, we show that stability and similarity are empirically uncorrelated ($\rho \approx 0.01$) and mechanistically distinct: similarity metrics collapse after removing the top principal components, while stability retains sensitivity to fine-grained manifold structure. This distinction yields actionable insights: for safety monitoring, stability acts as a functional geometric canary, detecting structural drift nearly $2\times$ more sensitively than CKA while filtering out the non-functional noise that triggers false alarms in rigid distance metrics; for controllability, supervised stability predicts linear steerability ($\rho = 0.89$-0.96); for model selection, stability dissociates from transferability, revealing a geometric tax that transfer optimization incurs. Beyond machine learning, stability predicts CRISPR perturbation coherence and neural-behavioral coupling. By quantifying *how reliably* systems maintain structure, geometric stability provides a necessary complement to similarity for auditing representations across biological and computational systems.

1 Introduction

Neural representation analysis has long focused on *similarity*: measuring how closely the internal representations of different models are aligned (Kornblith et al., 2019a; Raghu et al., 2017; Morcos et al., 2018; Kriegeskorte et al., 2008). Standard tools like Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) and Centered Kernel Alignment (CKA) (Kornblith et al., 2019a) have become the default for comparing representations between architectures, training runs, and domains. Yet, similarity answers only half of the question. Two models may align perfectly and still diverge under intervention. Two representations may match on content, yet differ entirely in robustness. What similarity misses is *geometric stability*: how consistently a model's representational structure holds under perturbation, resampling, or context shift.

Similarity is not stability. Consider a vast library. Its utility depends not merely on its inventory of books, but on the preservation of its indexing system. If every book were randomly reshuffled, the collection would remain identical, and any content-based audit would confirm that nothing is missing. Yet, the library would become functionally useless: the spatial relationships that enable retrieval have been destroyed. For foundation models, geometric stability plays the role of this index. Standard similarity metrics confirm that features are present, but fail to detect when the fine-grained relationships between them have fractured.

Preprint. Under review.

This dissociation is not hypothetical. Two models may appear identical under CKA. The first maintains a rigid geometric structure when evaluated on bootstrapped data or paraphrased inputs. The second shows a highly fluctuating geometry under the same conditions: its structure is brittle (Ilyas et al., 2019) despite perfect alignment with the first. We show that these dimensions are empirically *distinct*. Across 2,463 encoder configurations in 7 domains, stability and similarity show near-zero correlation ($\rho \approx 0.01$), encompassing all four quadrants of the stability-similarity space (Figure 6). Spectral deletion experiments offer an explanation. Standard similarity metrics collapse to noise after removing just the top principal components, whereas geometric stability retains sensitivity to fine-grained manifold structure distributed across the spectral tail.

Why stability matters. Geometric stability predicts what similarity cannot. In steering experiments, **supervised stability** strongly predicts intervention success ($\rho = 0.89$-0.96, $p < 10^{-18}$): models with task-aligned geometry accept linear steering vectors, while unstable models fracture under the same perturbations (Section 3.1). In vision, stability and transferability dissociate entirely: state-of-the-art models like DINOv2 rank last in stability despite showing the top performance for transfer (Section 3.2). In instruction tuning, geometric stability detects structural drift nearly $2\times$ more sensitively than CKA on average (up to $5.23\times$ in Llama), providing earlier warning of functional degradation in 73% of models while avoiding false alarms of rigid distance metrics such as Procrustes ($\rho \approx 0.93$; Section 3.3). These patterns extend beyond machine learning: in CRISPR screens, stability tracks the coherence of regulatory perturbations (Section 3.4); in brain recordings, it reveals region-specific dynamics invisible to similarity (Section 3.5). The common thread is that stability captures the *usability* of a geometric structure, not its presence.

Our approach: Shesha. We introduce *Shesha*, named for the Hindu deity representing the invariant remainder of the cosmos (Vogel, 1995; Daniélou, 1964; Dimmitt and van Buitenen, 1978), a framework for measuring geometric stability through *self-consistency*. Unlike similarity metrics that compare external representations *between* models, Shesha quantifies internal *within-model* reliability, providing an independent axis for diagnosing fine-tuning dynamics, safety audits, and high-dimensional interpretation.

Shesha operates on Representational Dissimilarity Matrices (RDMs) from RSA (Kriegeskorte et al., 2008), assessing the consistency of RDMs derived from perturbed or resampled views of the *same* representation rather than comparing across systems (Appendix 6.1). Our split-half approach adapts the statistical technique of noise ceiling estimation (Nili et al., 2014), but generalizes it for an alternate purpose: rather than bounding explainable variance to normalize cross-system comparisons, we treat internal consistency itself as the subject of interest. This reframing enables applications ranging from foundation model embeddings to CRISPR perturbation screens.

We employ two complementary variants (Appendix 6.1.1): (1) **Feature-Split Shesha** ($\text{Shesha}_{\text{FS}}$) measures internal geometric consistency by correlating RDMs from disjoint feature subsets, which is suited for drift detection and biological structure; (2) **Sample-Split Shesha** ($\text{Shesha}_{\text{SS}}$) measures robustness to input variation across disjoint data subsets, becoming **Supervised Shesha** ($\text{Shesha}_{\text{Sup}}$) when task labels are present. Unless noted, "Shesha" refers to the unsupervised $\text{Shesha}_{\text{FS}}$ variant.

2 Distinctness of Stability and Similarity

2.1 Geometric Intuition

Stability and similarity extract distinct information from representations. *Similarity* measures how a representation aligns with external references, an extrinsic property that quantifies correspondence between two representational spaces. *Stability* measures how consistently a representation preserves pairwise distances across feature subsets, an intrinsic property of the geometric manifold itself, independent of any reference. A model can match a reference on dominant structure while having degraded fine-grained geometry, or maintain rigid manifold structure while differing from references. PCA-reduced encoders exemplify this dissociation. They preserve high similarity to full models (dominant variance intact) while showing reduced stability (manifold structure compressed), yielding the negative correlation ($\rho = -0.47$) observed in our regime analysis. In contrast, geometry-preserving transformations, such as random projections, maintain both metrics ($\rho = +0.90$), consistent with the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2002).

This regime-dependent relationship (redundancy under geometry preservation, dissociation under compression) explains why aggregate near-zero correlation emerges and identifies when Shesha provides diagnostic value beyond similarity metrics.

2.2 Ground Truth Validation

We validated Shesha's construct and discriminant validity through controlled synthetic experiments (Appendix 7.1). First, representations with known stability levels (signal-to-noise ratios from 0 to 1) confirmed Shesha recovers ground truth with near-perfect accuracy ($\rho = 0.997$, $p < 10^{-86}$). Second, dissociation cases spanning all four quadrants of the stability-similarity space demonstrated that high stability can occur with low similarity and vice versa, including adversarial cases where CKA exceeds 0.97 despite near-zero Shesha. Using debiased CKA, Shesha and similarity showed weak correlation ($\rho = 0.22$). Third, spectral deletion experiments revealed that all similarity metrics (debiased CKA, PWCCA, Procrustes) collapsed below 0.4 after removing just the top principal component, while Shesha remained above 0.4 until $k = 26$ components were removed (Figure 1a). At $k = 30$, Shesha retained $110\times$ higher signal than CKA. This divergence held across preprocessing conditions, though whitening caused CKA to recover by artificially equalizing the spectrum. These results demonstrate that Shesha captures geometric structure distributed across the eigenspectrum that similarity metrics ignore entirely.

2.3 Experimental Design

7 domains, 2,463 encoder configurations. We validate that stability and similarity are distinct dimensions across seven computational domains covering 2,463 encoder configurations (Table 1). To ensure methodological consistency, we applied standardized encoder transformations (including PCA at various ranks, random projections, feature subsets, noise perturbations, and normalization variants) to base representations across all domains. This design tests whether stability and similarity remain distinct under controlled geometric operations.

Table 1: Distinctness across domains. Aggregate correlation is negligible ($\rho = -0.01$, CI within ± 0.06); four domains show negligible correlations ($|\rho| < 0.10$); Protein shows moderate negative correlation driven by low-dimensional encoders.

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

[a]Protein shows moderate negative correlation driven by PCA on low-dimensional sequence encoders (20–500 dims); see Appendix 7.6.

Machine learning domains (448 configurations): Language embeddings from sentence transformers (MiniLM, MPNet, DistilBERT, RoBERTa; $N = 127$), vision embeddings from ViT, CLIP, DeiT, and ResNet50 ($N = 129$), audio embeddings from Wav2Vec2 and HuBERT ($N = 64$), and video embeddings from TimeSformer, VideoMAE, and frame-level encoders ($N = 128$). **Biological domains** (2,015 configurations): Protein sequence encoders including compositional, spectral, and physicochemical features applied to Swiss-Prot sequences (Bateman et al., 2022) ($N = 402$), molecular embeddings of single-cell RNA-seq features from the pbmc3k dataset (Zheng et al., 2017) ($N = 767$), and neural population activity representations extracted from the Steinmetz recordings (Steinmetz et al., 2019) containing 26 sessions across multiple brain areas ($N = 846$). This scope empirically tests if distinctness generalizes across different types of representations and

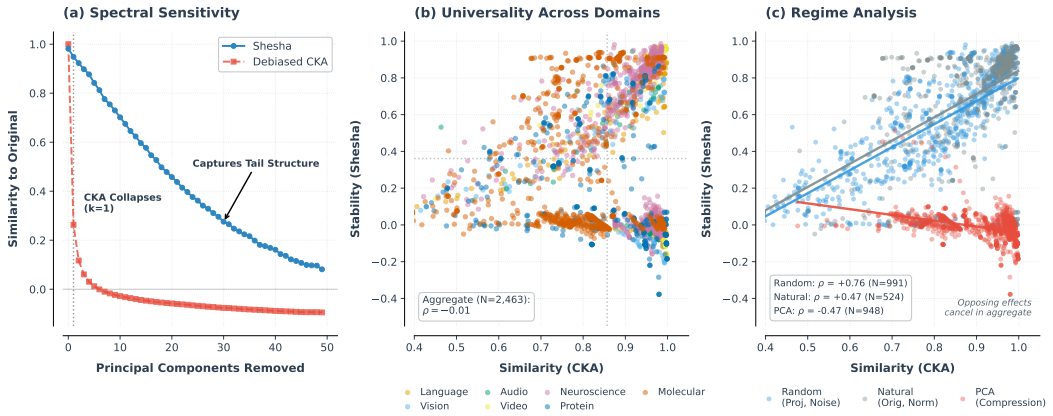

Figure 1: **Stability and similarity are independent dimensions of representational geometry. (a) Spectral Sensitivity:** CKA (red) collapses after removing just the single top principal component, while Shesha (blue) retains sensitivity to the spectral tail. CKA measures dominant variance; Shesha measures full manifold geometry. **(b) Universality:** Across 2,463 encoder configurations spanning seven domains, Shesha and CKA show negligible net correlation ($\rho = -0.01$, 95% CI $[-0.06, +0.03]$), confirming they capture distinct geometric properties. **(c) Regime Analysis:** Aggregate near-zero correlation emerges from opposing effects: random transformations yield positive correlation ($\rho = +0.76$), while PCA compression yields negative correlation ($\rho = -0.47$). These cancel in aggregate, revealing that Shesha specifically detects compression-induced damage invisible to CKA.

transformation regimes. For each encoder configuration, we measured Shesha (geometric stability) and CKA (similarity to domain-specific reference representations). The results were aggregated across 15 random seeds (Appendix 7).

Hypothesis and predictions. Our analysis establishes that geometric stability and cross-model similarity are empirically distinct dimensions. If this theoretical distinctness holds in practice, it should produce near-zero correlation between Shesha and CKA across encoder configurations. We evaluate this using Spearman correlation with 10,000 bootstrap replicates under the null hypothesis $H_0 : \rho = 0$. We predict: (i) domain-level correlations with $|\rho| < 0.3$, (ii) bootstrap confidence intervals containing zero, and (iii) an aggregate correlation consistent with negligible association ($|\rho| < 0.1$; Cohen, 1988). Consistent near-zero correlations across heterogeneous domains would constitute strong empirical evidence for distinctness. Systematic deviations would challenge this conclusion.

2.4 Results

Aggregate distinctness. The aggregate analysis of 2,463 encoder configurations yields $\rho = -0.01$ [95% CI: -0.06, $+0.03$], a correlation indistinguishable from zero ($p = 0.57$) and well below the threshold for negligible association ($|\rho| < 0.10$) (Cohen, 1988). The 95% confidence interval lies entirely within the negligible range, providing strong evidence that stability and similarity are statistically distinct dimensions.

To strictly control for dependencies among encoder configurations (e.g., multiple perturbed versions of ResNet50), we fitted a Linear Mixed-Effects Model (Stability $\sim$ Similarity $+ (1 \mid \text{BaseModel})$). The fixed effect of similarity on stability remains negligible ($\beta = 0.10$, $p < 0.001$, 95% CI $[0.06, 0.15]$), confirming that distinctness is robust to model family identity and not an artifact of clustering. The Intraclass Correlation Coefficient (ICC $= 0.10$) indicates that base model identity explains less than 10% of variance in stability scores, with the remaining 90% attributable to encoder-specific properties and residual variation.

Domain-level consistency. Six of seven domains show distinctness ($|\rho| < 0.30$). Four show negligible correlations: Neuroscience ($\rho = +0.01$, CI $[-0.06, +0.09]$), Language ($\rho = +0.03$),

Vision ($\rho = -0.03$), and Molecular ($\rho = +0.06$). The highest-powered domain (Neuroscience, $N = 846$) provides the strongest individual evidence for distinctness. Audio ($\rho = -0.26$) and Video ($\rho = -0.24$) show small negative correlations below threshold. Only Protein shows moderate correlation ($\rho = -0.36$), possibly driven by PCA's interaction with low-dimensional sequence encoders (20 to 500 dimensions), where aggressive compression disproportionately affects manifold structure.

Regime analysis: when do stability and similarity diverge? The aggregate near-zero correlation masks structured relationships across encoder types (Figure 1c). We identify three regimes:

Regime 1: Geometry-preserving transformations. Random projections ($\rho = +0.90$) and feature selection ($\rho = +0.92$) yield near-perfect positive correlation. This follows from the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2002): random projections approximately preserve pairwise distances, maintaining both stability and similarity. Noise injection behaves similarly ($\rho = +0.58$). In this regime, the metrics are redundant.

Regime 2: Compression transformations. PCA yields strong *negative* correlation ($\rho = -0.47$). Dimensionality reduction preserves dominant variance (maintaining high CKA) while discarding fine-grained manifold structure (Tenenbaum et al., 2000) (reducing Shesha). This dissociation demonstrates that CKA over-indexes on principal subspaces, whereas Shesha captures full geometric integrity including spectral tail structure.

Regime 3: Natural encoders. Original representations show weak positive correlation ($\rho \approx +0.31$ to $+0.34$), indicating that Shesha contributes approximately 90% unique information ($1 - \rho^2 \approx 0.90$) beyond CKA. Real-world encoders occupy this intermediate regime, consistent with the manifold hypothesis (Bengio et al., 2013), where both metrics offer complementary diagnostic value.

These opposing effects (positive correlation under geometry-preserving transformations, negative under compression) cancel in aggregate, producing net near-zero correlation. This explains when Shesha adds value: detecting compression-induced damage to manifold structure that remains invisible to similarity metrics optimized for dominant variance.

Statistical interpretation. The aggregate correlation ($\rho = -0.01$, $p = 0.57$) indicates that stability and similarity share less than 0.1% of variance ($R^2 < 0.001$), well within the negligible range by conventional standards (Cohen, 1988). This pattern across multiple unrelated domains confirms that distinctness is a fundamental property of the stability-similarity relationship. The regime analysis clarifies the mechanism: CKA tracks dominant variance while Shesha tracks full manifold geometry, making them jointly informative for representational analysis.

Robustness to metric choice. To verify that distinctness generalizes beyond CKA, we evaluated two alternative similarity metrics in the Language domain ($N = 127$): effective-rank PWCKA ($\rho = -0.22$, $p = 0.012$) and Procrustes similarity ($\rho = +0.28$, $p = 0.001$). All metrics maintain $|\rho| < 0.30$, confirming distinctness. Notably, Procrustes shows the strongest negative correlation for PCA encoders ($\rho = -0.86$), as its explicit optimization for rotational alignment makes it particularly sensitive to dimensional reduction. This reinforces our finding that Shesha captures geometric properties distinct from similarity metrics focused on dominant subspace structure (Appendix 7.8).

3 Discovered Phenomena

3.1 Stability Predicts Steering Performance

Motivation. Steering interventions enable targeted behavioral modifications without retraining by modifying model representations by adding scaled direction vectors to activations (Turner et al., 2023; Zou et al., 2023). We test whether geometric stability predicts which models can be reliably steered.

Experimental design. Experiment 1 (Synthetic) tested 69 sentence embedding models on controlled sentiment data on 15 random seeds ($n = 1{,}035$ observations). Experiment 2 (SST-2) evaluated 35 models on binary sentiment classification on 15 random seeds ($n = 525$ observations) and Experiment 3 (MNLI) tested the same 35 models on ternary natural language inference on 15 random seeds ($n = 525$ observations). Steering vectors were created by using logistic regression on class-labeled

embeddings. We used a strict half-split approach where we divided our data into two completely separate sets, with one for calculating the metrics and the other for steering evaluation. This was done to prevent any overlap in the information used for training and testing. The geometric metrics we tested included Shesha, both supervised and unsupervised, Fisher discriminant, silhouette score, Procrustes alignment, and anisotropy (Appendix 8).

Stability captures unique signal beyond separability. The raw correlations were consistently strong: $\rho = 0.89$ (Synthetic), 0.96 (SST-2), and 0.96 (MNLI), all $p < 10^{-18}$. To isolate the contribution of stability from mere class separation, we computed partial correlations to control for the Fisher discriminant and silhouette score. Shesha maintained large effect sizes across all settings: $\rho_{\text{partial}} = 0.67$ (Synthetic), 0.76 (SST-2), and 0.62 (MNLI), all $p < 0.001$. This demonstrates that **the reliability of class structure under perturbation, geometric consistency, is a distinct factor of steerability that separability metrics overlook**. Separability may be necessary for steering, but stability ensures control.

Task alignment is vital for real-world control. A notable dissociation emerged: unsupervised stability (feature-partition Shesha) predicted steering in synthetic settings ($\rho = 0.77$, $p < 10^{-14}$) but *failed completely* on real-world tasks ($\rho = 0.10$ for SST-2, $\rho = 0.35$ for MNLI; both n.s.). Unsupervised stability did not show residual signal after controlling for separability ($\rho_{\text{partial}} < 0.10$). This confirms that **intrinsic manifold rigidity alone is not enough; stability needs to be aligned with the task at hand**. In synthetic settings where the data manifold aligns with task structure, generic rigidity is adequate. In the wild, only supervised stability enables reliable intervention.

Metric robustness persists as control margins narrow. Negative controls validated our methodology while revealing how task complexity modulates steering. Shuffled-label controls confirmed that supervised metrics reflect genuine task structure: Shesha dropped from 0.60 to -0.001 (Synthetic), 0.23 to -0.001 (SST-2), and 0.02 to -0.001 (MNLI) under label permutation (all $p < 10^{-10}$). Random-direction controls quantified signal-to-noise: true directions produced $10.8\times$ (Synthetic), $2.7\times$ (SST-2), and $1.3\times$ (MNLI) larger accuracy drops than random directions. The ratio declines with increasing task complexity, however, Shesha accurately identify steerable models even when the margin for control is narrow.

Implications. These results establish supervised geometric stability as a prerequisite for reliable linear controllability, distinct from class separability. Model rankings revealed consistent patterns: supervised contrastive models (BGE, E5, GTE) were the most steerable, while unsupervised variants and retrieval-specialized models were the least steerable, suggesting that **supervised contrastive training produces the geometric stability required for reliable intervention**. Together, these findings provide practitioners with an *a priori* diagnostic: models with high supervised Shesha will steer reliably, while those with low stability will degrade under perturbation regardless of classification accuracy.

3.2 Visual Perception: The Architecture of Stability

Motivation. Transfer learning metrics like LogME (You et al., 2021) estimate how well pretrained features support downstream tasks. We test whether distinct properties are captured by transferability and geometric stability: feature *richness* (adaptability) vs. manifold *rigidity* (consistency).

Experimental design. We evaluated 93-94 pretrained vision models that encompassed diverse architectures (ViT, Swin, ConvNeXt, ResNet, EfficientNet) and training objectives (supervised, self-supervised, contrastive, generative) across six datasets in four visual domains: CIFAR-10/100 (natural images), Flowers-102 and Oxford Pets (fine-grained), DTD (texture), and EuroSAT (remote sensing). All metrics were computed on the features of the penultimate-layer.

Stability and transferability dissociate under task complexity. On CIFAR-10, Shesha-FS showed a negligible correlation with LogME ($\rho = -0.07$). However, as the task complexity increased, the relationship varied by domain: CIFAR-100 showed a negative trend ($\rho = -0.19$), and Flowers-102 showed a significant negative correlation ($\rho = -0.21$, $p < 0.05$). Oxford Pets exhibited the strongest negative correlation between Shesha-Var and Shesha-FS ($\rho = -0.42$, $p < 0.001$), indicating that

models with higher variance showed lower stability on fine-grained tasks. EuroSAT was a notable exception, showing a positive LogME-Shesha-FS correlation ($\rho = 0.45$, $p < 0.001$), possibly due to uniform texture distributions in satellite imagery.

The DINOv2 Paradox: universal across domains. The stability-transferability inversion persisted across all six datasets (Figure 15). DINOv2 achieved the highest mean LogME on 4/6 datasets but ranked last or near-last in geometric stability on all except EuroSAT. On CIFAR-100: LogME $= 1.36$ (rank 1/29 families), Shesha-FS $= 0.27$ (rank 28/29). On Flowers-102: LogME $= 2.47$ (rank 1/29), Shesha-FS $= 0.34$ (rank 29/29). On CIFAR-10, DINOv2 ranked first in LogME but last in stability (rank 29/29). Optimizing for state-of-the-art transfer appears to incur a "geometric tax" by collapsing the manifold structure required for stability. The sole exception was EuroSAT, where DINOv2 achieved top stability (Shesha-FS $= 0.95$, rank 3/29) while also ranking second in LogME (rank 2/29). This suggests that DINOv2's self-distillation objective may be uniquely suited to satellite imagery geometry.

Training objectives and architecture determine stability. Two factors consistently predicted geometric stability: **contrastive alignment and hierarchical architecture**. CLIP models outperformed self-supervised alternatives in 4 out of 6 domains ($p < 0.05$), and EVA-02 achieved the highest stability by reconstructing CLIP features instead of pixels. This confirms that alignment targets determine geometry. Hierarchical transformers (Swin, PVT, MaxViT, CoAtNet, PoolFormer) significantly outperformed columnar architectures (ViT, DeiT) on CIFAR-10 ($p = 0.011$), CIFAR-100 ($p = 0.007$), and Flowers-102 ($p < 0.001$), with multi-scale processing acting as implicit geometric regularization. However, this advantage did not generalize to DTD, EuroSAT, or Oxford Pets.

Stability rankings generalize within but not across domains. CIFAR-10/100 rankings correlated at $\rho = 0.92$, confirming that stability is intrinsic rather than task-specific. However, Flowers-102 vs. Oxford Pets showed no correlation ($\rho = -0.03$), suggesting fine-grained domains require distinct stability properties. Consistent patterns emerged across datasets: CLIP, EVA, and Inception families maintained high stability, while DINOv2, DeiT3, and ViT families showed low stability (except DINOv2 on EuroSAT).

3.3 Geometric Stability Detects Representational Drift

Motivation. Reliable deployment requires detecting representation degradation before it manifests as task failure. We test whether geometric stability provides earlier and more reliable drift detection than similarity metrics, evaluating magnitude of change, predictive validity, and false alarm rates across post-training alignment and structured perturbations.

Experimental design. Experiment 1 (Post-Training Drift) compared representations from 23 base/instruct model pairs spanning 11 families (Qwen, Llama, SmolLM, SmolLM2, Mistral, StableLM, Gemma, TinyLlama, Pythia, BLOOM, Falcon) ranging from 0.14B to 7B parameters, yielding 92 observations across four prompt types. Experiment 2 (Structured Perturbations) applied Gaussian noise ($\sigma \in [0.01, 0.5]$), quantization (INT8, INT4), and LoRA modifications to 16 causal LMs to characterize metric response curves. Experiment 3 (Canary Validation) injected progressive noise into 26 sentence embedding models, tracking geometric drift and SST-2 accuracy across 51 noise levels (1,326 observations). Experiment 4 (Extended Canary) replicated Experiment 3 on 15 causal LMs with Gaussian noise, quantization, and LoRA perturbations, specifically examining false alarm rates in the stable regime (Appendix 10).

Shesha detects greater geometric change in real-world alignment. In post-training shifts, Shesha measured **nearly 2× greater geometric change** than CKA (25.1% vs 12.8%, ratio: 1.96×). This held across prompt types: factual (2.37×), descriptive (2.28×), and conversational (1.82×). Sensitivity varied by family: Llama showed the largest discrepancy (**5.23**×), while BLOOM (1.14×) and Falcon (1.32×) showed near-parity. Procrustes detected moderate drift (15.0%, 1.17× vs CKA), but as we show below, this apparent sensitivity comes with significant drawbacks.

All metrics predict functional degradation. In the canary validation, all three metrics demonstrated strong predictive validity: CKA ($\rho = 0.937$), Procrustes ($\rho = 0.935$), and Shesha ($\rho = 0.927$)

correlated nearly identically with accuracy degradation. This replicates to causal LMs: Shesha ($\rho = 0.915$), CKA ($\rho = 0.912$), and Procrustes ($\rho = 0.903$). The consistently high correlations confirm that geometric drift reliably predicts functional degradation regardless of metric choice.

Shesha provides earlier warning than CKA. Despite equivalent validity, metrics differed in *when* they detected drift. Using a 5% threshold, Shesha provided **earlier warning in 73% of models** (19/26), CKA detected earlier in 0%, with 27% tied. When metrics diverged, Shesha detected first **100% of the time**. Mean detection thresholds confirmed this: Shesha triggered at $\sigma = 0.123$ vs CKA at $\sigma = 0.136$.

To validate detection performance independent of threshold selection, we performed an ROC analysis on the LoRA perturbation benchmark, where structural drift is subtle and detection is most challenging. Shesha achieves the highest performance (AUC $= 0.990$) compared to Procrustes (AUC $= 0.988$) and CKA (AUC $= 0.987$). Critically, at a strict 5% False Positive Rate, Shesha maintains a sensitivity of 90.2%, whereas Procrustes drops to 85.4%. This confirms that Shesha's earlier detection reflects genuine signal rather than threshold artifacts or noise susceptibility.

Procrustes triggers excessive false alarms. While Procrustes shows high predictive validity, it suffers from critical oversensitivity. In the stable regime (accuracy drop $< 1\%$), Procrustes triggered false alarms in **44% of cases** compared to only 7.3% for Shesha and CKA, a 6$\times$ difference. At minimal perturbation where functional performance is unchanged, Procrustes reported 1.50% drift versus 0.04% for Shesha, a **37$\times$ inflation**.

This oversensitivity arises from spectral properties rather than rigid transformation detection. While Procrustes is invariant to rotations, reflections, and global scaling by construction, it minimizes the Frobenius residual $\|XR - Y\|_F$ across the full spectrum. In high-dimensional representations, small perturbations accumulate in the spectral tail as noise that Procrustes attempts to align but cannot, inflating the distance score. By contrast, CKA effectively upweights dominant eigenvalues and discounts spectral tail variation, while Shesha's rank-based correlation is robust to low-magnitude perturbations that preserve relative distance ordering despite metric fluctuations. The ROC analysis reinforces this conclusion: Procrustes's spectral sensitivity leads to higher false alarms even under calibrated thresholds, with sensitivity dropping 4.8 percentage points below Shesha at the 5% FPR operating point.

Shesha achieves optimal balance. These experiments establish Shesha as the optimal drift metric: it detects real representational changes earlier than CKA, maintains equivalent predictive validity, and avoids the false alarm problem of Procrustes. The threshold-independent ROC analysis confirms this advantage is not an artifact of arbitrary detection criteria. For production monitoring, we recommend Shesha as the primary metric, with CKA providing stable confirmation and Procrustes serving only as an early (but noisy) alert.

3.4 CRISPR: Tracking Perturbation Magnitudes

Motivation. Single-cell perturbations induce high-dimensional state transitions that are often obscured by stochastic noise. We test whether geometric stability captures the structural coherence of these biological shifts by evaluating its correlation with transcriptional effect magnitude across diverse CRISPR screens.

Experimental design. We tested whether geometric stability captures systematic structure in single-cell perturbation responses using four large-scale CRISPR transcriptional datasets (Norman et al., 2019; Adamson et al., 2016; Dixit et al., 2016; Papalexi et al., 2021), covering activation, repression, combinatorial perturbations, and a high-variance pooled screen. For each perturbation, we measured the control-to-perturbation effect magnitude, the Shesha stability score, and two controls (sample size and intrinsic variance), embedding all cells into a consistent 50-dimensional PCA embedding computed per dataset (Appendix 12).

Strong and monotonic structure. Across all datasets, we found **uniformly positive magnitude-stability correlations**, ranging from $\rho = 0.746$ [0.640, 0.827] in high-variance screens to $\rho = 0.963$ [0.946, 0.974] in cleaner activation settings (Table 62). These effects were consistent across 811 perturbations of varying strength, indicating that larger transcriptional shifts consistently yield more

coherent geometric responses. Notably, combinatorial perturbations exhibited systematically higher stability than single-gene knockdowns (Dixit et al., 2016). Combinatorial guides achieved $\bar{S} = 0.55$ vs. $\bar{S} = 0.15$ for single-gene perturbations ($p < 10^{-9}$), with the magnitude-stability relationship holding within both categories ($\rho = 0.94$ and $\rho = 0.65$, respectively).

We note that Shesha is mathematically related to signal-to-noise ratio (SNR) estimation. However, in perturbation screens, this "reliability" is not a nuisance variable to be normalized away, but a primary signal of **regulatory coherence**. Perturbations with high Shesha scores induce reproducible state-space trajectories, whereas low-stability perturbations produce stochastic effects even when the mean effect size is non-zero.

Independence from confounds. Stability is effectively independent of sample size ($\rho \approx 0$) and is only weakly related to intrinsic variance ($\rho \approx -0.27$). A mixed-effects model confirmed that magnitude is the dominant predictor ($\beta = 0.123$ [0.166, 0.131], $p < 10^{-200}$), with the sample size contributing minimally ($\beta = -0.031$ [-0.039, -0.024], $\sim 4\times$ smaller effect). Partial correlation analysis controlling for SNR revealed dataset-specific heterogeneity (Norman: $\rho_{\text{partial}} = -0.859$; Dixit: $\rho_{\text{partial}} = 0.627$). The pooled estimate confirmed signal beyond confounding ($\rho_{\text{partial}} = 0.374$ [0.300, 0.440], $p < 10^{-27}$). Discordant cases, high magnitude but low stability, correspond to known pleiotropic regulators (e.g., CEBPA combinations), while high-stability but low-magnitude cases involve lineage-specific factors (e.g., KLF1), suggesting stability may index regulatory specificity.

Robustness across methods. To rule out the possibility that the results were artifacts of distance metric choice, we compared three approaches: Euclidean distance in PCA space, Mahalanobis-whitened distances, and k-nearest neighbor matched controls. All three resulted in strong correlations with tight bootstrap CIs: Euclidean ($\rho = 0.746$-0.953), Whitened ($\rho = 0.846$-0.976), and k-NN ($\rho = 0.912$-0.951). Ablations across PCA dimensions (10-100) and random seeds (15 per dataset) showed complete reproducibility ($r = 1.000$) and stable correlations across dimensionality choices ($\rho = 0.67$-0.96). The Adamson dataset ($n = 8$) showed appropriately wide CIs ([0.447, 1.000]), which reflects honest uncertainty with small samples.

Convergent evidence. Although the four datasets have differences in modality, noise level, and library design, all show the same monotonic pattern. After z-score calibration within datasets, the pooled correlation was $\rho = 0.915$ [0.897, 0.929], which confirms cross-dataset generalizability. This agreement suggests that geometric stability reflects a general property of cellular response manifolds. Stronger regulatory interventions produce more coherent state transitions, and Shesha isolates this structure while remaining insensitive to sample size and background variance.

3.5 Neuroscience: Tracking Drift and Behavior

Motivation. Biological neural representations drift over time, yet behavior remains stable. We test whether geometric stability captures the functional structure underlying this robustness better than simple temporal consistency.

Experimental design. We tested whether geometric stability captures functionally relevant structure in neural population dynamics using the Neuropixels decision-making dataset (Steinmetz et al., 2019). We analyzed 228 area-session observations across 26 sessions and 68 brain areas, computing condition-averaged population responses during the decision epoch (0-500 ms post-stimulus). Shesha was measured as split-half correlation of representational dissimilarity matrices. We additionally computed temporal drift-stability as cosine similarity between early and late trial centroids, validated against a permutation null model (Appendix 13).

Behavioral ground truth validation. Shesha predicted trial-by-trial neural-behavioral coupling ($\rho = 0.18$, 95% CI: $[0.05, 0.31]$, $p = 0.005$), indicating that regions with stable representational geometry exhibit a tighter relationship between neural activity and behavioral outcomes. This effect was specific to geometric stability: centroid drift showed no relationship ($\rho = 0.00$, $p = 0.98$), nor did WUC ($\rho = 0.09$, $p = 0.18$), suggesting that Shesha captures functionally relevant structures beyond simple temporal consistency or whitened similarity. There was no relationship between the change in accuracy and stability ($\rho = -0.05$, $p = 0.44$), which is consistent with prior work suggesting that drift at this timescale may be behaviorally silent (Rule et al., 2019).

Regional hierarchy and null model validation. Temporal drift-stability was significantly below permutation-based chance ($z = -44.7$; observed $S = 0.924$ vs. null $S = 0.995$). Shesha revealed a distinct regional hierarchy: Striatum (0.44), Motor (0.38), and Visual cortex (0.36) showed highest geometric stability, while Hippocampus exhibited the lowest (0.19). This pattern diverged from centroid drift, where sensory regions showed highest temporal stability (Thalamus: 0.95; Visual: 0.94) and Striatum showed the most drift (0.83). Drift accumulated gradually across sessions with no acceleration ($\Delta S = 0.001$, $p = 0.77$).

4 Discussion

The "Geometric Canary" in Safety Monitoring. Our findings demonstrate that standard similarity metrics greatly underestimate the structural impact of fine-tuning. In modern RLHF-aligned models (e.g., Llama), instruction tuning results in **5.23× greater geometric drift** in Shesha than in CKA (34.0% vs 6.5%). This indicates substantial manifold reorganization that rotation-invariant metrics miss. Notably, this increased sensitivity is achieved without losing reliability: Shesha, CKA, and Procrustes all predict functional degradation equally well. However, Procrustes triggers alarms in **44% of stable cases** versus 7.3% for Shesha, reflecting a 37× drift inflation at minimal perturbation. Meanwhile, Shesha provides **an early warning in 73% of models** (as opposed to 0% for CKA). Thus, Shesha serves as the **optimal primary metric**: sensitive enough to detect latent geometric fracturing before catastrophic failure, yet specific enough to avoid false alarms that undermine operator confidence.

The Stability-Transferability Trade-off in Vision. Our vision architecture analysis reveals a fundamental tension between transfer learning and geometric stability, modulated by task complexity. On simple tasks (CIFAR-10) or uniform imagery (EuroSAT), most models perform adequately on both axes. However, as class count increases or fine-grained discrimination is required, specialization becomes necessary. DINOv2 exemplifies this trade-off: it achieves state-of-the-art transferability (LogME rank 1/29 on 4/6 datasets) while ranking *last or near-last* in geometric stability on all but EuroSAT. This "geometric tax" suggests that optimizing for rich, adaptable features collapses the manifold structure required for predictable behavior. For practitioners, this implies distinct model selection criteria: high-LogME models (DINOv2, ResNets) for transfer learning pipelines; high-Shesha-FS models (CLIP, EVA-02, Inception) for zero-shot or safety-critical deployment where behavioral consistency matters more than adaptability.

Architectural and Training Determinants of Stability. Two factors consistently predict geometric stability across visual domains. First, **contrastive alignment**: CLIP models outperformed self-supervised alternatives on 4/6 datasets ($p < 0.05$), and EVA-02 achieved maximal stability by reconstructing CLIP features rather than pixels, confirming that alignment targets, not training mechanisms, determine geometry. Second, **hierarchical architecture**: Swin, PVT, and CoAtNet demonstrated superior performance to columnar architectures (ViT, DeiT) on 3/6 datasets (CIFAR-10, CIFAR-100, Flowers-102), which indicates that multi-scale processing acts as implicit geometric regularization, though this advantage did not generalize to texture, satellite, or fine-grained animal domains. These findings can assist practitioners in selecting architectures. When stability is required, prefer contrastively-aligned models with hierarchical feature pyramids.

The Stability-Alignment Trade-off. Our results across steering, transfer learning, and vision reveal a fundamental boundary condition for geometric stability. In domains where the task is intrinsic to the data manifold, such as drift detection and biological fitness, unsupervised stability is highly predictive. However, in high-level semantic tasks, intrinsic stability is insufficient. We observed an important dissociation in **Real-World Steering** (Section 3.1): supervised Shesha achieved near-perfect prediction ($\rho \geq 0.96$), proving that stability drives control, whereas unsupervised Shesha failed entirely ($\rho \leq 0.35$, n.s.); in **Vision** (Section 3.2), stability and transferability were empirically independent ($\rho \approx 0$) or negatively correlated on complex tasks ($\rho = -0.19$ to -0.42). This dissociates "rigidity" from "utility": a model can be geometrically rigid yet semantically misaligned (DINOv2: top transfer, bottom stability), necessitating either **supervised** stability metrics or explicit consideration of task alignment when predicting downstream control.

Stability in Adaptive Systems. The predictive power of stability appears to be linked to a system's ability to self-organize. In adaptive biological systems, stability tracks function: in **CRISPR** screens, it predicts perturbation coherence ($\rho = 0.75$-0.95); in **Neuroscience**, it predicts trial-by-trial behavioral coupling ($\rho = 0.18$, $p = 0.005$), distinguishing stable coding regions (Striatum) from plastic ones (Hippocampus) (Dhawale et al., 2017; Ziv et al., 2013; McClelland et al., 1995). This pattern suggests that stability metrics are most diagnostic in systems where the manifold must *learn* robustness through adaptation, such as biological evolution, neural plasticity, or gradient descent. The universality observed in natural and artificial adaptive systems confirms that geometric stability is not an artifact of machine learning but rather a fundamental property of robust information representation.

Limitations. Our primary metric operates on a global representational structure (RDMs), potentially overlooking fine-grained token-level dynamics. Transfer correlations are partially driven by architectural families (e.g., BGE/GTE vs. Jina), suggesting that metrics are most comparable within model classes. The vision analysis, while comprehensive, found only one model (weakly-supervised ResNeXt-101) achieving top-20 status across all metrics on 4/6 datasets, underscoring the scarcity of off-the-shelf models that excel in both stability and transferability for practitioners seeking unified solutions. Finally, the computation of stability requires multiple forward passes (perturbation/resampling), which increases inference costs compared to static similarity metrics, a necessary trade-off for quantifying reliability.

Acknowledgments

We thank Padma K. and Annapoorna Raju for generously supporting the computational resources used in this work. We thank the many institutions and individuals whose open-source datasets, frameworks, and models were used in our work. The authors acknowledge the use of large language models (specifically the GPT, Claude, and Gemini families) to assist with (limited) code factoring, debugging, and text polishing. All hypotheses, experimental designs, analyses, and interpretations were independently formulated and verified by the authors, and the authors assume full responsibility for all content and claims in this work.

Code Availability

The full code necessary to reproduce all experiments, benchmarks, and analysis described in this paper is publicly available at `https://github.com/prashantcraju/geometric-stability`. We have also released an open-source python library `https://pypi.org/project/shesha-geometry` (`pip install shesha-geometry`) with accompanying tutorials at `https://github.com/prashantcraju/shesha/tree/main?tab=readme-ov-file#tutorials`.

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C., Soatto, S., and Perona, P. (2019). Task2vec: Task embedding for meta-learning. *arXiv*.

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A., and Weissman, J. S. (2016). A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21.

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. (2020). Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Annual Meeting of the Association for Computational Linguistics*.

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv*.

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

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*.

Avellino, R. and Delwel, R. (2017). Expression and regulation of c/ebpα in normal myelopoiesis and in malignant transformation. *Blood*, 129(15):2083–2091.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukošiūtė, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., Dassarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T. J., Hume, T., Bowman, S., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T. B., and Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv*.

Bansal, Y., Nakkiran, P., and Barak, B. (2021). Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*.

Bao, Y., Li, Y., Huang, S.-L., Zhang, L., Zheng, L., Zamir, A., and Guibas, L. J. (2019). An information-theoretic approach to transferability in task transfer learning. *2019 IEEE International Conference on Image Processing*, pages 2309–2313.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., Kandasaamy, S., Lock, A., Luciani, A., Lugaric, M., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Mishra, A., Moulang, K., Nightingale, A., Pundir, S., Qi, G., Raj, S., Raposo, P., Rice, D. L., Saidi, R., Santos, R., Speretta, E., Stephenson, J., Totoo, P., Turner, E., Tyagi, N., Vasudev, P., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A. J., Aimo, L., Argoud-Puy, G., Auchincloss, A. H., Axelsen, K. B., Bansal, P., Baratin, D., Batista Neto, T. M., Blatter, M.-C., Bolleman, J. T., Boutet, E., Breuza, L., Gil, B. C., Casals-Casas, C., Echioukh, K. C., Coudert, E., Cuche, B., de Castro, E., Estreicher, A., Famiglietti, M. L., Feuermann, M., Gasteiger, E., Gaudet, P., Gehant, S., Gerritsen, V., Gos, A., Gruaz, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Kerhornou, A., Le Mercier, P., Lieberherr, D., Masson, P., Morgat, A., Muthukrishnan, V., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Poux, S., Pozzato, M., Pruess, M., Redaschi, N., Rivoire, C., Sigrist, C. J. A., Sonesson, K., Sundaram, S., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Ross, K., Vinayaka, C. R., Wang, Q., Wang, Y., and Zhang, J. (2022). Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, 1st edition.

Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Member of the Taylor and Francis Group, 2nd edition.

Cohen, U., Chung, S., Lee, D. D., and Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1).

Daniélou, A. (1964). *Hindu Polytheism*. Bollingen Series. Princeton University Press. Later republished as 'The Myths and Gods of India'.

Dasgupta, S. and Gupta, A. (2002). An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

Dhawale, A. K., Poddar, R., Wolff, S. B., Normand, V. A., Kopelowitz, E., and Ölveczky, B. P. (2017). Automated long-term recording and analysis of neural activity in behaving animals. *eLife*, 6.

Diedrichsen, J., Berlot, E., Mur, M., Schütt, H. H., Shahbazi, M., and Kriegeskorte, N. (2021). Comparing representational geometries using whitened unbiased-distance-matrix similarity. *Neurons, Behavior, Data analysis, and Theory*, 5(3).

Diedrichsen, J. and Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508.

Dimmitt, C. and van Buitenen, J. A. B. (1978). *Classical Hindu Mythology: A Reader in the Sanskrit Puranas*. Temple University Press, Philadelphia, PA.

Ding, F., Denain, J.-S., and Steinhardt, J. (2021). Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., and Regev, A. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17.

Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N., and Harvey, C. D. (2017). Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell*, 170(5):986–999.e16.

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1).

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Friedman, A. D. (2007). C/ebpα induces pu.1 and interacts with ap-1 and nf-κb to regulate myeloid development. *Blood Cells, Molecules, and Diseases*, 39(3):340–343.

Friedman, A. D. (2015). C/ebpα in normal and malignant myelopoiesis. *International Journal of Hematology*, 101(4):330–341.

Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A., and Miller, L. E. (2020). Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, 23(2):260–270.

Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. (2024). Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236. PMLR.

Helber, P., Bischke, B., Dengel, A., and Borth, D. (2018). 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.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. (2024). Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*.

Heumos, L., Ji, Y., May, L., Green, T. D., Peidli, S., Zhang, X., Wu, X., Ostner, J., Schumacher, A., Hrovatin, K., Müller, M., Chong, F., Sturm, G., Tejada, A., Dann, E., Dong, M., Pinto, G., Bahrami, M., Gold, I., Rybakov, S., Namsaraeva, A., Moinfar, A. A., Zheng, Z., Roellin, E., Mekki, I., Sander, C., Lotfollahi, M., Schiller, H. B., and Theis, F. J. (2025). Pertpy: an end-to-end framework for perturbation analysis. *Nature Methods*.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. *Conference on Modern Analysis and Probability*, page 189–206.

Khayatan, P., Shukor, M., Parekh, J., Dapogny, A., and Cord, M. (2025). Analyzing finetuning representation shift for multimodal llms steering. *arXiv*.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5(11):826–837.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019a). Similarity of neural network representations revisited. In *International Conference on Machine Learning*.

Kornblith, S., Shlens, J., and Le, Q. V. (2019b). Do better imagenet models transfer better? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report.

Kumar, A., Raghunathan, A., Jones, R. M., Ma, T., and Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Li, M. Z., Agrawal, K. K., Ghosh, A., Teru, K. K., Santoro, A., Lajoie, G., and Richards, B. A. (2025a). Tracing the representation geometry of language models from pretraining to post-training. In *ICML Workshop on High-dimensional Learning Dynamics*.

Li, Y., Zhang, J., Feng, S., Yuan, P., Wang, X., Shi, J., Zhang, Y., Tan, C., Pan, B., Hu, Y., and Li, K. (2025b). Revisiting self-consistency from dynamic distributional alignment perspective on answer aggregation. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Lin, Y., Tan, L., Lin, H., Xiong, W., Zheng, Z., Pi, R., Zhang, J., Diao, S., Wang, H., Dong, H., Zhao, H., Yao, Y., and Zhang, T. (2024). Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., Bossan, B., and Tietz, M. (2022). PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Masarotto, V., Panaretos, V. M., and Zemel, Y. (2018). Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhya A*, 81(1):172–213.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457.

Meng, K., Bau, D., Andonian, A. J., and Belinkov, Y. (2022). Locating and editing factual associations in GPT. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Miller, I. J. and Bieker, J. J. (1993). A novel, erythroid cell-specific murine transcription factor that binds to the caccc element and is related to the krüppel family of nuclear proteins. *Molecular and Cellular Biology*, 13(5):2776–2786.

Moalla, S., Miele, A., Pyatko, D., Pascanu, R., and Gulcehre, C. (2024). No representation, no trust: Connecting representation, collapse, and trust issues in PPO. In *Advances in Neural Information Processing Systems*.

Morcos, A., Raghu, M., and Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*.

Naitzat, G., Zhitnikov, A., and Lim, L.-H. (2020). Topology of deep neural networks. *Journal of Machine Learning Research*, 21(1).

Nguyen, C. V., Hassner, T., Seeger, M., and Archambeau, C. (2020). Leep: a new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*.

Nguyen, T., Raghu, M., and Kornblith, S. (2021). 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*.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553.

Nilsback, M.-E. and Zisserman, A. (2008). Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, Y. A., Xu, A., Jost, M., Gilbert, L. A., and Weissman, J. S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793.

O'Mahony, L., Grinsztajn, L., Schoelkopf, H., and Biderman, S. (2024). Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). 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.

Pándy, M., Agostinelli, A., Uijlings, J., Ferrari, V., and Mensink, T. (2022). Transferability estimation using bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Papalexi, E., Mimitou, E. P., Butler, A. W., Foster, S., Bracken, B., Mauck, W. M., Wessels, H.-H., Hao, Y., Yeung, B. Z., Smibert, P., and Satija, R. (2021). Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature Genetics*, 53(3):322–331.

Park, K., Choe, Y. J., Jiang, Y., and Veitch, V. (2025). The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*.

Park, K., Choe, Y. J., and Veitch, V. (2023). The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*.

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

Pilon, A. M., Arcasoy, M. O., Dressman, H. K., Vayda, S. E., Maksimova, Y. D., Sangerman, J. I., Gallagher, P. G., and Bodine, D. M. (2008). Failure of terminal erythroid differentiation in eklf-deficient mice is associated with cell cycle perturbation and reduced expression of e2f2. *Molecular and Cellular Biology*, 28(24):7394–7401.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*.

Ratzon, A., Derdikman, D., and Barak, O. (2024). Representational drift as a result of implicit regularization. *eLife*, 12.

Recanatesi, S., Farrell, M., Lajoie, G., Deneve, S., Rigotti, M., and Shea-Brown, E. (2021). Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature Communications*, 12(1).

Rohlf, F. J. and Slice, D. (1990). Extensions of the procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, 39(1):40.

Rosen, E. D., Hsu, C.-H., Wang, X., Sakai, S., Freeman, M. W., Gonzalez, F. J., and Spiegelman, B. M. (2002). C/ebpα induces adipogenesis through pparγ: a unified pathway. *Genes & Development*, 16(1):22–26.

Rule, M. E., O'Leary, T., and Harvey, C. D. (2019). Causes and consequences of representational drift. *Current Opinion in Neurobiology*, 58:141–147.

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

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

Schütt, H. H., Kipnis, A. D., Diedrichsen, J., and Kriegeskorte, N. (2023). Statistical inference on representational geometries. *eLife*, 12.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., and Zhang, F. (2014). Genome-scale crispr-cas9 knockout screening in human cells. *Science*, 343(6166):84–87.

Siatecka, M. and Bieker, J. J. (2011). The multifunctional role of eklf/klf1 during erythropoiesis. *Blood*, 118(8):2044–2054.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Empirical Methods in Natural Language Processing*.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47):1393–1434.

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., and Harris, K. D. (2019). Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273.

Subramani, N., Suresh, N., and Peters, M. (2022). Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581.

Tallack, M. R. and Perkins, A. C. (2010). Klf1 directly coordinates almost all aspects of terminal erythroid differentiation. *IUBMB Life*, 62(12):886–890.

Tallack, M. R., Whitington, T., Shan Yuen, W., Wainwright, E. N., Keys, J. R., Gardiner, B. B., Nourbakhsh, E., Cloonan, N., Grimmond, S. M., Bailey, T. L., and Perkins, A. C. (2010). A global role for klf1 in erythropoiesis revealed by chip-seq in primary erythroid cells. *Genome Research*, 20(8):1052–1063.

Tan, Y., Li, Y., and Huang, S.-L. (2021). Otce: A transferability metric for cross-domain cross-task representations. *arXiv*.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Tran, A. T., Nguyen, C. V., and Hassner, T. (2019). Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. (2023). Activation addition: Steering language models without optimization. *arXiv*.

Vogel, J. P. (1995). *Indian Serpent-Lore: Or, The Nāgas in Hindu Legend and Art*. Asian Educational Services, New Delhi. Reprint of the 1926 edition.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200.

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic screens in human cells using the crispr-cas9 system. *Science*, 343(6166):80–84.

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

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1).

You, K., Liu, Y., Wang, J., and Long, M. (2021). Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*.

You, K., Liu, Y., Zhang, Z., Wang, J., Jordan, M. I., and Long, M. (2022). Ranking and tuning pre-trained models: A new paradigm for exploiting model hubs. *Journal of Machine Learning Research*, 23(1).

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., and Houlsby, N. (2019). A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv*.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *arXiv*.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1).

Ziv, Y., Burns, L. D., Cocker, E. D., Hamel, E. O., Ghosh, K. K., Kitch, L. J., Gamal, A. E., and Schnitzer, M. J. (2013). Long-term dynamics of ca1 hippocampal place codes. *Nature Neuroscience*, 16(3):264–266.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, Z., and Hendrycks, D. (2023). Representation engineering: A top-down approach to ai transparency. *arXiv*.

Appendix

Appendix Outline

- Appendix 5: Related Work
- Appendix 6: The Shesha Metric
  - Appendix 6.1: Definition and Variants
  - Appendix 6.2: Metric Validation Tests
- Appendix 7: Evaluating Distinction of Stabiltiy and Similarity: Extended Methods and Results
- Appendix 8: Steering: Extended Methods and Results
- Appendix 9: Visual Perception Architecture: Extended Methods and Results
- Appendix 10: Representational Drift Detection: Extended Methods and Results
- Appendix 11: Transfer Learning and the Limits of Unsupervised Stability
- Appendix 12: CRISPR Perturbation Magnitudes: Extended Methods and Results
- Appendix 13: Neuroscience Drift and Behavior: Extended Methods and Results
- Appendix 14: Broader Impact

5 Related Work

Representation Similarity Metrics. Several methods exist for comparing the representations of neural networks. These include Centered Kernel Alignment (CKA) (Kornblith et al., 2019a), Singular Vector Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017), Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008), and Procrustes distance (Schönemann, 1966; Rohlf and Slice, 1990; Masarotto et al., 2018; Dryden and Mardia, 1998). These methods are used to measure *external* similarity: whether two representations encode similar pairwise structures or occupy aligned subspaces. Shesha addresses a fundamentally different question: *internal* stability. Rather than asking whether two representations are similar, Shesha asks whether a single representation's geometry is internally consistent. In other words, it determines whether the representation would be preserved under feature subsampling, data perturbation, or domain shift. This distinction matters because similarity and stability are empirically dissociable: representations can be highly similar yet geometrically fragile, or stable yet dissimilar (Section 2). Shesha thus provides a complementary diagnostic axis for representation analysis.

Geometric Properties of Representations. Prior work has studied the geometric properties of representations, including intrinsic dimensionality (Ansuini et al., 2019), linear decodability (Cohen et al., 2020; Recanatesi et al., 2021), hierarchical organization (Park et al., 2025), curvature, and topological structure (Naitzat et al., 2020; Jacot et al., 2018). Although many important findings related to structural attributes have been discovered, the focus has been on *static* properties of the representations. These approaches do not consider *stability*: whether geometric features are consistent across data splits, sampling noise, or distribution shift.

Representational Drift and Robustness. Representational drift has been studied extensively in machine learning, particularly during fine-tuning (Kumar et al., 2022; Aghajanyan et al., 2020; Khayatan et al., 2025), under adversarial perturbations (Ilyas et al., 2019), and across random seeds (Nguyen et al., 2021; Ratzon et al., 2024). While techniques like model stitching (Bansal et al., 2021) and cross-architecture comparisons (Nguyen et al., 2021) assess feature compatibility in the presence of such drift, they lack a unified approach to quantify the consistent stability of the underlying geometry. Additionally, work on RLHF and preference optimization has mainly focused on behavioral drift (Ouyang et al., 2022; Bai et al., 2022). Recent studies have begun to discover evidence of degradation of geometry during these stages (Li et al., 2025a). Shesha bridges this gap by directly estimating the reliability of a representation's geometry against these forms of drift.

Representation Steering and Control. Recent work has demonstrated that language models can be controlled by intervening directly on their internal activations, a technique often termed "activation engineering" or "steering" (Turner et al., 2023; Subramani et al., 2022; Hernandez et al., 2024). Approaches such as Representation Engineering (Zou et al., 2023) and causal interventions (Meng et al., 2022; Geiger et al., 2024) rely on the Linear Representation Hypothesis (Park et al., 2023, 2025), assuming that concepts are encoded as stable linear directions within the latent space. While these methods exploit geometric structure to generate behaviors, they typically do not quantify the *stability* of the geometry itself. If the representational structure is brittle or subject to drift, steering vectors may become unreliable across contexts or model updates. Shesha addresses this limitation by providing a metric to assess geometric stability, which allows it to serve as a diagnostic for the feasibility and robustness of steering interventions.

Alignment and Capability Tradeoffs. RLHF and related alignment methods are known to induce an "alignment tax" on broader capabilities (Ouyang et al., 2022; Bai et al., 2022; Lin et al., 2024). Studies have identified representation collapse under preference optimization (Moalla et al., 2024) and mode collapse in RLHF (O'Mahony et al., 2024) through behavioral metrics. Geometry remains less explored, though emerging evidence shows distributional and feature-space drift (Li et al., 2025b). Our analysis provides the first systematic quantification of post-alignment geometric drift (Section 3.3).

Perturbation Analysis in Systems Biology. Perturbation screens using tools like CRISPR-Cas9 (Shalem et al., 2014; Wang et al., 2014) are the gold standard in systems biology for inferring network topology and functional robustness. By systematically knocking out or modulating genes, researchers identify which components are essential and which are redundant, revealing the system's

underlying stability (Kitano, 2004; Barabási and Oltvai, 2004). This approach relies on the principle that a robust system preserves its phenotype despite local disruptions. Shesha adopts a similar epistemological stance for deep learning: rather than evaluating a model solely on its outputs, we employ systematic perturbations analogous to genetic screens to test the resilience of its internal representational geometry.

Neuroscience and Latent Dynamics. Long-term recording studies in biological neural networks have revealed that the specific neurons encoding a sensory stimulus change over days or weeks, even while behavioral performance remains stable, a phenomenon termed *representational drift* (Driscoll et al., 2017; Rule et al., 2019). Empirical evidence indicates that the brain maintains consistent behavior despite this drift by preserving the low-dimensional latent dynamics of population activity, even as the contribution of individual units fluctuates (Gallego et al., 2020). Shesha draws inspiration from these biological principles, shifting evaluation from the parametric stability of individual features to the geometric stability of the representational manifold.

Statistical Evaluation and Bootstrap Methods. Bootstrap resampling is a standard statistical tool for estimating uncertainty (Efron and Tibshirani, 1994) and has recently been applied to RDM variability in neuroscience (Schütt et al., 2023; Schütt, 2025). In machine learning, bootstrap methods are used for uncertainty estimation in model selection (Efron, 1979; Bishop, 2006) and distribution shift (Lakshminarayanan et al., 2017). These techniques have not, however, been applied to quantify the *geometric* stability of representations. Shesha formalizes this application, measuring RDM self-consistency under internal perturbations (feature splits, sample splits, or bootstrap resampling) to create a principled, task-agnostic metric of stability across domains.

Transfer Learning and Architecture Evaluation. A core goal of representation learning is to identify architectures that transfer well to downstream tasks (Kornblith et al., 2019b). Embedding-based transferability metrics such as LogME (You et al., 2021, 2022), H-Score (Bao et al., 2019), Task2Vec (Achille et al., 2019), and OTCE (Tan et al., 2021) rank models by feature richness or predict task similarity, while linear probing remains a standard diagnostic tool (Alain and Bengio, 2016). Large-scale benchmarks like VTAB (Zhai et al., 2019) and comparisons of self-supervised objectives (Caron et al., 2021; Radford et al., 2021) further prioritize adaptability. However, these evaluations overlook the *rigidity* of the representational manifold, and many require target-domain labels for supervised evaluation. Our empirical results (Section 9) reveal that transferability and stability can dissociate: architectures optimized for maximal transferability may paradoxically exhibit the lowest geometric stability, incurring a "geometric tax" that standard benchmarks fail to capture. To address this, Shesha offers a task-agnostic metric of stability that operates without downstream labels, predicting few-shot transfer performance purely from geometric resilience (Appendix 11).

6 The Shesha Metric

6.1 Definition and Variants

Shesha quantifies the geometric stability of a representation by measuring the self-consistency of its pairwise distance structure under controlled internal perturbations. The core insight is that a geometrically stable representation should yield similar Representational Dissimilarity Matrices (RDMs) when computed from different "views" of the same underlying structure.

General formulation. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote a matrix of n samples with d-dimensional representations. An RDM $\mathbf{D} \in \mathbb{R}^{n \times n}$ captures pairwise dissimilarities, typically computed as:

$$D_{ij} = 1 - \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$$

for cosine distance, or using correlation distance for centered representations.

Shesha operates by constructing two RDMs from complementary views of $\mathbf{X}$ and measuring their agreement:

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

where ρ_s denotes Spearman's rank correlation and $\text{vec}(\cdot)$ extracts the upper triangular elements. The choice of how to construct views $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ defines distinct Shesha variants, each probing different aspects of geometric stability.

6.1.1 Variants

We organize Shesha variants along two axes: (1) the *source of variation* used to construct complementary views, and (2) whether the metric incorporates *label information*. This taxonomy reflects distinct notions of stability relevant to different analytical contexts.

Feature-Split Shesha (Unsupervised). This variant measures *internal geometric consistency* by partitioning feature dimensions:

$$\text{Shesha}_{\text{FS}}(\mathbf{X}) = \frac{1}{K} \sum_{k=1}^{K} \rho_s\big(\text{vec}(\mathbf{D}_{\mathcal{F}_k^{(1)}}), \text{vec}(\mathbf{D}_{\mathcal{F}_k^{(2)}})\big)$$

where $\mathcal{F}_k^{(1)}, \mathcal{F}_k^{(2)} \subset \{1, \ldots, d\}$ are disjoint random partitions of feature indices at iteration k, and $\mathbf{D}_{\mathcal{F}}$ denotes the RDM computed using only features in $\mathcal{F}$. Averaging over K random splits (typically $K = 30$) provides a stable estimate.

Feature-Split Shesha captures whether geometric structure is *distributed* across dimensions rather than concentrated in a few features. High values indicate that arbitrary subsets of features encode consistent relational information, a signature of robust, redundant representations. This variant requires no labels and is well-suited for intrinsic analyses including representational drift detection, biological perturbation studies, and scientific discovery applications where ground-truth labels may be unavailable or ill-defined.

Sample-Split Shesha (Unsupervised). This variant measures *robustness to input variation* by partitioning data points:

$$\text{Shesha}_{\text{SS}}(\mathbf{X}) = \frac{1}{K} \sum_{k=1}^{K} \rho_s\big(\text{vec}(\mathbf{D}_{\mathcal{S}_k^{(1)}}), \text{vec}(\mathbf{D}_{\mathcal{S}_k^{(2)}})\big)$$

where $\mathcal{S}_k^{(1)}, \mathcal{S}_k^{(2)} \subset \{1, \ldots, n\}$ are disjoint random partitions of sample indices. Because the two RDMs are computed over different items, correlation is performed on the subset of pairs where both samples appear in both partitions, or alternatively via anchor-based approaches (see below).

Sample-Split Shesha assesses whether the representation's distance structure generalizes across different subsets of the data distribution. Low values may indicate overfitting to spurious patterns or sensitivity to sampling noise.

Supervised Variants. When class labels $\mathbf{y} \in \{1, \ldots, C\}^n$ are available, Shesha can be adapted to measure *task-aligned* geometric stability:

- **Label-Conditioned Sample-Split**: Partitions are constructed to be class-balanced, and RDMs are computed on class centroids rather than individual samples. This variant, related to Whitened Unbiased Cosine (WUC) metrics in RSA (Diedrichsen et al., 2021), measures whether class-level geometry is stable across data splits.
- **Supervised RDM Alignment**: Directly correlates the model's RDM with an ideal RDM derived from labels:

$$\text{Shesha}_{\text{sup}}(\mathbf{X}, \mathbf{y}) = \rho_s\big(\text{vec}(\mathbf{D}_\mathbf{X}), \text{vec}(\mathbf{D}_\mathbf{y})\big)$$

  where $\mathbf{D}_\mathbf{y}$ encodes label dissimilarity (e.g., Hamming distance on one-hot encodings). This variant quantifies how well the representation's geometry aligns with task structure.
- **Variance Ratio**: A computationally efficient approximation measuring the ratio of between-class to total variance:

$$\text{Shesha}_{\text{var}}(\mathbf{X}, \mathbf{y}) = \frac{\sum_{c=1}^{C} n_c \|\boldsymbol{\mu}_c - \boldsymbol{\mu}\|^2}{\sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}\|^2}$$

  where $\boldsymbol{\mu}_c$ is the centroid of class c and $\boldsymbol{\mu}$ is the global mean.
- **Class Separation Ratio**: Measures the ratio of between-class to within-class distances:

$$\text{Shesha}_{\text{sep}}(\mathbf{X}, \mathbf{y}) = \frac{\bar{d}_{\text{between}}}{\bar{d}_{\text{within}}}$$

  where $\bar{d}_{\text{between}}$ is the mean pairwise distance between samples of different classes and $\bar{d}_{\text{within}}$ is the mean pairwise distance between samples of the same class. To reduce computational cost and improve stability, we estimate this ratio via bootstrap subsampling ($B = 50$ iterations at 50% subsampling). This variant is related to Fisher's discriminant ratio but operates in distance space rather than variance space
- **LDA Subspace Stability**: Measures the consistency of the linear discriminant direction under resampling:

$$\text{Shesha}_{\text{LDA}}(\mathbf{X}, \mathbf{y}) = \frac{1}{B} \sum_{b=1}^{B} \left| \hat{\mathbf{w}}^\top \hat{\mathbf{w}}^{(b)} \right|$$

  where $\hat{\mathbf{w}}$ is the unit-normalized LDA discriminant direction fitted on the full dataset and $\hat{\mathbf{w}}^{(b)}$ is the direction fitted on bootstrap subsample b. The absolute value accounts for sign ambiguity in the discriminant axis. High values indicate that the optimal linear decision boundary is robust to sampling variation. Low values suggest the discriminant subspace is unstable, potentially indicating overfitting to source domain structure.

Supervised variants are appropriate for transfer learning prediction, steering vector analysis, and other settings where alignment with semantic categories is the property of interest.

Domain-Specific Adaptations. For specialized applications, Shesha admits natural extensions:

- **Perturbation Stability (CRISPR)**: In single-cell genomics, stability is measured as the directional consistency of perturbation effects. Given control and perturbed cell populations, we compute the cosine similarity between individual cell shift vectors and the mean perturbation direction:

$$\text{Shesha}_{\text{pert}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \frac{(\mathbf{x}_i - \boldsymbol{\mu}_{\text{ctrl}})^\top \bar{\mathbf{d}}}{\|\mathbf{x}_i - \boldsymbol{\mu}_{\text{ctrl}}\| \|\bar{\mathbf{d}}\|}$$

  where $\bar{\mathbf{d}}$ is the mean shift direction. Variants include Mahalanobis-scaled (whitened) and k-NN matched control formulations.
- **Temporal Drift Detection**: For monitoring representation drift across model updates or fine-tuning, Shesha compares RDMs at different time points on a fixed held-out set, quantifying geometric divergence as $1 - \text{Shesha}(\mathbf{X}_{t_1}, \mathbf{X}_{t_2})$.

- **Trial-Split Shesha (Neuroscience)**: For neural recordings with repeated trials per condition, stability is measured by partitioning trials and comparing condition-level RDMs:

$$\text{Shesha}_{\text{trial}}(\mathbf{X}) = \rho_s\big(\text{vec}(\mathbf{D}^{\text{odd}}), \text{vec}(\mathbf{D}^{\text{even}})\big)$$

  where $\mathbf{D}^{\text{odd}}$ and $\mathbf{D}^{\text{even}}$ are RDMs computed on condition centroids derived from odd and even trials, respectively. This variant, standard in RSA (Kriegeskorte et al., 2008; Walther et al., 2016), is similar to the noise ceiling measurement (Nili et al., 2014). It measures whether the representational geometry of stimulus conditions is reliable across independent neural measurements.

- **Perturbation Stability (Latent Space)**: For models processing noisy inputs, stability is measured as the consistency of latent representations under stochastic input perturbations:

$$\text{Shesha}_{\text{pert}}(\mathbf{x}) = \frac{1}{1+\bar{d}}, \quad \bar{d} = \frac{1}{K}\sum_{k=1}^{K}\left\|\frac{f(\tilde{\mathbf{x}}_k)}{\|f(\tilde{\mathbf{x}}_k)\|} - \frac{f(\mathbf{x})}{\|f(\mathbf{x})\|}\right\|_2$$

  where $f(\cdot)$ denotes the encoder, $\tilde{\mathbf{x}}_k$ are stochastically perturbed versions of input $\mathbf{x}$, and $\bar{d}$ is the mean Euclidean displacement in the normalized latent space across K perturbations. High values indicate representations that are robust to input noise. When confounds correlate with stability (e.g., physical observables in scientific applications), residualized variants can be computed by regressing out confound effects.

Implementation Details. Across all variants, we use the following defaults unless otherwise specified: $K = 30$ random splits, cosine distance for RDM computation, Spearman correlation for RDM comparison, and subsampling to $n_{\max} = 1600$ samples when computational constraints require. All implementations use fixed random seeds for reproducibility. The choice of Spearman (rank) correlation makes Shesha robust to monotonic transformations of distances and insensitive to outliers in the distance distribution.

Relationship to Existing Metrics. Shesha complements, rather than replaces cross-model similarity metrics. Centered Kernel Alignment (CKA) (Kornblith et al., 2019a), Projected Weight CCA (PWCCA) (Morcos et al., 2018), Projected Weight CKA (PWCKA) (Morcos et al., 2018; Kornblith et al., 2019a), Procrustes distance (Schönemann, 1966), Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008), and others measure whether two representations encode *similar* geometry. Shesha measures whether a *single* representation's geometry is *internally consistent*. As demonstrated in Section 2, these properties are empirically dissociable. Representations can be similar but unstable, or stable but dissimilar. This independence establishes Shesha as a distinct diagnostic axis for representation analysis.

Relationship to Split-Half Reliability. Shesha adapts the principle of the split-half reliability estimator from the RSA literature (Nili et al., 2014), but applies it along a fundamentally different axis and with new variants. Traditional split-half reliability partitions *observations* (trials or subjects) to estimate a noise ceiling: how well could any model RDM correlate with neural data given measurement variability? This answers a question about *data quality*. Shesha instead partitions *features* (neurons or embedding dimensions) to measure whether geometric structure is distributed redundantly across the representation. This answers a question about *representational architecture*: is the geometry holographically encoded such that arbitrary subsets of dimensions preserve relational structure, or is it concentrated in fragile, low-dimensional subspaces?

The mathematical machinery is identical; the interpretive frame is distinct. A low noise ceiling diagnoses unreliable measurements. A low Shesha score diagnoses brittle geometry, which may reflect architectural choices (e.g., sparse coding), training dynamics (e.g., feature collapse), or intrinsic properties of the domain (e.g., high-dimensional but low-rank structure). Notably, Shesha requires no repeated measurements, enabling stability assessment for systems like pretrained embeddings or single-cell profiles where observation-level replication is unavailable or undefined.

6.2 Metric Validation Tests

We conducted a comprehensive suite of tests to assess the reliability, robustness, and interpretability of the Shesha metric. Unless noted, all of the experiments described here used the **feature-split**

variant. This variant measures Shesha by randomly partitioning feature dimensions into two disjoint halves and computes pairwise cosine distance matrices (RDMs) for each half after L2 normalization and measures the Spearman rank correlation between the two RDMs. The final score is averaged over $S = 30$ random splits. This unsupervised form measures the internal geometric consistency of learned representations without requiring any class labels.

We evaluated 15 pretrained vision models spanning different architectural families: ResNets (resnet18, resnet34, resnet50, seresnet50), efficiency-focused architectures (densenet121, mobilenetv3_large_100), EfficientNets (efficientnet_b0, efficientnet_b2), ConvNeXt variants (convnext_tiny, convnext_small), and vision transformers (vit_tiny_patch16_224, vit_small_patch16_224, vit_base_patch16_224, swin_tiny_patch4_window7_224, deit_small_patch16_224). Embeddings were extracted from CIFAR-10 and CIFAR-100 test sets using each model's canonical preprocessing pipeline via the `timm` library (Wightman, 2019) with `num_classes=0` to obtain penultimate layer features. All experiments enforced strict determinism through fixed random seeds (seed=320), single-threaded execution (`OMP_NUM_THREADS=1`, `MKL_NUM_THREADS=1`), and deterministic CUDA operations (`torch.use_deterministic_algorithms(True)`, `CUBLAS_WORKSPACE_CONFIG=:4096:8`).

6.2.1 Metric Properties and Robustness

Test 1: Convergence and Sample Efficiency. To assess whether Shesha estimates converge reliably as sample size varies, we measured the metric at two sample sizes ($n \in \{400, 1600\}$) across all 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. We computed Shesha on the subsample, and measured the drift $\Delta = \text{Shesha}_{n=400} - \text{Shesha}_{n=1600}$. Stability was defined as $|\Delta| < 0.05$.

Shesha estimates demonstrated excellent convergence properties across all architectures (Table 2, Figure 2). At $n = 400$, scores ranged from 0.254 (vit_tiny_patch16_224 on CIFAR-100) to 0.803 (efficientnet_b2 on CIFAR-10) with mean 0.622. At $n = 1600$, scores ranged from 0.288 (vit_tiny_patch16_224 on CIFAR-100) to 0.790 (swin_tiny on CIFAR-10) with mean 0.622, showing negligible change in the distribution. The mean absolute drift across all 30 model-dataset combinations was $|\bar{\Delta}| = 0.0115$, well below the 0.05 stability threshold. When averaged per-model across both datasets, drifts ranged from 0.0002 (resnet50, most stable) to 0.0176 (vit_tiny_patch16_224, least stable), with mean 0.0077. All 15 models achieved stable estimates across both datasets. These results demonstrate that reliable Shesha measurements can be obtained with as few as 400 samples.

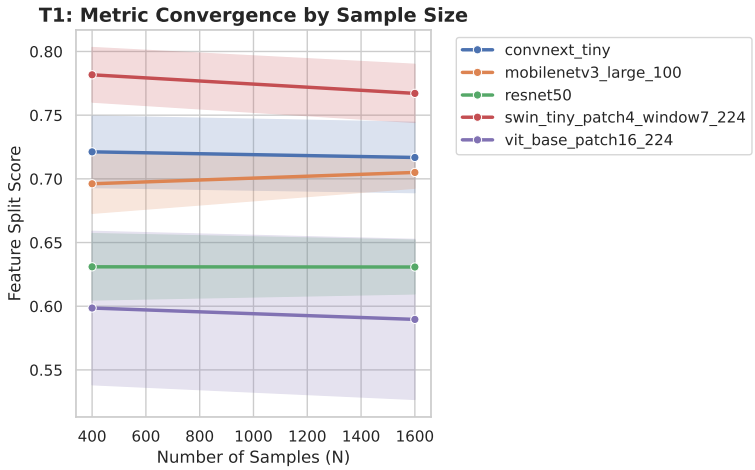

Figure 2: **Metric Convergence.** Shesha estimates remain stable as sample size increases from 400 to 1600 across representative architectures. The flat trajectories confirm rapid convergence and numerical reliability at modest sample sizes.

Table 2: **Convergence Analysis.** Mean score drift (Δ) between $n = 400$ and $n = 1600$ samples, averaged across CIFAR-10 and CIFAR-100. All models satisfy stability threshold ($|\Delta| < 0.05$).

Model	**Mean Δ**	**Stable**
resnet50	$+0.0002$	✓
efficientnet_b0	$+0.0023$	✓
densenet121	-0.0032	✓
resnet18	-0.0039	✓
convnext_tiny	$+0.0044$	✓
deit_small_patch16_224	-0.0070	✓
efficientnet_b2	$+0.0086$	✓
vit_small_patch16_224	-0.0086	✓
resnet34	$+0.0087$	✓
mobilenetv3_large_100	-0.0089	✓
vit_base_patch16_224	$+0.0089$	✓
convnext_small	$+0.0094$	✓
seresnet50	-0.0095	✓
swin_tiny_patch4_window7_224	$+0.0146$	✓
vit_tiny_patch16_224	-0.0176	✓
Mean $\lvert\Delta\rvert$	**0.0077**	**15/15**

Test 2: Model Leaderboard. To create a benchmark ranking that verifies Shesha's ability to differentiate among model architectures in a consistent and meaningful way, we estimated scores for 15 different models on the CIFAR-10 and CIFAR-100 datasets. For each model-dataset combination, we used the first 2000 test examples (capped at $n = 1600$ for the feature split computation due to the $O(n^2)$ RDM calculation). The individual dataset score and the combined score for both datasets were calculated to evaluate the degree of cross-dataset consistency.

The model leaderboard revealed substantial and meaningful variation in geometric stability across architectures (Table 3, Figure 3). Total scores ranged from 0.699 (vit_tiny_patch16_224) to 1.550 (swin_tiny_patch4_window7_224), spanning a dynamic range of 0.851 with a mean 1.244 ± 0.227. The top-performing model was swin_tiny_patch4_window7_224 (CIFAR-10: 0.794, CIFAR-100: 0.757, Total: 1.550), followed by efficientnet_b2 (0.782, 0.735, 1.517) and convnext_small (0.757, 0.708, 1.466). The bottom tier was dominated by smaller vision transformer variants: vit_tiny_patch16_224 (0.421, 0.279, 0.699) and vit_small_patch16_224 (0.536, 0.367, 0.903), suggesting that geometric stability in transformers scales with model capacity. Mobilenetv3_large_100 exhibited good performance (1.400) as an architecture with a strong focus on efficiency. Additionally, the high correlation between the CIFAR-10 and CIFAR-100 rankings (Spearman $\rho = 0.93$, $p < 0.0001$) confirms that Shesha is capturing more than just dataset-specific properties of the architecture. Instead it appears to capture intrinsic architectural properties rather than dataset-specific artifacts.

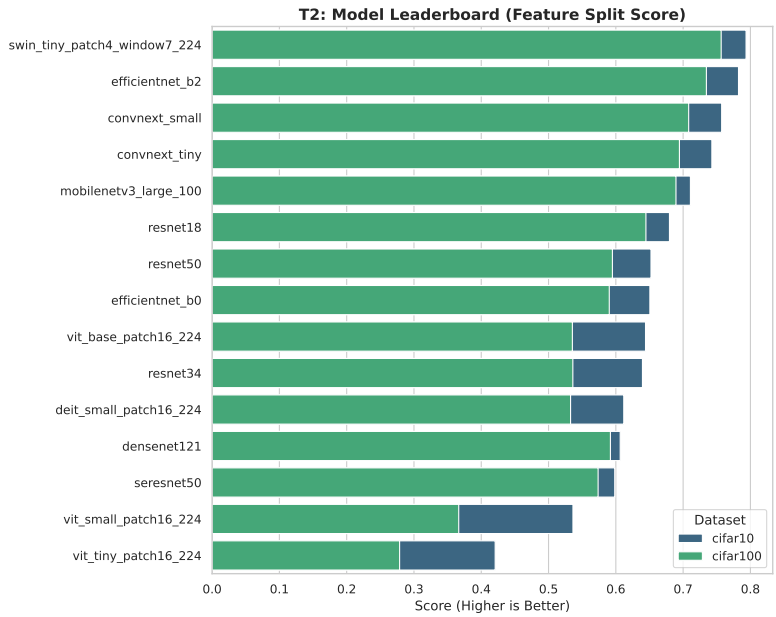

Figure 3: **Model Leaderboard.** Ranking of 15 architectures by Shesha score (feature split). Bar segments show contributions from CIFAR-10 (teal) and CIFAR-100 (blue). Modern architectures with attention or dense connectivity achieve higher geometric stability.

Table 3: **Model Leaderboard.** Shesha scores (feature split variant) computed on 1600 samples per dataset. Models ranked by total score.

Rank	Model	CIFAR-10	CIFAR-100	Total
1	swin_tiny_patch4_window7_224	0.794	0.757	1.550
2	efficientnet_b2	0.782	0.735	1.517
3	convnext_small	0.757	0.708	1.466
4	convnext_tiny	0.743	0.694	1.437
5	mobilenetv3_large_100	0.711	0.689	1.400
6	resnet18	0.680	0.645	1.324
7	resnet50	0.652	0.595	1.247
8	efficientnet_b0	0.651	0.590	1.241
9	densenet121	0.606	0.592	1.198
10	vit_base_patch16_224	0.644	0.535	1.180
11	resnet34	0.640	0.536	1.176
12	seresnet50	0.598	0.574	1.172
13	deit_small_patch16_224	0.612	0.533	1.145
14	vit_small_patch16_224	0.536	0.367	0.903
15	vit_tiny_patch16_224	0.421	0.279	0.699
	Mean ± Std	0.655 ± 0.098	0.589 ± 0.132	1.244 ± 0.227

Test 3: Determinism. To confirm that Shesha produces bit-identical outputs after evaluating the same input data multiple times, we measured the metric twice using identical inputs for the two different measurements. We measured the random viewing size metric for the different models, a total of 15, trained on the CIFAR-10 and CIFAR-100 datasets, from the first 400 samples. We then created three separate Shesha metrics: variance, zscore, and feature_split, and performed all three variant metrics with the same input dataset. We computed the absolute differences of these results: $|v_1 - v_2|$.

The metric achieved perfect bitwise determinism across all 90 test conditions (15 models $\times$ 2 datasets $\times$ 3 variants). Every pairwise difference was exactly 0.0 to floating-point precision. This confirms that (1) the random number generators are properly seeded and produce identical sequences across calls, (2) the feature split procedure with seed-based RNG produces identical partitions, and (3) no stochastic operations corrupt reproducibility. The 100% pass rate validates that the implementation produces reproducible results suitable for scientific benchmarking.

Test 4: Numerical Validity. To ensure that the metric produces valid floating-point outputs across the full diversity of embedding distributions encountered in practice, we systematically tested for numerical pathologies. For each of the 15 models on both CIFAR-10 and CIFAR-100, we randomly sampled 1000 examples using a fixed random generator, computed all three Shesha variants (variance, zscore, feature_split), and verified that each output satisfied: (1) not NaN (`np.isnan`), (2) not infinite (`np.isinf`), and (3) within expected theoretical bounds.

All 90 test conditions (15 models $\times$ 2 datasets $\times$ 3 variants) produced valid numerical outputs (Table 4). No NaN or infinite values were observed across any model-dataset-variant combination. The feature_split variant produced scores in the range $[0.285, 0.792]$ with mean 0.620, consistent with Spearman correlations bounded in $[-1, 1]$. The variance variant (the ratio of between-class to total variance) ranged from $[0.152, 0.398]$, with a mean of 0.277, indicating that 15-40% of total variance is explained by class structure depending on the architecture. The zscore variant ranged from $[0.053, 0.146]$ with mean 0.096. These ranges confirm that the metric handles the diversity of embedding magnitudes (ranging from $O(1)$ for normalized features to $O(10^2)$ for unnormalized activations), dimensionalities (512 to 2048 depending on architecture), and distributional shapes produced by modern vision architectures without numerical instability.

Table 4: **Numerical Validity.** Score ranges across all 90 test conditions. All outputs were valid floats (no NaN/Inf).

Variant	Min	Max	Mean	Valid
feature_split	0.285	0.792	0.620	30/30
variance	0.152	0.398	0.277	30/30
zscore	0.053	0.146	0.096	30/30
Total	-	-	-	**90/90**

This test was expanded into a larger and more comprehensive experiment. See Section 3.2 and Appendix 9 for more details.

Test 5: Dimensionality Sensitivity. To assess how Shesha 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 model-dataset combination, we extracted 400 samples, fit PCA with `n_components=64` and `random_state=320`, transformed the embeddings, and recomputed all three Shesha variants on the reduced representations.

Dimensionality reduction produced systematic and variant-specific effects (Table 5). The feature_split variant was most severely affected, producing negative mean scores (-0.112, range: $[-0.204, -0.055]$) after PCA reduction. This occurs because PCA concentrates variance into the top principal components, creating strong correlations across the feature dimensions. When these correlated features are randomly split, the two halves capture overlapping rather than complementary geometric information. The slightly negative scores reflect that principal components are orthogonal by construction; randomly splitting them can separate paired variance components, leading to near-zero or slight noise-induced anti-correlation in the resulting RDMs. The supervised variants remained positive but showed altered distributions: variance increased from 0.277 to 0.426 (range: $[0.228, 0.634]$) because PCA preserves class-discriminative directions, while zscore increased from 0.096 to 0.173 (range: $[0.083, 0.292]$). These results establish that Shesha (feature split) measurements should be performed on full-dimensional embeddings.

Table 5: **Dimensionality Sensitivity.** Effect of PCA reduction (64 components) on Shesha variants. Statistics across 30 model-dataset conditions.

Variant	Original Mean	Reduced Mean	Reduced Range
feature_split	$+0.620$	-0.112	$[-0.204, -0.055]$
variance	$+0.277$	$+0.426$	$[+0.228, +0.634]$
zscore	$+0.096$	$+0.173$	$[+0.083, +0.292]$

Test 6: Label Noise Sensitivity. To verify that supervised Shesha variants appropriately reflect label-geometry alignment and degrade when semantic structure is destroyed, we compared scores computed with true labels versus fully randomized labels. For each model-dataset combination, we extracted 500 samples with their ground-truth labels, computed the supervised variants (variance, zscore), then randomly permuted the label vector using a fixed random generator and recomputed both variants. This test does not apply to the unsupervised feature_split variant.

Both supervised variants showed dramatic score reductions under complete label randomization, with dataset-dependent effect sizes (Table 6). On CIFAR-10 (10 classes), the variance formulation dropped from 0.216 to 0.018 (91.6% reduction), and zscore dropped from 0.078 to 0.006 (92.7% reduction). For both variance and zscore, the reductions were smaller on CIFAR-100 with variance dropping 0.415 to 0.199 (52.2% reduction) and zscore dropping from 0.155 to 0.067 (56.7% reduction). This effect is due to the increase in the number of classes for CIFAR-100, which results in an increased number of partitions in the embedding space. Overall means showed variance dropping from 0.316 to 0.108 (65.7% reduction) and zscore from 0.117 to 0.036 (68.8% reduction). These results confirm that supervised variants capture genuine label-geometry correspondence rather than spurious statistical artifacts.

Table 6: **Label Noise Sensitivity.** Score comparison between true labels (0% noise) and fully randomized labels (100% noise). Scores averaged across all 15 models.

Dataset	Variant	True Labels	Random Labels	Δ	Drop
CIFAR-10	variance	0.216	0.018	0.198	91.6%
CIFAR-10	zscore	0.078	0.006	0.072	92.7%
CIFAR-100	variance	0.415	0.199	0.216	52.2%
CIFAR-100	zscore	0.155	0.067	0.088	56.7%
Overall	**variance**	**0.316**	**0.108**	**0.207**	**65.7%**
Overall	**zscore**	**0.117**	**0.036**	**0.080**	**68.8%**

Test 7: Class Imbalance Robustness. To simulate a real-world application, we tested the metric stability under severe class imbalance. We constructed maximally imbalanced subsets by sampling 100 examples from the first class and only 5 examples from each remaining class, yielding a total of 145 samples for CIFAR-10 ($100 + (9 \times 5)$) and a total of 595 samples for CIFAR-100 ($100 + (99 \times 5)$). This creates a 20:1 imbalance ratio. We computed the supervised Shesha variants on these imbalanced subsets.

The metric remained computable and interpretable even under a major 20:1 class imbalance, with dataset-dependent behavior reflecting the underlying class structure (Table 7). On CIFAR-10, the variance formulation showed tight clustering across models (mean: 0.178, std: 0.022, range: $[0.144, 0.216]$), while zscore was similarly stable (mean: 0.144, std: 0.018, range: $[0.120, 0.175]$). On CIFAR-100, variance scores were substantially higher and more variable (mean: 0.407, std: 0.040, range: $[0.339, 0.473]$), reflecting the greater between-class separation achievable with 100 categories even under imbalance. The zscore formulation remained remarkably stable across datasets (CIFAR-10: 0.144 ± 0.018; CIFAR-100: 0.144 ± 0.020), demonstrating robustness to both imbalance and class count. These results indicate that Shesha handles severely imbalanced data without numerical failure, though cross-dataset comparisons should account for class-count effects.

Table 7: **Class Imbalance Robustness.** Stability under 20:1 imbalance (100 samples from one class, 5 from others). Statistics across 15 models.

Dataset	Variant	n	Mean	Std	Min	Max
CIFAR-10	variance	145	0.178	0.022	0.144	0.216
CIFAR-10	zscore	145	0.144	0.018	0.120	0.175
CIFAR-100	variance	595	0.407	0.040	0.339	0.473
CIFAR-100	zscore	595	0.144	0.020	0.112	0.178
Overall	**variance**	-	**0.292**	**0.121**	**0.144**	**0.473**
Overall	**zscore**	-	**0.144**	**0.019**	**0.112**	**0.178**

Test 8: Input Perturbation Robustness. To assess sensitivity to embedding-level noise, which simulates measurement error, adversarial perturbations, or representation drift, we added isotropic Gaussian noise to the embeddings and measured score stability. For each model-dataset combination, we extracted 400 samples, computed Shesha on the clean embeddings, then added noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.1$ and recomputed. The drift was measured as $|\text{Shesha}_{\text{clean}} - \text{Shesha}_{\text{noisy}}|$.

Models exhibited highly variable robustness to input perturbations, revealing clear architectural patterns (Table 8). For the feature_split variant, the mean drift across all 30 model-dataset combinations was 0.0176 ± 0.0227, but individual drifts spanned three orders of magnitude from 0.00004 to 0.073. The most robust architecture was densenet121, achieving drift of only 0.00004 on CIFAR-100 and 0.00005 on CIFAR-10. This is approximately 1800$\times$ more stable than the least robust. Vision transformers also showed excellent robustness: vit_small_patch16_224 (0.00009 on CIFAR-100), vit_tiny_patch16_224 (0.00009 on CIFAR-100). The least robust architectures were ResNet variants: seresnet50 (0.0728 on CIFAR-10), resnet50 (0.0704 on CIFAR-10), resnet34 (0.0552 on CIFAR-10), resnet18 (0.0521 on CIFAR-10). This pattern suggests that dense connectivity (DenseNet) and attention mechanisms (ViT) create representations that are more robust to additive noise than standard residual connections, possibly because they distribute information across more pathways.

Table 8: **Perturbation Robustness.** Score drift (feature_split) under Gaussian noise ($\sigma = 0.1$). Top 5 most/least robust combinations shown.

Model	Dataset	Clean	Drift
Most Robust:			
densenet121	CIFAR-100	0.596	0.00004
densenet121	CIFAR-10	0.607	0.00005
vit_small_patch16_224	CIFAR-100	0.374	0.00009
vit_tiny_patch16_224	CIFAR-100	0.261	0.00009
vit_tiny_patch16_224	CIFAR-10	0.404	0.00088
Least Robust:			
seresnet50	CIFAR-10	0.603	0.0728
resnet50	CIFAR-10	0.645	0.0704
resnet34	CIFAR-10	0.649	0.0552
resnet18	CIFAR-10	0.680	0.0521
resnet50	CIFAR-100	0.596	0.0451
Mean $\pm$ Std (all 30)			$\mathbf{0.0176 \pm 0.0227}$

Test 9: Seed Stability. To test that the stochastic feature splitting procedure produces consistent estimates on different random initializations, we measured the sensitivity of Shesha, which was computed using two independent seeds. For each model-dataset combination, we extracted the first 1600 samples and computed the feature_split variant twice: once with `seed=100` and once with `seed=200`. Each seed generates a different sequence of 30 random feature partitions. We measured sensitivity as $|\text{Shesha}_{\text{seed}=100} - \text{Shesha}_{\text{seed}=200}|$.

The metric demonstrated excellent seed stability across all architectures and datasets (Figure 4, Table 9). The mean sensitivity across all 30 model-dataset combinations was 0.0047, with a maximum

of 0.0142 (resnet34 on CIFAR-100) and a minimum of 0.00015 (resnet50 on CIFAR-10). The five most stable measurements were: resnet50/CIFAR-10 (0.00015), swin_tiny/CIFAR-10 (0.00017), vit_small/CIFAR-10 (0.00053), efficientnet_b2/CIFAR-100 (0.00071), and resnet18/CIFAR-100 (0.00095). All 30 combinations resulted in sensitivity that was well below the 0.05 stability threshold, with 25/30 below 0.01. These results confirm that averaging over $S = 30$ random splits provides sufficient variance reduction to yield highly reproducible estimates, with typical seed-to-seed variation of less than 1% of the score magnitude.

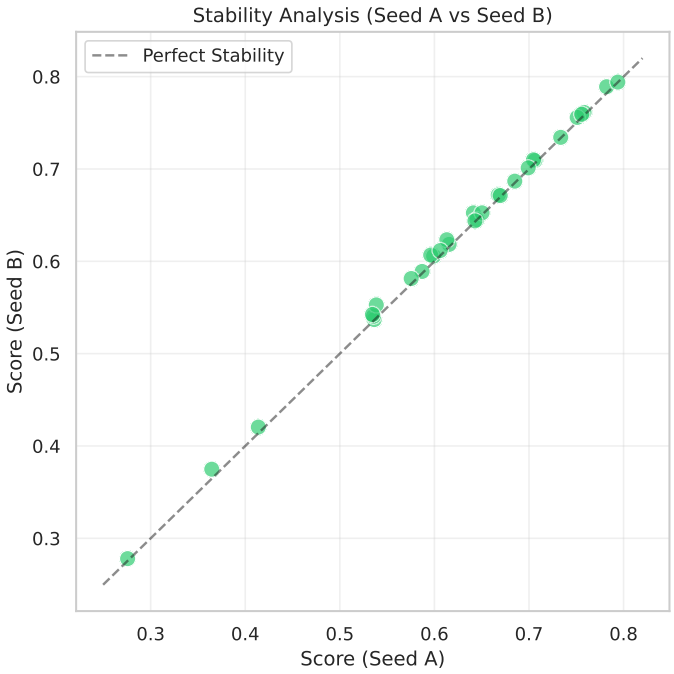

Figure 4: **Seed Stability.** Comparison of Shesha scores computed with two different random seeds (Seed A=100 vs. Seed B=200). Points align closely with the diagonal identity line, indicating high reproducibility across random initializations.

Table 9: **Seed Stability.** Sensitivity of feature_split scores to random seed choice. Top 5 most stable combinations shown. All 30 achieved $|\Delta| < 0.05$.

Model	**Dataset**	**Seed A**	$\lvert\Delta\rvert$	$\lvert\Delta\rvert/\mu$
resnet50	CIFAR-10	0.645	0.00015	0.02%
swin_tiny_patch4_window7_224	CIFAR-10	0.794	0.00017	0.02%
vit_small_patch16_224	CIFAR-10	0.536	0.00053	0.10%
efficientnet_b2	CIFAR-100	0.734	0.00071	0.10%
resnet18	CIFAR-100	0.643	0.00095	0.15%
Mean Sensitivity (all 30)			**0.0047**	**0.76%**
Max Sensitivity			**0.0142**	**2.64%**
Count $\lvert\Delta\rvert < 0.01$			**25/30**	

Test 10: Sanity Baseline. To confirm that Shesha captures learned geometric structure rather than statistical artifacts inherent to high-dimensional spaces, we computed the metric on pure random data with no learned structure. We generated random Gaussian embeddings $X \sim \mathcal{N}(0, I)$ with $n = 500$ samples and $d = 128$ dimensions (chosen to approximate typical embedding sizes), along with random integer labels $y \sim \text{Uniform}\{0, 1, \ldots, 9\}$. We computed all three Shesha variants on this null distribution.

Random embeddings produced near-zero scores across all metric variants, confirming that high scores require genuine geometric structure (Table 10). The feature_split variant yielded 0.0027, indicating essentially zero correlation between RDMs computed on random feature partitions, as expected when features are i.i.d. Gaussian. The variance variant produced 0.0192, reflecting the small spurious between-class variance that arises from the random partitioning of isotropic data into 10 groups. The zscore variant yielded 0.0060. In comparison to the trained model scores (Table 4), the feature_split baseline (0.0027) is 106× lower than the lowest trained model score (0.285 for vit_tiny on CIFAR-100) and 293× lower than the highest (0.792 for swin_tiny on CIFAR-10). This establishes a clear floor: Shesha scores near zero indicate an absence of learned structure, while scores above 0.3 indicate meaningful geometric organization.

Table 10: **Sanity Baseline.** Shesha scores on random Gaussian embeddings ($n = 500$, $d = 128$) with random labels. Ratios computed against lowest trained model score from Table 4.

Variant	**Baseline**	**Lowest Trained**	**Ratio**
feature_split	0.0027	0.285	106×
variance	0.0192	0.152	7.9×
zscore	0.0060	0.053	8.8×

7 Evaluating Distinction of Stabiltiy and Similarity: Extended Methods and Results

This appendix contains all the methods and additional analyses the distinctness validation experiments outlined in Section 2. We establish that geometric stability (Shesha) and representational similarity (CKA) measure fundamentally different properties through three complementary analyses. First through a ground-truth validation we demonstrate that Shesha tracks known stability levels with high fidelity. Second, we show cross-domain evaluation across 2,463 encoder configurations in seven machine learning and biological domains to show minimal correlation between stability and similarity metrics. Finally, we present a mechanistic decomposition, revealing how geometry-preserving versus geometry-altering transformations produce opposing stability-similarity relationships that cancel in aggregate.

7.1 Ground Truth Validation of Shesha

To establish that Shesha measures a construct distinct from existing similarity metrics, we conducted four validation experiments using synthetic representations with known properties. Shesha was computed using $N = 50$ split-half iterations with correlation distance for RDM construction. Tests 1, 2, and 4 used random seed 320 for reproducibility; Test 3 (spectral deletion) used deterministic operations requiring no random seed.

7.1.1 Test 1: 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 ($\rho = 0.987$), confirming it accurately measures internal representational consistency.

7.1.2 Test 2: Debiased CKA Baseline

Standard linear CKA exhibits a positive bias for random high-dimensional matrices due to diagonal Gram matrix terms, yielding spurious similarity scores of $\sim$0.4 for independent noise. We implemented debiased CKA using the unbiased estimator of HSIC (Song et al., 2012), which implies zeroing the Gram matrix diagonals, as applied to CKA by Kornblith et al. (2019a). We verified that debiased CKA correctly returns near-zero for independent random representations $X, Y \sim \mathcal{N}(0, I)^{200 \times 256}$ (debiased CKA $= -0.006$, Shesha $= -0.010$; seed $\mathcal{S}[1]$ for generation, $\mathcal{S}[2]$ for Shesha splits). Negative values are expected for an unbiased estimator when true similarity is near zero. All subsequent ground-truth analyses use debiased CKA.

7.1.3 Test 3: Spectral Sensitivity

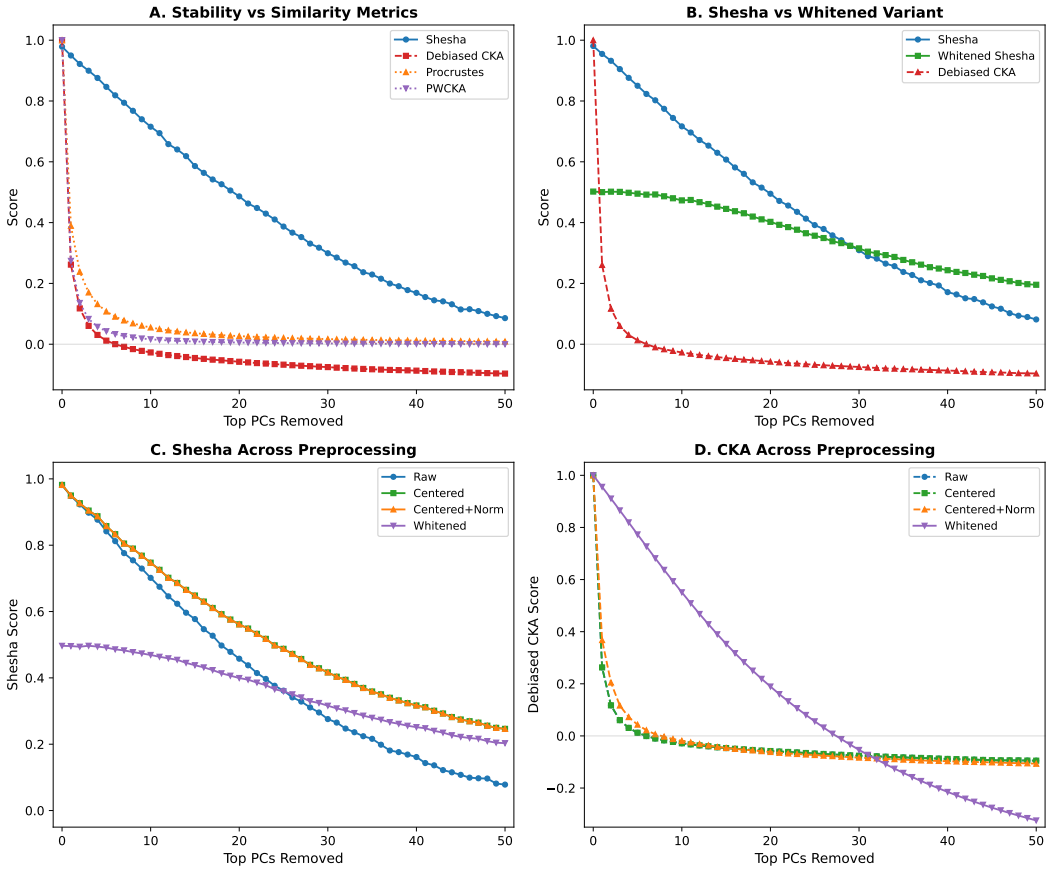

Figure 5: **Spectral Sensitivity Analysis.** We measure metric responses as the top k principal components are progressively removed from a power-law representation. **(A)** Shesha degrades gracefully while all similarity metrics (CKA, PWCKA, Procrustes) collapse after removing just 1 PC. **(B)** Comparison with whitened Shesha shows high correlation ($\rho = 0.999$), though whitening reduces baseline stability. **(C)** Shesha robustness across preprocessing conditions (raw, centered, normalized, whitened). **(D)** CKA behavior across preprocessing; notably, whitening causes CKA to recover sensitivity by equalizing the spectrum.

To test whether Shesha and similarity metrics capture different aspects of representational structure, we generated representations with power-law eigenspectra (mimicking trained neural networks). Specifically, we constructed $X = USV^\top$ where $U \in \mathbb{R}^{200\times200}$ and $V \in \mathbb{R}^{256\times256}$ are random orthogonal matrices (via QR decomposition) and S is diagonal with $S_{ii} = 100/(i+1)$. We then progressively removed the top k principal components by zeroing the first k columns of the PCA-transformed representation.

Comparison with Multiple Similarity Metrics. We compared Shesha against debiased CKA (Kornblith et al., 2019a), PWCKA (Morcos et al., 2018; Kornblith et al., 2019a), and Procrustes similarity (Schönemann, 1966) under identical preprocessing. Table 11 shows the results.

Table 11: Metric values after removing top k principal components. All similarity metrics collapse immediately while Shesha degrades gracefully, retaining sensitivity to spectral tail structure.

PCs Removed (k)	Shesha	Debiased CKA	PWCKA	Procrustes
0	0.979[a]	1.000	1.000	1.000
1	0.950	0.262	0.274	0.389
2	0.922	0.118	0.136	0.238
3	0.900	0.060	0.083	0.170
4	0.876	0.031	0.057	0.132
5	0.846	0.012	0.043	0.108
6	0.819	0.000	0.033	0.091
7	0.794	−0.009	0.027	0.078
8	0.768	−0.016	0.022	0.069
9	0.740	−0.022	0.019	0.061
10	0.715	−0.027	0.016	0.055
15	0.586	−0.045	0.009	0.036
20	0.486	−0.058	0.006	0.027
25	0.387	−0.067	0.004	0.021
30	0.299	−0.075	0.002	0.017
35	0.229	−0.082	0.002	0.014
40	0.169	−0.087	0.001	0.012
45	0.114	−0.092	0.000	0.010
50	0.086	−0.097	0.000	0.009

[a] Shesha at $k = 0$ reflects split-half reliability rather than trivial self-similarity.

All similarity metrics collapse to near-zero after removing just 1–2 dominant components (first k where metric < 0.5: CKA at $k = 1$, PWCKA at $k = 1$, Procrustes at $k = 1$, Shesha at $k = 20$). At $k = 30$, Shesha retains meaningful signal (0.299) while debiased CKA has collapsed to negative values (-0.075), indicating no detectable similarity.

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 12).

Table 12: 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

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.

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. This divergence is robust across preprocessing choices and mechanistically explained by spectral anisotropy.

7.1.4 Test 4: Dissociation with Balanced Quadrant Sampling

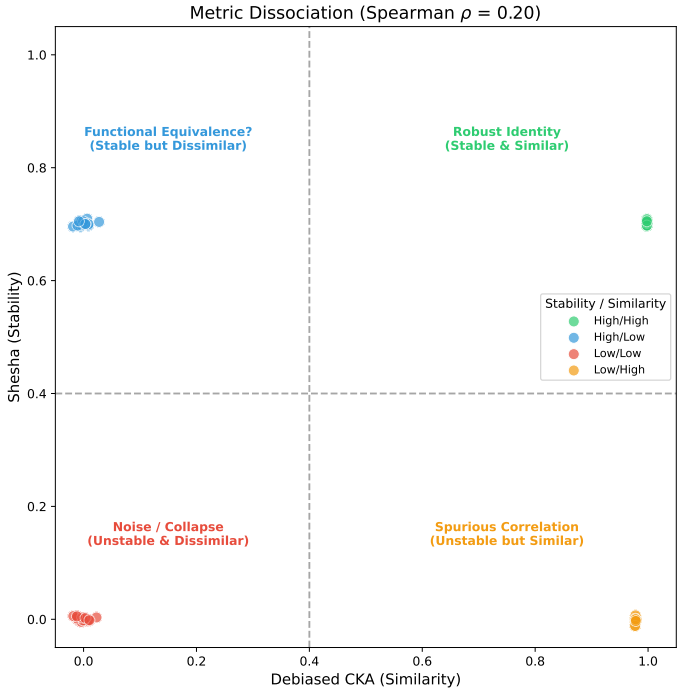

Figure 6: **Metric Dissociation.** A scatter plot of Shesha vs. Debiased CKA using balanced sampling across four stability/similarity quadrants. The presence of distinct clusters, particularly the High Stability/Low Similarity quadrant (blue), confirms that stability is mathematically distinct from similarity. The low correlation ($\rho = 0.20$) indicates that Shesha measures intrinsic geometric consistency largely independent of pairwise alignment.

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 6.

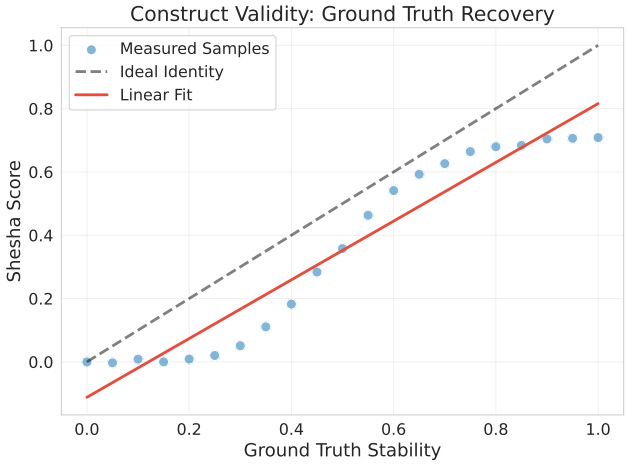

Figure 7: **Construct Validity: Ground Truth Recovery.** Shesha scores plotted against parametrically controlled stability levels (signal-to-noise ratio) in synthetic representations. The metric shows a near-perfect monotonic response ($\rho = 0.990$) to the underlying ground truth, confirming high sensitivity to geometric consistency.

7.1.5 Invariance Properties

We additionally verified that Shesha exhibits expected invariances:

- **Rotation**: Shesha is invariant to orthogonal transformations ($\Delta < 0.003$).
- **Isotropic scaling**: Shesha is invariant to uniform scaling ($\Delta < 0.002$).
- **Translation**: Shesha is invariant to mean shifts ($\Delta < 0.001$).
- **Noise injection**: Shesha degrades monotonically with additive noise, as expected for a stability metric.

These results establish that Shesha possesses construct validity (accurately tracking known stability), discriminant validity (distinct to similarity metrics), and appropriate invariance properties.

7.2 Experimental Design

Our validation framework employs encoder transformations as controlled interventions on neural representations. Rather than comparing heterogeneous pretrained models directly, which conflates architectural differences with representational properties, we extract base representations from established models and apply a standardized suite of mathematical transformations (Table 14). This design follows the perturbation analysis paradigm established in representational similarity research (Ding et al., 2021; Kornblith et al., 2019a), where controlled manipulations with known effects serve as ground-truth benchmarks for metric validation. The approach addresses a core challenge in representational analysis: two architecturally identical networks trained from different initializations learn functionally equivalent but geometrically distinct feature spaces, making direct coordinate-wise comparison meaningless without alignment (Raghu et al., 2017).

Our encoder transformation strategy provides several methodological advantages. First, it satisfies the "sanity check" criterion formalized by Kornblith et al. (Kornblith et al., 2019a), which requires that valid metrics identify correct correspondences under controlled conditions. Second, it incorporates random baselines (random projections, noise injection) justified by the Johnson-Lindenstrauss lemma, which establishes that random projections preserve distance structure in high-dimensional spaces (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2002). Third, geometry-preserving

transformations (random projections, noise injection) should affect stability and similarity in parallel, while geometry-altering transformations (aggressive PCA, feature selection) may induce dissociations, following the sensitivity-specificity framework of Ding et al. (Ding et al., 2021). The encoder approach thus provides two critical advantages: (1) it eliminates confounds from comparing models trained on different objectives or architectures, and (2) it enables systematic stress-testing of whether Shesha and similarity metrics respond differently to operations with predictable geometric consequences. Our ground-truth experiments (Section 7.1) validate this logic by demonstrating that Shesha tracks known stability levels with near-perfect fidelity ($\rho = 0.987$) while maintaining distinction to similarity metrics under balanced sampling ($\rho = 0.087$).

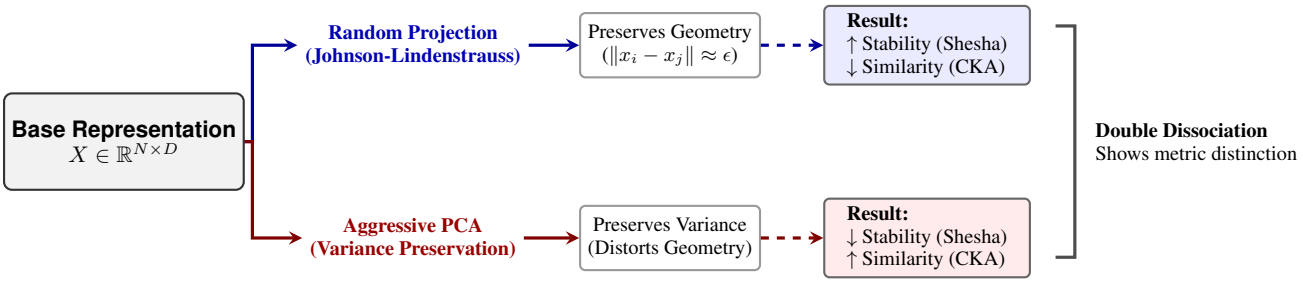

Figure 8: **Experimental Logic for Metric Independence.** We employ a double dissociation strategy to validate Shesha against CKA. Following the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2002), random projections (top) preserve geometric structure (high Stability) while discarding specific semantic features (low Similarity). Conversely, aggressive PCA (bottom) preserves dominant variance (high Similarity) but collapses the geometric manifold (low Stability). This design satisfies the specificity requirements of Kornblith et al. (Kornblith et al., 2019a).

7.3 Data Sources and Preprocessing

All representations were processed with `Float64` precision for accurate ranking and correlation computations. GPU acceleration (via CUDA) was utilized where available for PCA, Shesha, and CKA calculations. Experiments were run across 15 random seeds to ensure the stability of correlation estimates.

7.3.1 Probe Stimuli and Base Representations

For each domain, the base representations were extracted from pretrained models or structured data sources before the standardized encoder transformations were applied.

Table 13: Data sources and base models across domains.

Domain	Data Source	N samples	Base Models
Language	SST-2 validation set	500	MiniLM, MPNet, DistilBERT, RoBERTa
Vision	CIFAR-100	400	ViT, CLIP, DeiT, ResNet50
Audio	LibriSpeech dev-clean	200	Wav2Vec2, HuBERT
Video	Jellyfish sample	100	TimeSformer, VideoMAE, ViT, CLIP
Neuroscience	Steinmetz et al. (2019)	varies[a]	26 sessions
Protein	Swiss-Prot (human)	200	Sequence encoders[b]
Molecular	pbmc3k	1000	scRNA-seq

[a] Neuroscience data filtered for sessions with $\geq$20 neurons and $\geq$50 trials.

[b] Multiple encoding schemes: amino acid composition (20-dim), dipeptide frequency (400-dim), hydrophobicity profiles, charge profiles, and k-mer spectra.

Language. Sentences from the SST-2 validation (Socher et al., 2013) set were tokenized by using each model's default tokenizer with padding and truncation (max length: 64 tokens). The representations were extracted from the final hidden layer and mean-pooled

across tokens using attention masks. Base models: `all-MiniLM-L6-v2`, `all-mpnet-base-v2`, `distilbert-base-nli-stsb-mean-tokens`, and `paraphrase-distilroberta-base-v1`.

Vision. 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. Base models: `google/vit-base-patch16-224`, `openai/clip-vit-base-patch32`, `facebook/deit-base-patch16-224`, and `ResNet50` (ImageNet-V2 weights).

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

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

Neuroscience. Neural population recordings from the Steinmetz et al. dataset (Steinmetz et al., 2019), which consist of high-density Neuropixels recordings from 29,134 neurons across 42 brain regions in awake mice. For reliability of the estimation of representational geometry, sessions were filtered to include only those with at least 20 neurons and 50 trials ($N = 26$ qualified sessions). Spike counts were binned at 20ms resolution and averaged across time bins.

Protein. Protein sequences were obtained from Swiss-Prot (UniProt reviewed human proteins) (Bateman et al., 2022) and were filtered to lengths between 50-2000 residues. Multiple encoding schemes were applied: 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). This approach provides diverse representations that span different aspects of sequence properties.

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

7.4 Encoder Transformations

To test the distinctness of stability and similarity, we applied a standardized set of transformations to each base representation. This resulted in 2,463 unique encoder configurations across all seven domains (aggregated across 15 seeds: {3, 7, 9, 11, 12, 18, 103, 108, 320, 411, 724, 1754, 1991, 2222, 7258}).

Table 14: Encoder transformations applied to base representations.

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

7.5 Metrics and Statistical Analysis

Shesha (Stability). We measured geometric stability using the Feature-Split variant of Shesha (Appendix 6.1.1). For a given representation $\mathbf{X}$, the features were randomly divided into two halves a total of $S = 30$ times. For each split, we calculated the Spearman correlation between the Representational Dissimilarity Matrices (RDMs) of the two halves. The final Shesha score is determined by averaging the correlations obtained from all splits.

CKA (Similarity). We measured cross-representation similarity using linear Centered Kernel Alignment (CKA). Using three reference representations from the base model (the original, a PCA projection of $k = 100$ or the closest available, and a Z-scored version), we computed the CKA scores for each of the encoder configurations against each of the three reference representations. We averaged these scores across the three references in order to minimize the effects of single-reference artifacts.

Distinctness Criterion. We measured distinctness using the Spearman rank correlation (ρ) between Shesha and CKA scores within each domain. By following established standard effect size guidelines (Cohen, 1988), we interpret $|\rho| < 0.10$ as negligible/trivial and $|\rho| < 0.30$ as a small effect. Distinctness is supported if the aggregate correlation falls within the negligible-to-small range. Additionally, for each of the correlation estimates, we generated bootstrap confidence intervals based on 10,000 iterations.

7.6 Summary of Results

7.6.1 Aggregate Distinctness

The aggregate analysis ($N = 2{,}463$) provides robust evidence for distinctness, with $\rho = -0.01$ and a 95% CI $[-0.06, +0.03]$ falling entirely within the negligible range ($|\rho| < 0.10$). This suggests that, globally, stability and similarity capture fundamentally different geometric properties.

7.6.2 Domain-Specific Results

Domain-specific analysis reveals that 6 of 7 domains exhibit distinctness ($|\rho| < 0.30$):

- **Negligible correlations** ($|\rho| < 0.10$): Neuroscience ($\rho = +0.01$), Language ($\rho = +0.03$), Vision ($\rho = -0.03$), Molecular ($\rho = +0.06$)
- **Small correlations** ($0.10 \leq |\rho| < 0.30$): Audio ($\rho = -0.26$), Video ($\rho = -0.24$)
- **Moderate correlation**: Protein ($\rho = -0.36$)

The highest-powered domain (Neuroscience, $N = 846$) results in the tightest confidence interval $[-0.06, +0.09]$. This strongly supports the distinctness hypothesis. The Protein domain shows a moderate negative correlation, likely driven by the interaction of PCA with low-dimensional sequence encoders (20-500 dimensions), where aggressive dimensionality reduction has outsized effects.

7.6.3 Mechanistic Analysis: Encoder Type Effects

Analysis by encoder type reveals that the aggregate near-zero correlation arises from opposing effects across transformation categories:

Table 15: 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	$+0.90$ $[+0.87, +0.92]$
Noise Injection	395	$+0.58$ $[+0.50, +0.66]$
Top Variance	287	$+0.64$ $[+0.55, +0.71]$
Normalization	158	$+0.34$ $[+0.17, +0.50]$
Original	79	$+0.31$ $[+0.05, +0.56]$
PCA	948	-0.47 $[-0.52, -0.42]$

This pattern demonstrates a significant mechanistic explanation, as the use of **geometry-preserving transformations**, such as random projection and noise added back into the data, produce a similar strong positive correlation. This occurs because geometry-preserving transformations maintain the structure of the manifold that both metrics capture. By contrast, **PCA-based compression** results in a large negative correlation because it preserves dominant variance (maintaining high CKA) while destroying the fine-grained manifold structure (reducing Shesha). Therefore, the two opposing effects of geometry-preserving transformations and PCA-based compression produce a near-zero correlation due to their aggregate negative and positive impacts.

7.6.4 Robustness Checks

To verify that the results are not driven by any single domain or encoder type, we conducted several robustness analyzes:

Table 16: Robustness checks for aggregate distinctness.

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

[a]Language, Vision, Audio, Video. [b]Neuroscience, Protein, Molecular.

All robustness checks maintain $|\rho| < 0.10$, confirming that the distinction between similarity and stability is not an artifact of any particular subset. The slight negative shift when excluding the Neuroscience domain ($\rho = -0.09$) reflects the loss of the largest domain, which has near-zero correlation. The result remains within the negligible range. When the Protein domain is excluded, the aggregate shifts in the positive direction ($\rho = +0.05$). The Protein domain itself achieves distinctness ($\rho = +0.15$, CI $[-0.00, +0.30]$) when PCA encoders are excluded, confirming that its moderate negative correlation is driven specifically by compression effects on low-dimensional sequence encoders rather than a fundamental domain property.

7.7 Descriptive Statistics

Table 17: Distribution of Shesha and CKA scores across all encoder configurations.

Metric	**Mean**	**SD**	**Median**	**IQR**	**Min**	**Max**
Shesha	0.36	0.39	0.25	[0.01, 0.78]	-0.38	0.99
CKA	0.86	0.15	0.91	[0.79, 0.97]	0.15	1.00

The distributions reveal that CKA scores cluster in the high range (median 0.91) with moderate variance, while Shesha scores span a wider range with a higher amount of variance. This is consistent with its sensitivity to fine-grained geometric structures that vary substantially across encoder configurations.

7.8 Alternative Similarity Metrics

To test whether the distinction between stability and similarity is specific to CKA, we tested whether it generalizes to other similarity measures. We evaluated two alternative similarity metrics in the Language domain ($N = 127$ encoder configurations).

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:

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

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:

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

Results. The alternative metrics yield distinctness patterns similar to those of standard CKA:

Similarity Metric	ρ with Shesha	p-value	Distinct?
CKA (standard)	$+0.03$	0.74	Yes
PWCKA	-0.22	0.012	Yes
Procrustes	$+0.28$	0.001	Yes

All three metrics maintain weak correlations with Shesha ($|\rho| < 0.30$), confirming the distinction between similarity and stability. PWCKA shows a weak negative correlation, while Procrustes shows a weak positive correlation, suggesting that these metrics capture slightly different aspects of representational structure in relation to geometric stability.

Analysis by encoder type reveals consistent patterns across metrics:

Encoder Type	ρ vs CKA	ρ vs PWCKA	ρ vs Procrustes
PCA	-0.78	-0.25	-0.86
Random Projection	$+0.99$	$+0.99$	$+0.99$
Noise Injection	$+0.85$	$+0.79$	$+0.80$
Random Features	$+0.95$	$+0.95$	$+0.92$
Top Variance	$+0.91$	$+0.91$	$+0.86$
Normalization	$+0.60$	$+0.02$	$+0.64$

The most notable patterns emerge for PCA encoders, where all metrics show negative correlations with Shesha; however, Procrustes shows the strongest negative relationship ($\rho = -0.86$) compared to CKA ($\rho = -0.78$) and PWCKA ($\rho = -0.25$). This occurs because Procrustes explicitly optimizes for alignment under rotation, making it particularly sensitive to the dimensional reduction caused by PCA. For Normalization encoders, PWCKA shows near-zero correlation ($\rho = +0.02$) compared to CKA ($\rho = +0.60$) and Procrustes ($\rho = +0.64$), suggesting that effective-rank projection eliminates normalization-specific artifacts.

These results confirm that the distinction between geometric stability and representational similarity is robust to the choice of similarity metric and is not an artifact of CKA's specific formulation.

8 Steering: Extended Methods and Results

This appendix provides complete experimental details for the steering experiments described in Section 3.1. We tested the relationship between geometric stability and linear controllability through three experiments. The first was by using synthetic sentiment data and running it through 69 embedding models to map out the geometry of the data in a controlled environment with a known ground truth. The second involved using a split-half protocol on the binary sentiment classification SST-2 to completely separate the processes of metric calculation and steering evaluation. The third experiment tackled the more complex task of ternary natural language inference (MNLI) to test Shesha's resilience. All three experiments compared Shesha to a well-rounded set of geometric baselines. We have included null tests to rule out circularity and coincidence.

8.1 Experiment 1: Synthetic Steering Analysis

8.1.1 Methods

Models. We conducted an evaluation of **69 sentence embedding models** that span multiple architecture families, including MiniLM, DistilBERT, MPNet, BERT, RoBERTa, DeBERTa, E5, BGE, GTE, UAE, and different versions of SimCSE. The set of models covered three size tiers (small, base, and large) and included both supervised contrastive and unsupervised/pretrained models.

Synthetic dataset. We generated a corpus of 1,000 sentiment-laden sentences using a combinatorial grammar that was generated by a large language model to ensure broad geometric coverage:

```
template: "{context}, the {noun} was {adjective}"
contexts = ["in my opinion", "overall", "considering everything", "to be honest"]
nouns = ["aspect", "element", "part", "feature", "component", "unit", "item",
"factor"]
adj_pos = ["adequate", "fine", "good", "decent", "solid", "excellent", "superb",
"exceptional"]
adj_neg = ["poor", "bad", "mediocre", "lacking", "subpar", "terrible", "awful",
"dreadful"]
```

This produces $4 \times 8 \times 8 = 256$ unique positive and $4 \times 8 \times 8 = 256$ unique negative sentences per polarity. The synthetic nature avoids lexical memorization effects.

Data split. For each seed, we split the 1,000 samples into disjoint sets, with 500 samples for metric computation (Set A) and 500 samples for steering evaluation (Set B). The steering set was then split into 250 samples for probe training and 250 samples for testing.

Multi-seed protocol. We repeated all experiments on 15 random seeds: {3, 7, 9, 11, 12, 18, 103, 108, 320, 411, 724, 1754, 1991, 2222, 7258}. Each seed controlled: (1) data sampling, (2) train/test splits, and (3) probe initialization. This gave us $69 \times 15 = \mathbf{1{,}035}$ total observations.

Metrics computed on Set A.

- **Shesha (Supervised)**: RDM correlation between model geometry and label structure (Label-Conditioned Sample-Split variant)
- **Shesha (Unsupervised)**: Feature-partition stability (Feature-Split variant)
- **Fisher Discriminant**: Ratio of between-class to within-class variance
- **Silhouette Score**: Cluster cohesion and separation
- **Procrustes Alignment**: Similarity to ideal one-hot geometry after optimal rotation
- **Anisotropy**: Variance explained by the first principal component

Steering protocol on Set B. For each model and seed combination:

1. Train a logistic regression probe on 250 samples from Set B

2. Extract the weight vector $\mathbf{w}$ as the steering direction

3. For $\alpha \in \{-2, -1.5, ..., 1.5, 2\}$, compute the steered embeddings: $\mathbf{e}' = \mathbf{e} + \alpha\hat{\mathbf{w}}$

4. Evaluate the probe accuracy on the remaining 250 test samples

5. Record **max_drop** = $\text{acc}_0 - \min_\alpha \text{acc}(\alpha)$

Negative controls.

- **Shuffled labels**: Recompute all supervised metrics with permuted labels
- **Random directions**: Average max_drop over 20 random unit vectors per split

8.1.2 Results

Primary finding: Stability predicts steerability. Supervised geometric stability showed a strong correlation with steering effectiveness:

$$\rho(\text{Shesha}_{\text{sup}}, \text{max\_drop}) = 0.894, \quad p < 10^{-24}$$

Table 18: **Experiment 1: Metric correlations with steering (Synthetic).** Spearman ρ between geometric metrics and max_drop, aggregated by model ($n = 69$).

Category	Metric	Raw ρ	p-value	Partial ρ^a
Stability	**Shesha (Supervised)**	**0.894**	$< 10^{-24}$	**0.665**
	Shesha (Unsupervised)	0.767	$< 10^{-14}$	0.053
	Procrustes	0.797	$< 10^{-16}$	-0.089
Separability	Fisher Discriminant	0.888	$< 10^{-24}$	-
	Silhouette Score	0.889	$< 10^{-24}$	-
Structure	Anisotropy	0.710	$< 10^{-11}$	-

[a] Partial correlation controlling for Fisher + Silhouette

Key finding: Stability provides a unique signal beyond separability. After controlling for the Fisher discriminant and silhouette score, supervised Shesha maintained a large partial correlation ($\rho_{\text{partial}} = 0.665$, $p < 0.0001$), while unsupervised Shesha dropped to near zero ($\rho_{\text{partial}} = 0.053$, $p = 0.66$). This shows that **task-aligned geometric consistency captures unique variance that separability metrics miss**.

Negative controls validate methodology.

- **Shuffled labels**: Shesha dropped from 0.600 to -0.001 ($p < 10^{-12}$)
- **Random directions**: True directions produced $10.8\times$ larger effects than random ($p < 10^{-12}$)

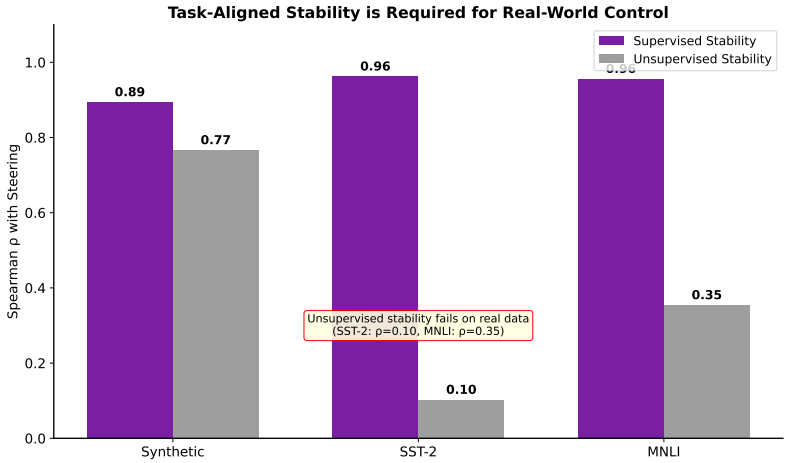

Figure 9: **Task-aligned stability is required for real-world control.** Comparison of supervised Shesha (task-aligned geometric consistency) versus unsupervised stability (feature-partition consistency) in predicting steering effectiveness. A significant gap: unsupervised stability predicts steering in the synthetic setting ($\rho = 0.77$) where the data manifold is fully aligned with the task structure but completely fails in the real-world (SST-2: $\rho = 0.10$; MNLI: $\rho = 0.35$). This indicates that for semantic control, intrinsic representational rigidity is insufficient. Stability must be aligned with the task manifold for reliable linear intervention.

8.2 Experiment 2: Binary Sentiment (SST-2)

8.2.1 Methods

Models. 35 sentence embedding models, including the MiniLM, MPNet, DistilBERT, BGE, E5, GTE, and SimCSE families, were evaluated. All models were evaluated using mean-pooled outputs from the base encoder (not from the official sentence-transformer heads) for consistency.

Split-half protocol. For each run, 800 SST-2 sentences (Socher et al., 2013) were partitioned into two **completely disjoint** sets:

- **Set A** ($n = 400$): Metric computation only
- **Set B** ($n = 400$): Steering evaluation only (200 training, 200 testing).

This design ensures that Shesha, when computed on Set A, cannot trivially predict steering on Set B through shared samples.

Multi-seed protocol. 35 models $\times$ 15 seeds = **525 total observations**.

Steering protocol. Identical to Experiment 1, with bidirectional flip rates computed at $\alpha = \pm 2.0$.

8.2.2 Results

Primary finding: Near-perfect prediction of steerability. Supervised Shesha achieved an exceptional correlation with steering effectiveness.

$$\rho(\text{Shesha}_{\text{sup}}, \text{max\_drop}) = 0.962, \quad p < 10^{-19}$$

This outperformed the Fisher discriminant ($\rho = 0.885$).

Table 19: **Experiment 2: Metric correlations with steering (SST-2).** Spearman ρ between geometric metrics and max_drop ($n = 35$ models).

Category	Metric	Raw ρ	p-value	Partial ρ^a
Stability	**Shesha (Supervised)**	**0.962**	$< 10^{-19}$	**0.764**
	Shesha (Unsupervised)	0.103	n.s.	-0.036
	Procrustes	0.976	$< 10^{-23}$	0.522
Separability	Fisher Discriminant	0.885	$< 10^{-12}$	-
	Silhouette Score	0.884	$< 10^{-12}$	-
Structure	Anisotropy	0.396	0.019	-

[a] Partial correlation controlling for Fisher + Silhouette.
Separability metrics serve as controls; partial correlations are not applicable.

Unique signal confirmed. The partial correlation of $\rho = 0.764$ after controlling for separability confirms that geometric stability captures a distinct mechanism for enabling control. Unsupervised stability did not show any predictive power ($\rho = 0.103$, n.s.).

Negative controls.

- **Shuffled labels**: Shesha dropped from 0.227 to -0.001 ($p < 10^{-10}$)
- **Random directions**: True directions produced $2.7\times$ larger effects ($p < 10^{-10}$)

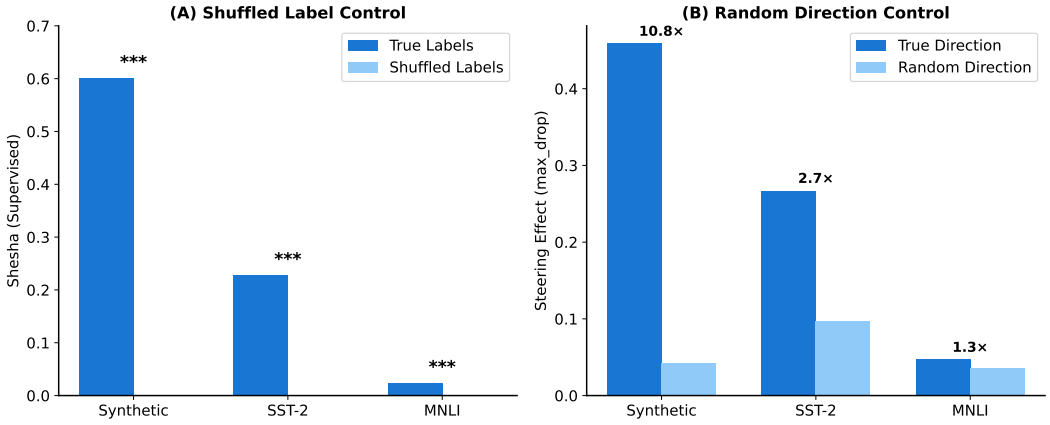

Figure 10: **Negative controls validate methodology.** **(A)** Shuffled label control: Supervised Shesha computed with true labels (dark) versus randomly permuted labels (light). The complete collapse of Shesha under label shuffling ($0.60 \rightarrow -0.001$ for Synthetic, $0.23 \rightarrow -0.001$ for SST-2, $0.02 \rightarrow -0.001$ for MNLI; all $p < 10^{-10}$) confirms that the metric captures genuine task-relevant structure rather than spurious geometric patterns. **(B)** Random direction control: Steering effect using the true probe direction (dark) versus averaged effect over 20 random unit vectors (light). True directions produce $10.8\times$ (Synthetic), $2.7\times$ (SST-2), and $1.3\times$ (MNLI) larger effects than random, which confirms direction-specific controllability. The decreasing ratio reflects narrowing steering margins as task complexity increases.

8.3 Experiment 3: Ternary NLI (MNLI)

8.3.1 Methods

Models and protocol. We repeated the identical procedure described in Experiment 2, over the same 35 models $\times$ 15 seeds = 525 observations on the MNLI dataset (Williams et al., 2018).

Tokenization. Premise-hypothesis pairs were tokenized using proper pair encoding: `tokenizer(premise, hypothesis, ...)` to ensure correct handling of separator tokens.

Multiclass steering direction. For the 3-class logistic regression, we extracted the steering direction as the **top right singular vector** of the coefficient matrix:

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad \Rightarrow \quad \hat{\mathbf{w}} = \mathbf{v}_1$$

This captures the principal direction of class discrimination (typically the Entailment-Contradiction axis) in embedding space.

8.3.2 Results

Primary finding: Stability remains predictive on complex tasks. Despite the increased complexity of the tasks, Shesha achieved strong correlation:

$$\rho(\text{Shesha}_{\text{sup}}, \text{max\_drop}) = 0.956, \quad p < 10^{-18}$$

Table 20: **Experiment 3: Metric correlations with steering (MNLI).** Spearman ρ between geometric metrics and max_drop ($n = 35$ models).

Category	Metric	Raw ρ	p-value	Partial ρ^a
Stability	**Shesha (Supervised)**	**0.956**	$< 10^{-18}$	**0.620**
	Shesha (Unsupervised)	0.354	0.037	0.100
	Procrustes	0.886	$< 10^{-12}$	0.491
Separability	Fisher Discriminant	0.952	$< 10^{-18}$	-
	Silhouette Score	0.650	$< 10^{-5}$	-
Structure	Anisotropy	0.343	0.044	-

[a] Partial correlation controlling for Fisher + Silhouette

Unique signal persists in complex settings. The partial correlation of $\rho = 0.620$ ($p = 0.0001$) indicates that stability results in a robust level of predictive power beyond separability, even for ternary classification. This contrasts with earlier analyzes suggesting "convergence" on complex tasks.

Narrower steering margins. The random direction control showed only $1.3\times$ signal-to-noise ratio (compared to $10.8\times$ for synthetic and $2.7\times$ for SST-2), which indicates that as complexity increases, steering margins become narrower. However, Shesha continues to identify robustly in such a challenging mode.

Anisotropy inconsistency. The global geometric structure (Anisotropy) showed inconsistent behavior on each of the tasks, decreasing from $\rho = 0.71$ in the synthetic setting to $\rho = 0.34$ for MNLI. This confirms that steering relies on local, task-aligned geometry rather than global embedding density.

Negative controls.

- **Shuffled labels**: Shesha dropped from 0.023 to -0.001 ($p < 10^{-10}$)
- **Random directions**: True directions produced $1.3\times$ larger effects ($p < 10^{-6}$)

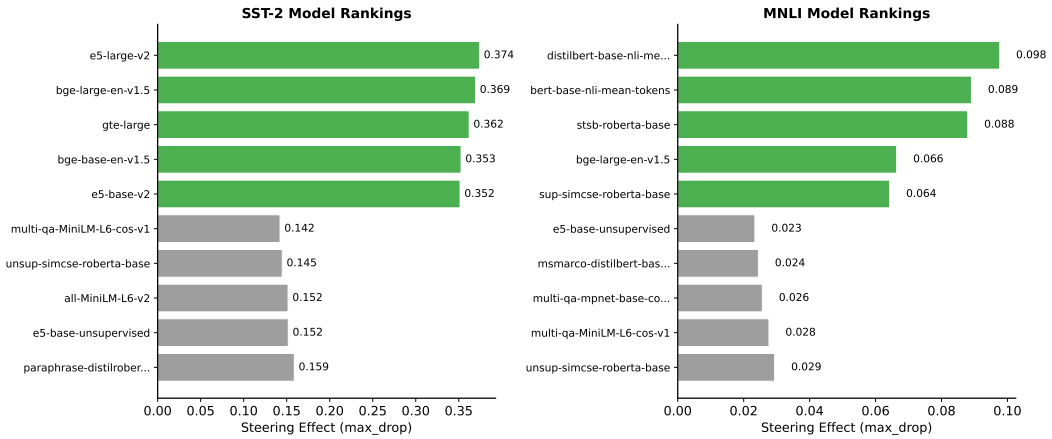

Figure 11: **Model characteristics associated with steerability.** Top 5 (colored) and bottom 5 (gray) models ranked by steering effectiveness (max_drop) for SST-2 (left) and MNLI (right). Consistent patterns emerge across tasks: the most steerable models are from the BGE, E5, and GTE families, all trained with supervised contrastive objectives. The least steerable models are unsupervised variants (`unsup-simcse`, `e5-base-unsupervised`) and retrieval-specialized models (`multi-qa-*`). This suggests that supervised contrastive training produces representations with the geometric stability required for reliable linear intervention, while unsupervised or task-misaligned training yields brittle geometries that fracture under steering.

8.4 Summary and Synthesis

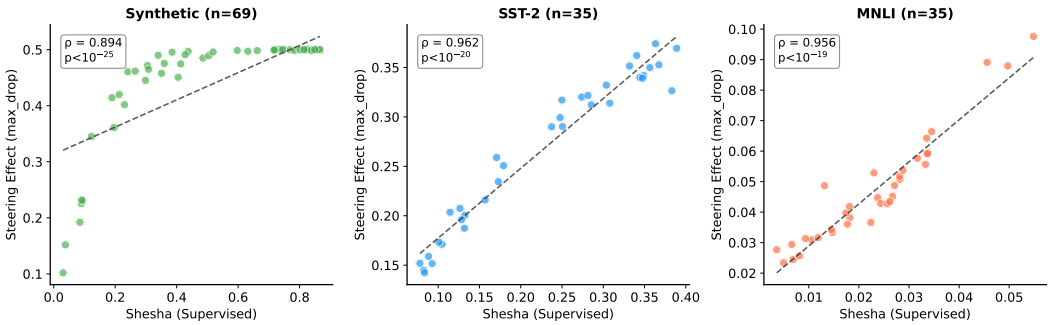

Figure 12: **Geometric stability predicts linear steerability across all experimental settings.** Scatter plots show supervised Shesha (computed on held-out Set A) versus steering effectiveness (max_drop, evaluated on disjoint Set B) for each model. **(A)** Synthetic sentiment data ($n = 69$ models): $\rho = 0.894$, $p < 10^{-24}$. **(B)** SST-2 binary sentiment ($n = 35$ models): $\rho = 0.962$, $p < 10^{-19}$. **(C)** MNLI ternary NLI ($n = 35$ models): $\rho = 0.956$, $p < 10^{-18}$. Dashed lines show linear fits. The near-perfect correlations establish Shesha as a state-of-the-art predictor of controllability, effective across both synthetic and naturalistic settings.

Table 21: **Summary: Shesha predicts steering across all experimental settings.** Partial correlations control for Fisher discriminant and silhouette score.

Dataset	Models	Obs.	Shesha (Raw)	Fisher (Raw)	Shesha (Partial)	Verdict
Synthetic	69	1,035	0.894	0.888	**0.665*****	Primary Predictor
SST-2	35	525	0.962	0.885	**0.764*****	Independent Signal
MNLI	35	525	0.956	0.952	**0.620*****	Task Invariant

***$p < 0.001$

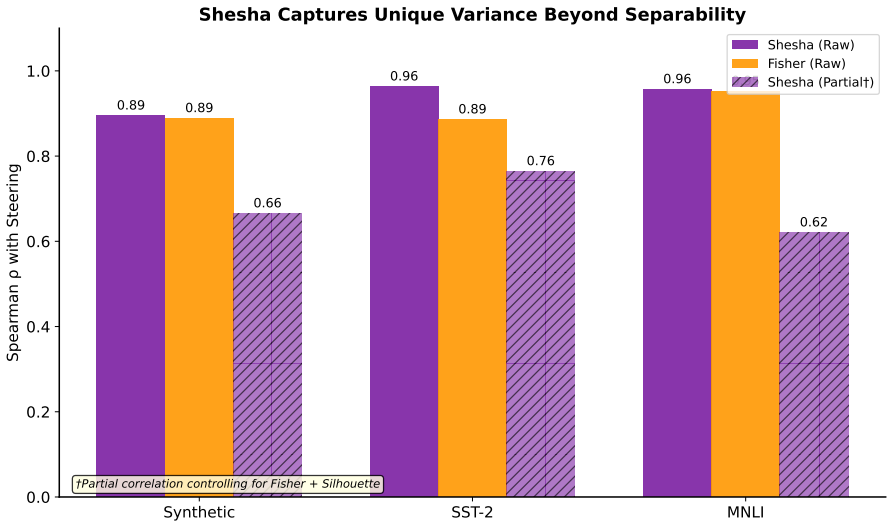

Figure 13: **Shesha captures unique variance beyond class separability.** Comparison of raw Spearman correlations (solid bars) and partial correlations controlling for Fisher discriminant and silhouette score (hatched bars). While Shesha and Fisher show similar raw correlations with steering effectiveness, Shesha maintains large partial correlations ($\rho = 0.62$-0.76, all $p < 0.001$) after controlling for separability. This demonstrates that geometric consistency (the reliability of class structure under perturbation) captures a distinct mechanism enabling control that static separability metrics miss. Separability may be necessary for steering, but stability is what guarantees it.

Key findings.

1. **State-of-the-art prediction**: Supervised Shesha achieves $\rho > 0.89$ with steering effectiveness across all settings, matching or exceeding the Fisher discriminant.
2. **Unique geometric signal**: Partial correlations of $\rho \in [0.62, 0.76]$ after controlling for separability show that stability is detecting something that separability measures miss. This shows that geometric consistency, rather than class separation, is a causal driver of controllability.
3. **Task alignment is essential**: Unsupervised stability predicted steering in synthetic settings ($\rho = 0.77$), but it failed on real-world tasks ($\rho \approx 0.10$-0.35). For semantic control, stability must be task-aligned.
4. **Methodology is sound**: Negative controls confirm that (a) supervised metrics reflect genuine task structure (shuffled labels destroy signal), and (b) steering effects are direction-specific (true directions outperform random by 1.3-$10.8\times$).

Model characteristics. Analysis of model rankings revealed that supervised contrastive models (BGE, E5, and GTE families) were consistently the most steerable, while unsupervised variants (`unsup-simcse` and `e5-base-unsupervised`) and retrieval-specialized models (`multi-qa-*`) were the least steerable. This suggests that supervised contrastive training results in a type of geometric stability that allows for reliable linear intervention.

Implications. These experiments establish supervised geometric stability as a universal prerequisite for reliable linear controllability. Practitioners can use Shesha as *a priori* diagnostic: models with high supervised stability will steer reliably, while those with low stability will fracture under perturbation, regardless of their classification accuracy.

9 Visual Perception Architecture: Extended Methods and Results

This appendix provides complete experimental details and supplementary analyses for the visual architecture benchmark described in Section 3.2. We evaluate geometric stability as a diagnostic for transfer learning through systematic comparison of 93-94 pretrained vision models across six datasets. The first analysis established the relationship between stability (Shesha-FS) and transferability (LogME, LEEP) across natural image domains (CIFAR-10, CIFAR-100). The second tested generalization to fine-grained recognition (Flowers-102, Oxford Pets), texture classification (DTD), and remote sensing (EuroSAT). The third characterized architecture family differences in geometric stability profiles.

9.1 Experimental Methods

9.1.1 Model Selection and Architecture Coverage

To evaluate the relationship between geometric stability and transferability in vision models, we created a comprehensive test which spanned 94 pretrained models from the `timm` library (Wightman, 2019). Our selection choice ensured broad coverage across four critical axes: (1) **Training Objectives:** self-supervised (DINO, DINOv2, MAE), contrastive (CLIP), generative/MIM (EVA-02, BEiT), and fully supervised (ImageNet-1k/21k); (2) **Architectural Families:** columnar vision transformers (ViT, DeiT), hierarchical transformers (Swin, SwinV2, PVT-v2), hybrid architectures (CoAtNet, MaxViT), and convolutional networks (ResNet, ConvNeXt, EfficientNet, RegNet, DenseNet); (3) **Model Scales:** from compact models (MobileNetV3-Small, ConvNeXt-Atto) to large-scale models (ViT-Giant/14 DINOv2, Swin-Large); and (4) **Training Paradigms:** standard supervised training, distillation (DeiT), data augmentation (AugReg), and foundation model pretraining.

Models were grouped into semantic families for aggregate analysis (Table 30). When training objective and architecture conflicted, we prioritized training objective for family assignment (e.g., ViT-CLIP was assigned to "CLIP" rather than "ViT").

9.1.2 Evaluation Datasets and Domain Coverage

To investigate how the stability-transferability relationship varies across task complexity and visual domains, we evaluated all models on six diverse datasets spanning four domain categories (Table 22).

Table 22: **Benchmark Dataset Summary.** Six datasets spanning four visual domains with varying class granularity.

Dataset	Domain	Classes	Models	Samples
CIFAR-10 (Krizhevsky, 2009)	Natural	10	94	5,000
CIFAR-100 (Krizhevsky, 2009)	Natural	100	94	5,000
Flowers-102 (Nilsback and Zisserman, 2008)	Fine-grained	102	93	5,000
DTD (Cimpoi et al., 2014)	Texture	47	93	1,600
EuroSAT (Helber et al., 2018)	Remote Sensing	10	93	5,000
Oxford Pets (Parkhi et al., 2012)	Fine-grained	37	93	1,500

This design enables systematic analysis across multiple axes: task complexity (CIFAR-10 vs. CIFAR-100 with identical image distributions), domain transfer (natural images vs. specialized domains), and fine-grained recognition (Flowers-102 and Oxford Pets test subtle inter-class discrimination).

9.1.3 Feature Extraction Protocol

All models were evaluated on fixed random subsets (seed 320) of 1,500-5,000 images from each dataset. We employed a dual-model extraction strategy: (1) models were instantiated with `num_classes=0` to obtain penultimate-layer representations, with features pooled using standard conventions (global average pooling for CNNs, CLS token or mean pooling for ViTs); (2) separate model instances retained classification heads for LEEP computation, yielding 82-83 models with valid LEEP scores per dataset. All evaluations used deterministic settings with disabled cuDNN benchmarking to ensure exact reproducibility.

Single seed. Due to computational constraints, we report results for a single random seed (320). Preliminary experiments with additional seeds showed consistent metric rankings.

9.1.4 Metric Computation

Shesha Feature-Split (Stability). We computed stability using the Feature-Split variant of Shesha:

$$\text{Shesha}_{\text{FS}} = \frac{1}{S}\sum_{s=1}^{S}\rho\left(\mathbf{D}_1^{(s)}, \mathbf{D}_2^{(s)}\right)$$

where $S = 10$ splits were averaged (validated in Appendix 6.2).

Shesha-Variance (Class Separability). The variance-based metric measures between-class to total variance ratio:

$$\text{Shesha}_{\text{Var}} = \frac{\text{SS}_{\text{between}}}{\text{SS}_{\text{total}}}$$

LogME and LEEP. Transferability was estimated using LogME (You et al., 2021, 2022) applied to penultimate-layer features, and LEEP (Nguyen et al., 2020) for models with valid classification heads.

9.2 Complete Results

9.2.1 Correlation Structure Across Datasets

Table 23 presents Spearman correlations across all six datasets, revealing domain-dependent stability-discriminability relationships.

Table 23: **Spearman Correlations Across All Datasets.** The stability-discriminability tradeoff varies by domain.

Metric Pair	CIFAR-10	CIFAR-100	Flowers-102	DTD	EuroSAT	Oxford Pets
LEEP vs LogME	0.83^{***}	0.81^{***}	0.11	0.39^{***}	0.19	0.14
LEEP vs Shesha-Var	0.64^{***}	0.58^{***}	0.17	0.06	0.40^{***}	-0.17
LEEP vs Shesha-FS	-0.06	-0.26^{*}	-0.10	-0.16	-0.16	-0.05
LogME vs Shesha-Var	0.59^{***}	0.55^{***}	0.47^{***}	-0.03	0.33^{**}	0.36^{***}
LogME vs Shesha-FS	-0.07	-0.19	-0.21^{*}	0.17	0.45^{***}	-0.08
Shesha-Var vs Shesha-FS	-0.01	-0.24^{*}	-0.14	0.06	0.12	-0.42^{***}

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

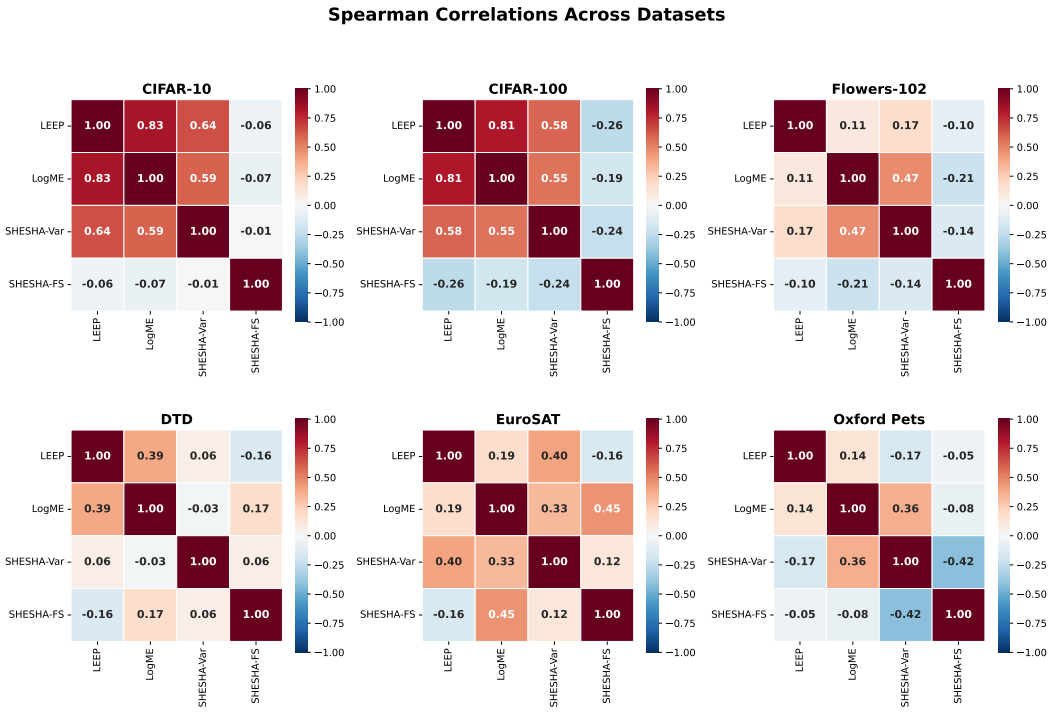

Figure 14: **Correlation Structure Across All Datasets.** Spearman correlation heatmaps for all six datasets.

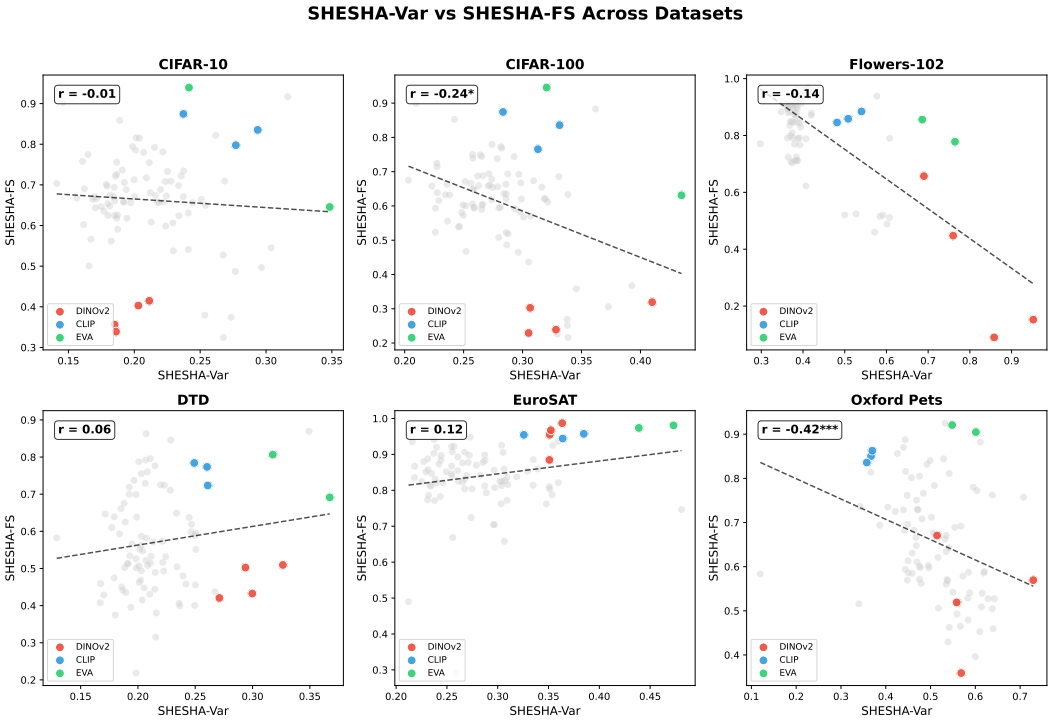

Figure 15: **Shesha-Var vs. Shesha-FS Across Datasets.** DINOv2 models (red) cluster in the high-Var/low-FS region; CLIP models (blue) maintain high stability.

9.2.2 Descriptive Statistics

Table 24: **Descriptive Statistics Across All Datasets.** Mean ± Std for each metric.

Metric	CIFAR-10	CIFAR-100	Flowers-102	DTD	EuroSAT	Oxford Pets
LEEP	-0.89 ± 0.20	-2.27 ± 0.37	-3.32 ± 0.46	-2.01 ± 0.22	-1.66 ± 0.10	-0.76 ± 0.10
LogME	0.50 ± 0.21	1.09 ± 0.10	1.35 ± 0.35	0.95 ± 0.51	0.48 ± 0.07	1.37 ± 0.70
Shesha-Var	0.21 ± 0.04	0.28 ± 0.04	0.43 ± 0.11	0.22 ± 0.04	0.29 ± 0.05	0.51 ± 0.09
Shesha-FS	0.66 ± 0.12	0.61 ± 0.15	0.82 ± 0.17	0.57 ± 0.14	0.84 ± 0.10	0.66 ± 0.13

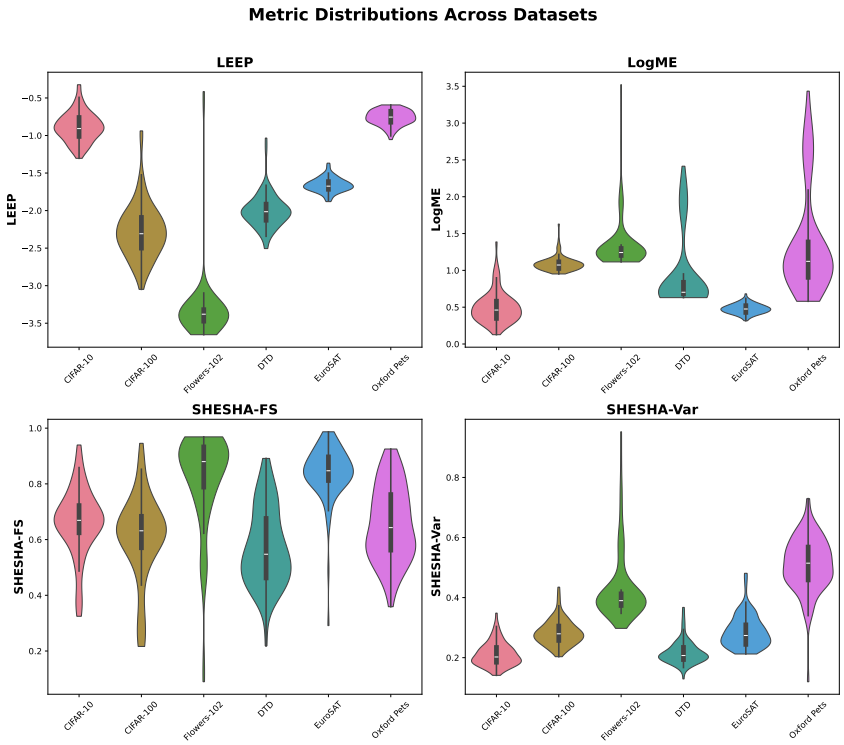

Figure 16: **Metric Distributions Across Datasets.** Violin plots comparing distributions.

9.2.3 The DINOv2 Paradox

Table 25: **The DINOv2 Paradox Across Datasets.** DINOv2-giant achieves highest LogME on 4/6 datasets while consistently ranking in the bottom quartile for Shesha-FS, except on EuroSAT where it ranks first on both metrics.

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

Table 26: **The DINOv2 Paradox (Family Averages).** DINOv2 family achieves highest LogME on multiple datasets while ranking last or near-last in Shesha-FS, except on EuroSAT.

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

9.2.4 Contrastive vs. Self-Supervised Stability

Table 27: **Contrastive vs. Self-Supervised Stability (Shesha-FS).** Mann-Whitney U tests comparing CLIP ($n = 3$) to SSL models ($n = 9$).

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

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

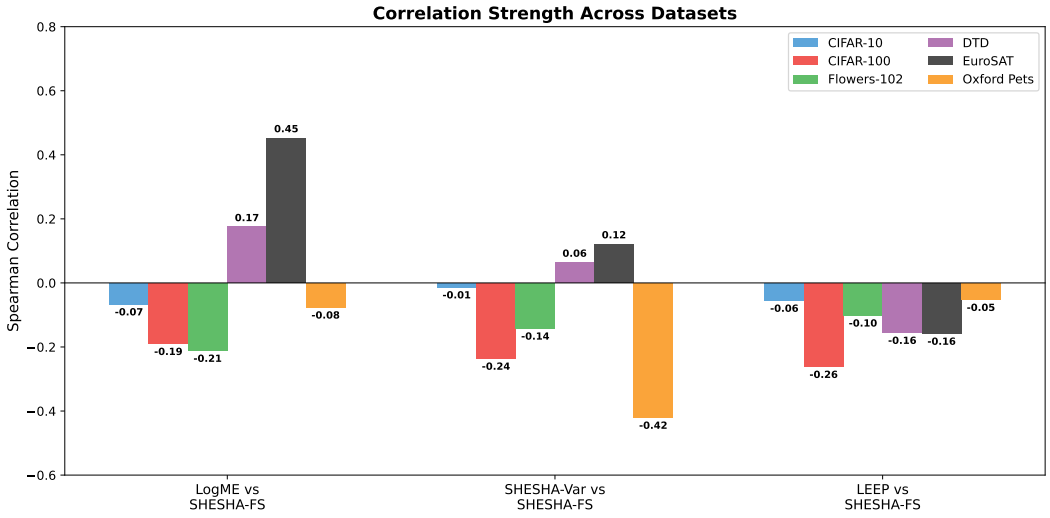

Figure 17: **Cross-Dataset Correlation Comparison.** Spearman correlations between metric pairs across all datasets.

9.2.5 Hierarchical vs. Columnar Transformers

Table 28: **Hierarchical vs. Columnar Transformer Stability.** Mann-Whitney U tests.

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

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

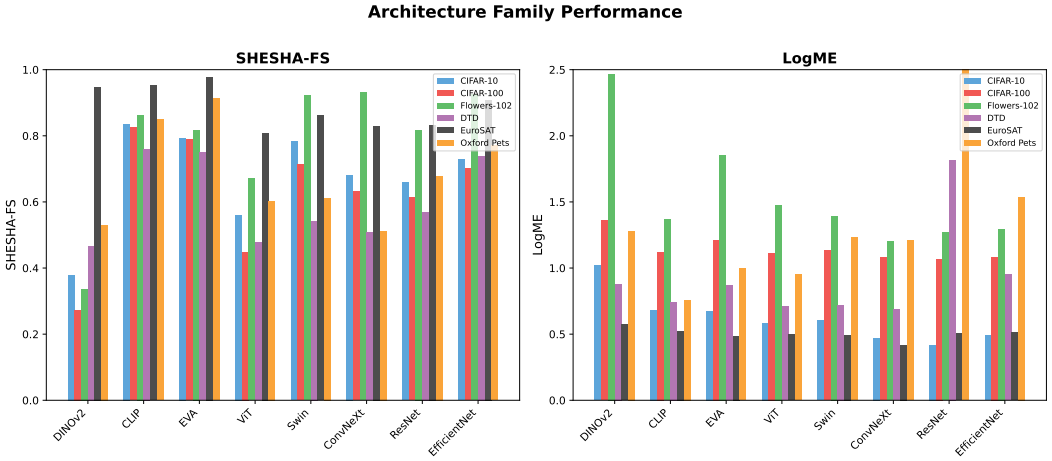

Figure 18: **Architecture Family Performance Across Datasets.** Mean Shesha-Var (left) and Shesha-FS (right) by family.

9.2.6 Cross-Dataset Rank Stability

Table 29: **Cross-Dataset Shesha-FS Rank Correlations.** $n = 93$ common models per pair.

	CIFAR-10	**CIFAR-100**	**Flowers**	**DTD**	**EuroSAT**	**Pets**
CIFAR-10	1.00	0.92	0.56	0.62	0.51	0.46
CIFAR-100	-	1.00	0.50	0.76	0.52	0.55
Flowers	-	-	1.00	0.28	0.20	-0.03
DTD	-	-	-	1.00	0.62	0.78
EuroSAT	-	-	-	-	1.00	0.47
Pets	-	-	-	-	-	1.00

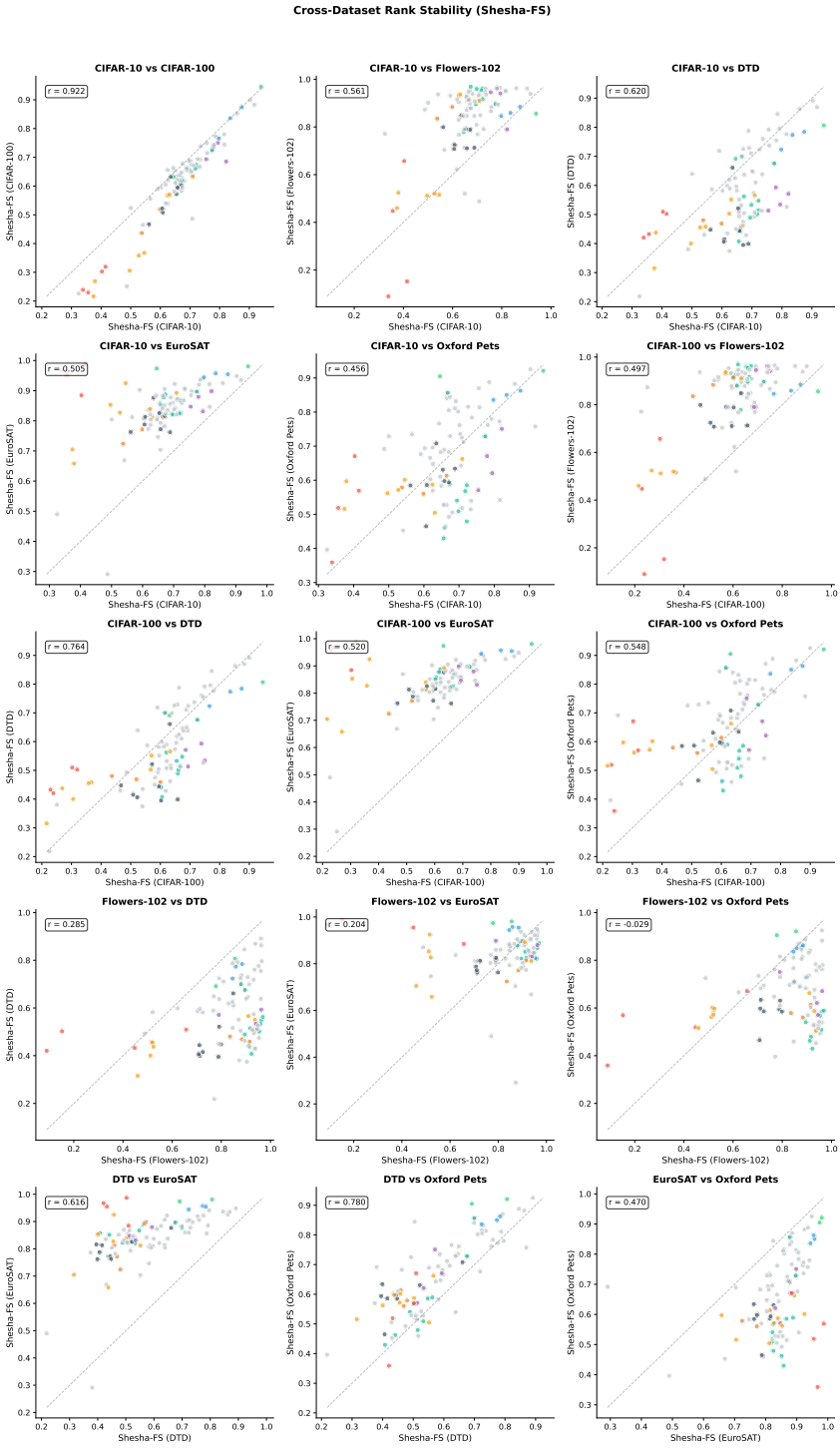

Figure 19: **Cross-Dataset Rank Stability.** Scatter plots comparing Shesha-FS rankings between dataset pairs.

9.2.7 Architecture Family Stability Profiles

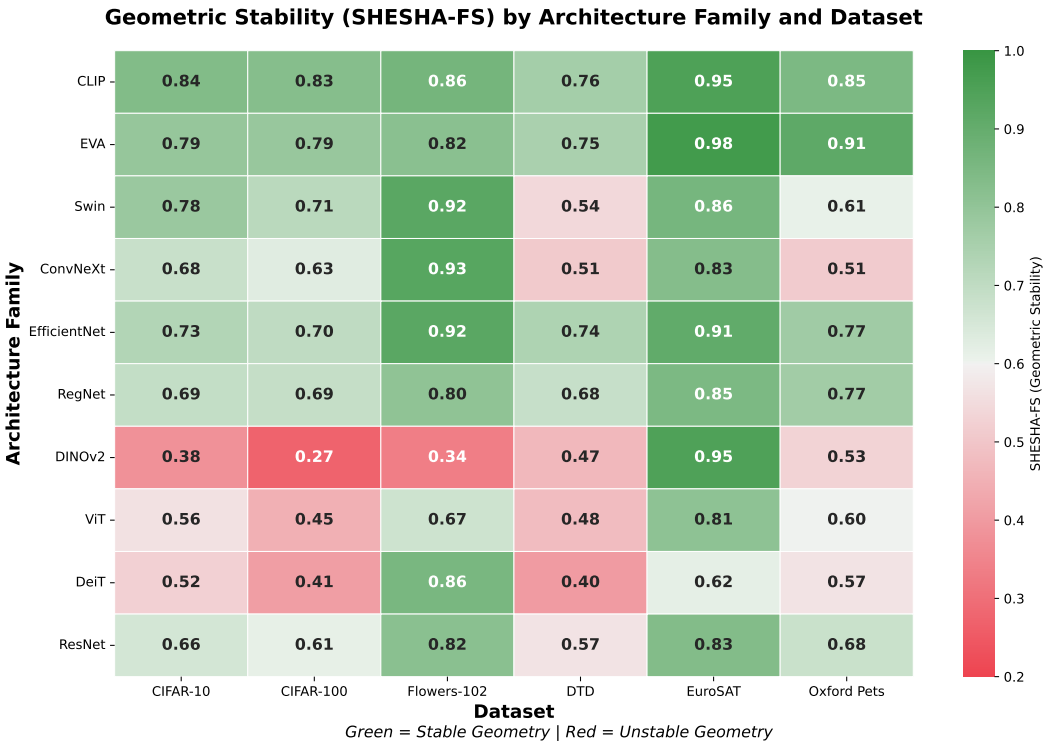

Figure 20: **Geometric Stability Heatmap: Family × Dataset.** Red = unstable, green = stable.

Table 30: **Family Shesha-FS Scores Across Datasets.** Ordered by mean cross-dataset stability.

Family	CIFAR-10	CIFAR-100	Flowers-102	DTD	EuroSAT	Oxford Pets	Mean
Inception	0.88	0.88	0.96	0.88	0.94	0.90	**0.91**
CLIP	0.84	0.83	0.86	0.76	0.95	0.85	**0.85**
EVA	0.79	0.79	0.82	0.75	0.98	0.91	**0.84**
EfficientNetV2	0.73	0.71	0.94	0.76	0.92	0.76	0.80
RegNetY	0.71	0.71	0.83	0.72	0.86	0.80	0.77
Swin	0.79	0.72	0.91	0.55	0.86	0.65	0.75
ConvNeXt	0.69	0.64	0.94	0.50	0.85	0.52	0.69
PVTv2	0.67	0.62	0.92	0.53	0.82	0.58	0.69
ResNet	0.64	0.57	0.76	0.46	0.80	0.60	0.64
DeiT	0.60	0.52	0.88	0.47	0.77	0.58	0.64
ViT	0.54	0.41	0.66	0.46	0.81	0.58	0.58
DINOv2	0.38	0.27	0.34	0.47	**0.95**	0.53	0.49
DeiT3	0.41	0.24	0.82	0.30	0.39	0.54	0.45

9.2.8 Top-20 Rank Agreement

Table 31: **Top-20 Rank Agreement (Jaccard Index) Across Datasets.**

Metric Pair	CIFAR-10	CIFAR-100	Flowers-102	DTD	EuroSAT	Oxford Pets
LEEP vs LogME	0.43	0.43	0.11	0.29	0.14	0.29
LEEP vs Shesha-Var	0.38	0.29	0.18	0.05	0.11	0.11
LEEP vs Shesha-FS	0.08	0.05	0.14	0.11	0.05	0.11
LogME vs Shesha-Var	0.33	0.29	0.60	0.05	0.21	0.11
LogME vs Shesha-FS	0.11	0.05	0.05	0.18	0.29	0.14
Shesha-Var vs Shesha-FS	0.14	0.11	0.05	0.14	0.29	0.08

Models achieving top-20 status for all four metrics: CIFAR-10 (1 model: Swin-Large), CIFAR-100 (0 models), Flowers-102 (1 model: ResNeXt-101), DTD (1 model: ResNeXt-101), EuroSAT (1 model: RegNetY-032), Oxford Pets (0 models). The weakly-supervised ResNeXt-101 appears as a "universal generalist" on 2/6 datasets (Flowers-102, DTD).

9.2.9 Statistical Tests Summary

Table 32: **Statistical Test Summary: Architecture Comparisons on Shesha-FS.**

Dataset	Kruskal-Wallis H	Hier. $>$ Col.	Modern $>$ Classic	CLIP $>$ SSL
CIFAR-10	58.3^{***}	$p = 0.001^{**}$	$p = 0.867$	$p = 0.013^{*}$
CIFAR-100	63.3^{***}	$p < 0.001^{***}$	$p = 0.996$	$p = 0.013^{*}$
Flowers-102	61.6^{***}	$p < 0.001^{***}$	$p = 0.092$	$p = 0.220$
DTD	58.6^{***}	$p = 0.009^{**}$	$p = 1.000$	$p = 0.013^{*}$
EuroSAT	55.2^{***}	$p = 0.795$	$p = 0.850$	$p = 0.261$
Oxford Pets	63.5^{***}	$p = 0.417$	$p = 1.000$	$p = 0.055$

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

9.2.10 Correlation with Embedding Dimension

Table 33: **Correlation with Embedding Dimension.** Spearman ρ values.

Metric	CIFAR-10	CIFAR-100	Flowers-102	DTD	EuroSAT	Oxford Pets
LEEP	0.11	0.10	0.12	0.31^{**}	0.19	0.20
LogME	0.12	0.23^{*}	0.60^{***}	0.79^{***}	0.67^{***}	0.71^{***}
Shesha-Var	-0.29^{**}	-0.35^{***}	-0.01	-0.30^{**}	-0.04	-0.19
Shesha-FS	0.24^{*}	0.24^{*}	0.09	0.42^{***}	0.45^{***}	0.46^{***}

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

9.3 Model Rankings

9.3.1 Top 10 Most Stable Models

Table 34: **Top 10 Shesha-FS: CIFAR-10 and CIFAR-100.**

CIFAR-10			CIFAR-100		
Model	**FS**	**LogME**	**Model**	**FS**	**LogME**
eva02_large	0.94	0.40	eva02_large	0.95	1.10
resnext101_32x8d	0.92	0.51	inception_v3	0.90	1.01
inception_v3	0.90	0.29	resnext101_32x8d	0.88	1.12
vit_large_clip	0.87	0.84	vit_large_clip	0.87	1.17
inception_v4	0.86	0.40	inception_v4	0.85	1.04
vit_base_clip	0.84	0.63	vit_base_clip	0.84	1.10
swin_large	0.82	0.99	tf_efficientnetv2_s	0.78	1.07
swinv2_tiny	0.82	0.46	regnety_032	0.77	1.03
tf_efficientnetv2_s	0.82	0.46	efficientnetv2_rw_m	0.77	1.15
vit_base32_clip	0.80	0.58	vit_base32_clip	0.77	1.09

Table 35: **Top 10 Shesha-FS: Flowers-102, DTD, EuroSAT.**

Flowers-102		DTD		EuroSAT	
Model	**FS**	**Model**	**FS**	**Model**	**FS**
convnext_large	0.97	inception_v3	0.89	vit_giant_dinov2	0.99
efficientnetv2_rw_m	0.96	resnext101_32x8d	0.87	eva02_large	0.98
swin_base	0.96	inception_v4	0.86	eva02_base	0.97
inception_v4	0.96	regnety_032	0.85	vit_large_dinov2	0.97
convnext_base	0.96	eva02_large	0.81	vit_base_clip	0.96
inception_v3	0.96	efficientnetv2_rw_m	0.80	vit_base_dinov2	0.95
tf_efficientnetv2_m	0.96	regnety_064	0.79	vit_large_clip	0.95
convnextv2_base	0.96	wide_resnet101_2	0.79	inception_v3	0.95
convnext_small	0.96	vit_large_clip	0.78	vit_base32_clip	0.94
convnext_tiny	0.95	tf_efficientnetv2_s	0.78	efficientnetv2_rw_m	0.94

Table 36: **Top 10 Shesha-FS: Oxford Pets.**

Model	**Shesha-FS**	**LogME**
inception_v3	0.93	2.80
eva02_large	0.92	0.99
eva02_base	0.90	1.01
regnety_032	0.89	0.89
inception_v4	0.88	1.40
regnety_064	0.87	0.90
vit_large_clip	0.86	0.89
efficientnet_b0	0.86	0.97
vit_base_clip	0.85	0.71
vit_base_dino	0.84	0.80

9.3.2 Bottom 10 Least Stable Models

Table 37: **Bottom 10 Shesha-FS (CIFAR-100).** All four DINOv2 variants appear.

Model	Shesha-FS	LogME
vit_base.augreg_in21k	0.22	1.17
deit3_small	0.23	1.11
vit_base_patch14_dinov2	0.23	1.28
vit_large_patch14_dinov2	0.24	1.41
deit3_base	0.25	1.17
vit_tiny.augreg_ft	0.27	1.00
vit_small_patch14_dinov2	0.30	1.13
vit_base.augreg_ft	0.31	1.21
vit_giant_patch14_dinov2	0.32	1.63
vit_small.augreg_ft	0.36	1.10

9.4 Composite Scores

Table 38: **Top 5 Models by Composite Score (Equal-Weight Z-Score).**

CIFAR-10		CIFAR-100		Flowers-102	
Model	**Score**	**Model**	**Score**	**Model**	**Score**
beit_base	1.21	beit_base	1.01	vit_base.augreg	1.13
swin_large	1.00	swin_large	0.87	beit_base	1.02
vit_large_clip	0.91	vit_large_clip	0.86	inception_v3	0.27
vit_giant_dinov2	0.72	convnextv2_base	0.55	efficientnetv2_rw_m	0.23
efficientnetv2_rw_m	0.54	efficientnetv2_rw_m	0.54	poolformer_m36	0.22

Table 39: **Top 5 Models by Composite Score (Continued).**

DTD		EuroSAT		Oxford Pets	
Model	**Score**	**Model**	**Score**	**Model**	**Score**
inception_v3	1.33	beit_base	1.21	wide_resnet101_2	0.98
wide_resnet101_2	1.22	vit_giant_dinov2	1.06	densenet201	0.94
efficientnetv2_rw_m	0.96	vit_large_dinov2	0.71	resnet152	0.93
wide_resnet50_2	0.80	swin_large	0.62	resnet50.a2	0.93
resnext50_32x4d	0.77	vit_large_clip	0.60	inception_v3	0.91

9.5 Outlier Detection

Table 40: **Consistent Outliers Across Datasets.** Models with $|z| > 2.5$ on multiple datasets.

Model	Metric	Direction	Datasets
beit_base_patch16_224.in22k_ft_in22k	LEEP	HIGH	5/6
eva02_base_patch14_224.mim_in22k	Shesha_Var	HIGH	5/6
vit_base_patch16_224.augreg_in21k	LEEP	HIGH	4/6
vit_giant_patch14_dinov2.lvd142m	LogME	HIGH	4/6
deit3_small_patch16_224.fb_in1k	Shesha_FS	LOW	4/6
vit_large_patch14_dinov2.lvd142m	LogME	HIGH	3/6

9.6 Family Ranking Consistency

Table 41: **Family Shesha-FS Rankings by Dataset.** Rank 1 = highest stability.

Family	CIFAR-10	CIFAR-100	Flowers-102	DTD	EuroSAT	Oxford Pets
Inception	1	1	1	1	4	2
CLIP	2	2	18	2	2	3
EVA02	3	3	22	4	1	1
EfficientNetV2	7	6	4	3	5	10
MobileNetV3	8	9	12	8	6	8
ResNetRS	13	13	9	6	8	5
EfficientNet	9	10	13	9	7	7
RegNetY	10	7	20	5	13	6
ResNeXt	5	4	17	10	11	15
Swin	4	5	10	15	12	16
ResNetV2	14	11	19	11	9	11
SwinV2	6	8	2	19	14	26
DINO	16	19	14	20	19	4
ConvNeXt	12	14	5	21	15	27
WideResNet	26	20	16	7	17	9
PVTv2	17	15	7	16	22	22
RegNetX	19	12	24	12	21	14
CoAtNet	15	18	3	26	16	28
BEiT	11	26	28	22	10	12
MaxViT	23	24	6	17	18	23
ConvNeXtV2	18	17	8	18	25	29
DenseNet	25	21	25	13	20	13
PoolFormer	21	23	11	28	26	17
MAE	20	16	27	14	28	21
DeiT	24	25	15	23	27	20
ResNet	22	22	23	25	24	18
DINOv2	29	28	29	24	3	25
ViT	27	27	26	27	23	19
DeiT3	28	29	21	29	29	24

9.7 Complete Family Statistics: LogME

Table 42: **Complete Architecture Family Statistics: LogME.**

Family	CIFAR-10	CIFAR-100	Flowers-102	DTD	EuroSAT	Oxford Pets
ResNetRS	0.42	1.07	1.32	2.24	0.57	3.05
ResNeXt	0.45	1.09	1.30	2.13	0.47	3.00
WideResNet	0.38	1.05	1.26	2.11	0.49	2.80
DINOv2	1.02	1.36	2.47	0.88	0.57	1.28
ResNetV2	0.50	1.09	1.28	1.70	0.53	2.40
ResNet	0.40	1.06	1.25	1.59	0.50	2.21
Inception	0.35	1.03	1.21	1.40	0.47	2.10
BEiT	0.90	1.22	1.91	0.78	0.57	1.00
EfficientNetV2	0.52	1.10	1.29	1.04	0.51	1.70
EVA02	0.68	1.21	1.86	0.87	0.49	1.00
Swin	0.64	1.15	1.47	0.73	0.50	1.23
DenseNet	0.31	1.01	1.23	1.13	0.49	1.50
ViT	0.67	1.15	1.64	0.73	0.51	0.94
CoAtNet	0.62	1.14	1.22	0.69	0.44	1.25
SwinV2	0.52	1.10	1.25	0.71	0.48	1.23
MaxViT	0.58	1.12	1.20	0.68	0.40	1.24
DeiT3	0.66	1.14	1.17	0.68	0.40	1.17
CLIP	0.68	1.12	1.37	0.74	0.52	0.76
ConvNeXtV2	0.50	1.09	1.21	0.69	0.43	1.24
ConvNeXt	0.46	1.08	1.20	0.69	0.41	1.19
DINO	0.53	1.09	1.30	0.71	0.59	0.80
EfficientNet	0.40	1.04	1.29	0.69	0.52	1.04
PVTv2	0.48	1.07	1.19	0.68	0.41	1.06
PoolFormer	0.38	1.03	1.21	0.69	0.45	0.95
DeiT	0.46	1.06	1.17	0.66	0.39	0.94
MobileNetV3	0.29	1.01	1.29	0.67	0.49	0.85
RegNetY	0.29	1.01	1.23	0.67	0.47	0.81
RegNetX	0.22	0.99	1.18	0.66	0.43	0.73
MAE	0.13	0.95	1.19	0.63	0.52	0.58

10 Representational Drift Detection: Extended Methods and Results

This appendix contains all the methods and additional analyses related to the drift detection experiments described in Section 3.3. The analysis of geometric stability measures whether representational drift can be reliably detected across neural network architectures undergoing various forms of modification through four complementary experiments. First, we establish how the magnitude of geometric change relates to stability during real-world instruction tuning. Second, we test the robustness of our methods under structured perturbations, including quantization, LoRA insertion, and Gaussian noise injection. Third, we validate *predictive validity* through controlled parameter noise injection experiments (canary) on both sentence embedding models. Finally, we extend the canary validation from the third test to causal language models using the same structured perturbations from the second test.

10.1 Experiment 1: Post-Training Drift (Base → Instruct)

Models. We compared representations from 23 base/instruct model pairs spanning 11 model families: Qwen, Llama, SmolLM, SmolLM2, Mistral, StableLM, Gemma, TinyLlama, Pythia, BLOOM, and Falcon. Model sizes ranged from 0.14B to 7B parameters. Table 53 lists all pairs.

Prompt Sets. For each model pair, we embedded four semantically coherent prompt sets, with 50 prompts each, for a total of 200 prompts per model:

- **Factual**: Declarative statements about scientific facts (e.g., "The Earth orbits around the Sun.")
- **Descriptive**: Vivid scene descriptions (e.g., "Waves crashed against the rocky shoreline.")
- **Instructions**: Explanatory requests (e.g., "Explain how photosynthesis works in plants.")
- **Conversational**: Casual dialogue prompts (e.g., "How was your day today?")

All prompts were generated synthetically using large language models. This was done to ensure that each prompt was semantically coherent within its category while maintaining diversity. The complete prompt sets are available in our code repository.

Embedding Extraction. For each model, we extracted hidden states from the final layer (layer -1) using mean pooling over non-padding tokens, followed by L2 normalization. We applied chat templates where available (detected via `tokenizer.chat_template`), with `add_generation_prompt=False` to avoid assistant prefix bias. The maximum sequence length was 256 tokens. For models where `AutoModel` did not return hidden states, we automatically fell back to `AutoModelForCausalLM`.

Metrics. We computed the following representation similarity metrics between base and instruct embeddings. For interpretability, we report *dissimilarity* (i.e., drift magnitude) as $1 -$ similarity, so higher values indicate greater geometric change:

- **Shesha Drift**: $1 - \rho_{\text{Spearman}}(\text{RDM}_{\text{base}}, \text{RDM}_{\text{instruct}})$, where RDM is the pairwise cosine distance vector
- **CKA Drift**: $1 - \text{CKA}_{\text{debiased}}$, using the unbiased HSIC estimator (Song et al., 2012)
- **Procrustes Drift**: $1 -$ Procrustes similarity after optimal rotation, centering, and scaling
- **Wasserstein**: Sliced Wasserstein distance with 100 random projections (inherently a distance metric)
- **MMD**: Maximum Mean Discrepancy with RBF kernel (inherently a distance metric)

All metrics used per-pair-set stable random seeds for reproducibility. Bootstrap confidence intervals (100 resamples) were computed for the main metrics.

Results. Shesha detected **nearly 2× greater geometric change** than CKA on average (25.1% vs 12.8% drift, ratio: 1.96×), with a large amount of variation between families. Of the family of models evaluated, Llama had the greatest difference in performance (**5.23**×), while BLOOM (1.14×) and

Falcon (1.32×) showed near-parity. This suggests that instruction tuning causes some geometric reorganization at the family level, which Shesha can capture in greater detail than CKA.

The advantage of magnitude is present across all prompts: factual (2.37×), descriptive (2.28×), conversational (1.82×), and instructions (1.44×). The lower ratio of instruction-based prompts may be due to the optimization of instruction-tuned models for those specific distributions. This supports the idea that instruction tuning optimizes the representation space to reflect the way instruction-following semantics are structured in relation to instructions, effectively "straightening" the manifold for this specific task while producing more extensive geometric reorganizations for out-of-distribution input prompts.

Table 43: Geometric drift between base and instruction-tuned model pairs. Drift metrics (as percentages) averaged across four prompt types (factual, descriptive, instructions, conversational). Higher values indicate greater representational change from instruction tuning.

Model Pair	**Size**	**Shesha (%)**	**CKA (%)**	**Procrustes (%)**
SmolLM-135M → Instruct	135M	28.3	18.3	13.4
SmolLM2-135M → Instruct	135M	29.6	12.8	10.0
SmolLM-360M → Instruct	360M	41.6	24.5	17.5
SmolLM2-360M → Instruct	360M	30.5	19.1	14.1
Qwen2-0.5B → Instruct	0.5B	10.9	4.9	2.9
Qwen1.5-0.5B → Chat	0.5B	14.2	6.0	4.4
BLOOM-560M → BLOOMZ	560M	35.5	31.6	23.1
Pythia-1B → Deduped	1.0B	10.6	3.5	2.6
Llama-3.2-1B → Instruct	1.0B	34.9	6.7	4.2
TinyLlama-1.1B → Chat	1.1B	18.7	4.7	3.1
BLOOM-1.1B → BLOOMZ	1.1B	10.3	8.6	5.8
Qwen2-1.5B → Instruct	1.5B	13.0	5.0	22.3
StableLM-1.6B → Zephyr	1.6B	23.0	11.8	19.4
SmolLM-1.7B → Instruct	1.7B	30.4	19.6	26.2
SmolLM2-1.7B → Instruct	1.7B	35.2	19.6	28.0
Qwen1.5-1.8B → Chat	1.8B	21.1	7.2	15.1
Gemma-2B → IT	2.0B	32.4	14.6	22.0
Gemma-2-2B → IT	2.0B	41.4	22.5	25.6
Llama-3.2-3B → Instruct	3.0B	33.1	6.3	13.8
Qwen1.5-4B → Chat	4.0B	15.5	6.4	13.8
Qwen2-7B → Instruct	7.0B	15.6	8.6	17.5
Mistral-7B → Instruct	7.0B	18.5	6.5	14.1
Falcon-7B → Instruct	7.0B	34.1	25.8	26.8

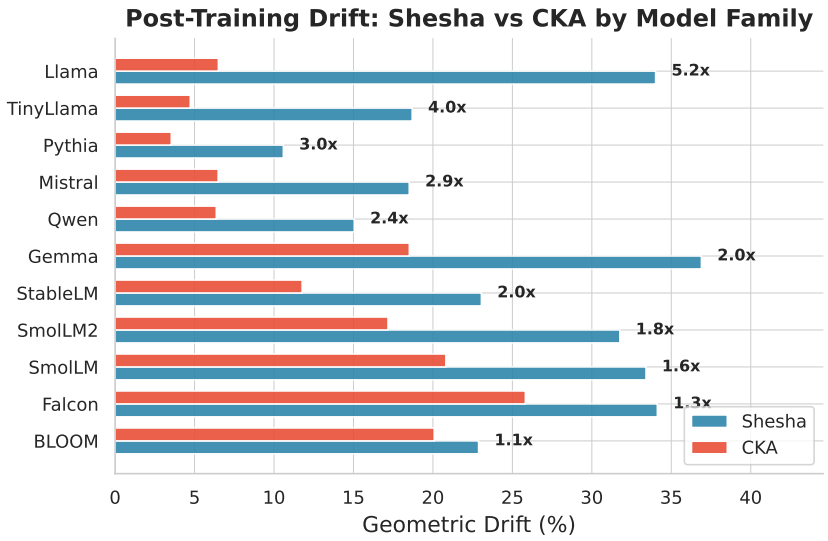

Figure 21: **Post-training drift varies substantially across model families.** Geometric drift between base and instruction-tuned model pairs, measured by Shesha and CKA, aggregated by model family (23 pairs total). The Shesha/CKA ratio ranges from 1.1× (BLOOM) to 5.2× (Llama), indicating that Shesha consistently detects greater representational reorganization than CKA. Families with larger ratios exhibit more distributed geometric changes that CKA's top-principal-component weighting underestimates. This pattern is consistent across model scales within each family.

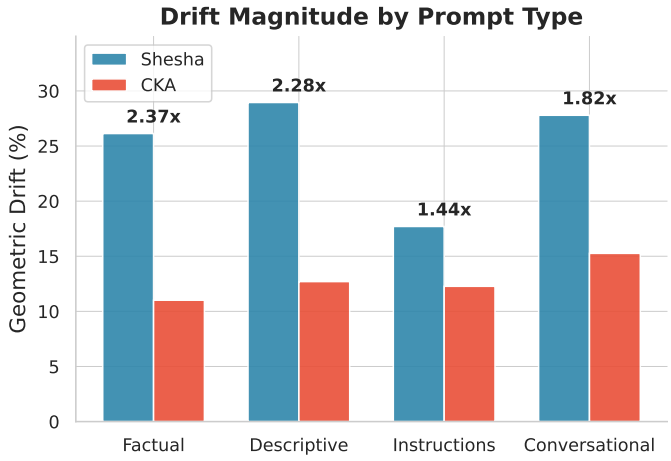

Figure 22: **Drift magnitude varies by prompt type.** Mean geometric drift across 23 base-instruct pairs, stratified by prompt category. Factual and descriptive prompts induce the largest Shesha/CKA ratios (2.37× and 2.28×), while instruction prompts show the smallest ratio (1.44×). This pattern suggests that instruction tuning most strongly reshapes representations for instruction-following inputs (reducing the Shesha-CKA gap) while introducing greater reorganization for out-of-distribution prompt types where the tuning objective provides less direct supervision.

10.2 Experiment 2: Structured Perturbation Analysis

This experiment measures representational drift under three types of structured perturbations that simulate real-world model modifications: quantization, LoRA adapter insertion, and Gaussian noise injection. Unlike the base-to-instruct comparison (Experiment 1), these perturbations allow for specific control over the magnitude of the perturbations.

Models. We evaluated 16 causal language models the covered multiple families and scales: SmolLM (135M, 360M, 1.7B), SmolLM2 (135M, 360M, 1.7B), Qwen2 (0.5B, 1.5B, 7B), Llama-3.2 (1B, 3B), TinyLlama (1.1B), StableLM-2 (1.6B), Gemma (2B), Gemma-2 (2B), and Mistral-7B. These models were selected because they support both quantization through `bitsandbytes` and LoRA through the `peft` library (Mangrulkar et al., 2022).

Prompt Sets. We used the same four sets of prompts as in Experiment 1 (factual, descriptive, instructions, conversational), with 50 prompts each, totaling 200 total prompts per model.

Embedding Extraction. For each model, we extracted hidden states from the final layer using mean pooling over non-padding tokens, followed by L2 normalization. The maximum sequence length was 256 tokens. All models were loaded using `AutoModelForCausalLM` with appropriate padding side detection based on model architecture.

Perturbation Protocols.

Quantization. We compared FP16 (baseline) representations against INT8 and INT4 (NF4) quantized versions using bitsandbytes:

- **FP16**: Baseline precision (no quantization)
- **INT8**: 8-bit integer quantization via `load_in_8bit=True`
- **INT4 (NF4)**: 4-bit NormalFloat quantization via `load_in_4bit=True` with `bnb_4bit_quant_type="nf4"`

For each quantization level, we embedded all prompt sets and computed drift metrics relative to the FP16 baseline.

LoRA Adapter Insertion. We applied randomly initialized LoRA adapters (Hu et al., 2022) to attention projection layers and measured representational drift as a function of adapter rank and initialization scale:

- **Ranks tested**: $r \in \{1, 2, 4, 8, 16, 32, 64\}$ with $\alpha = 2r$ and fixed initialization scale 0.01. This fixed initialization scale ensures that the per-parameter variance remains constant, isolating the effect of topological complexity (rank) from raw magnitude.
- **Initialization scales tested**: $s \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1\}$ at fixed rank $r = 8$
- **Target modules**: Architecture-specific attention projections (e.g., `q_proj`, `v_proj`, `k_proj`, `o_proj` for Llama-style models)
- **Initialization**: LoRA-A matrices initialized with Kaiming-style scaling; LoRA-B matrices initialized with $\mathcal{N}(0, s^2)$

This protocol simulates the representational impact of LoRA fine-tuning at various capacity levels without actual training.

Gaussian Noise Injection. We injected Gaussian noise into the model parameters at 10 levels: $\alpha \in \{0.00, 0.01, 0.02, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50\}$:

- For each parameter tensor θ, noise was added: $\theta' = \theta + \mathcal{N}(0, \alpha \cdot \text{std}(\theta))$
- Clean weights were restored between noise levels to ensure independence
- Each (model, α) combination used a deterministic seed for reproducibility

This protocol simulates parameter corruption resulting from quantization error, bit rot, or fine-tuning drift, allowing for unambiguous control over perturbation magnitude.

Metrics. We computed the following metrics between clean/baseline and perturbed embeddings:

- **Shesha**: Spearman correlation of RDM vectors (reported as $1 - \rho$ for drift)
- **RDM-Pearson**: Pearson correlation of RDM vectors

- **CKA**: Debiased linear CKA
- **Procrustes**: Procrustes similarity after optimal orthogonal alignment
- **Wasserstein**: Sliced Wasserstein distance (100 projections)
- **MMD**: Maximum Mean Discrepancy with RBF kernel
- **Subspace Overlap**: Mean squared cosine of principal angles at $k \in \{5, 10, 20\}$
- **Eigenspectrum Similarity**: Cosine similarity of normalized singular value vectors
- **Participation Ratio**: Effective dimensionality measure

Bootstrap confidence intervals (100 resamples) were computed for the primary metrics (Shesha, CKA, Procrustes).

Results.

Gaussian Noise. Table 44 shows the mean drift across all models and prompt types as a function of noise level. As expected, all metrics increased monotonically with the magnitude of noise. At low noise levels ($\sigma \leq 0.05$), Procrustes demonstrated the highest sensitivity (0.085 at $\sigma = 0.05$), followed by Shesha (0.049) and CKA (0.032). At higher noise levels, Shesha exhibited the largest drift values, reaching 0.716 at $\sigma = 0.5$, compared to 0.430 for CKA and 0.414 for Procrustes. This pattern confirms that Procrustes detects early displacement while Shesha captures the full extent of geometric reorganization.

Table 44: Mean drift by Gaussian noise level (averaged across 16 models and 4 prompt types).

Noise Level (σ)	Shesha	CKA	Procrustes
0.01	0.003	0.002	0.019
0.02	0.012	0.007	0.037
0.05	0.049	0.032	0.085
0.10	0.119	0.074	0.141
0.15	0.225	0.146	0.206
0.20	0.361	0.238	0.278
0.30	0.630	0.380	0.362
0.40	0.714	0.432	0.396
0.50	0.716	0.430	0.414

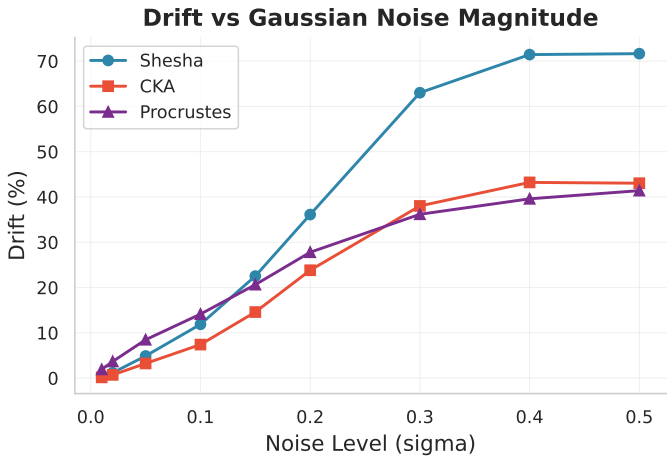

Figure 23: **Metric response to Gaussian noise perturbation.** Mean drift across 16 causal LMs as noise magnitude increases ($\sigma \in [0, 0.5]$). Shesha exhibits the steepest response curve, reaching 71% drift at $\sigma = 0.5$ compared to 43% for CKA and 42% for Procrustes. At low noise levels ($\sigma < 0.1$), Procrustes shows elevated sensitivity relative to Shesha and CKA, foreshadowing the false alarm behavior characterized in Experiment 4. The divergence between metrics grows with perturbation magnitude, with Shesha capturing approximately 1.7× more drift than CKA at high noise levels.

Quantization. Table 45 shows drift induced by quantization. INT8 quantization caused minimal representational drift (Shesha: 2.1%, CKA: 1.4%, Procrustes: 5.3%), while INT4 quantization induced approximately 3× more drift (Shesha: 6.2%, CKA: 4.2%, Procrustes: 9.7%). Procrustes showed the highest sensitivity to quantization effects. However, given the minimal functional degradation observed in Experiment 4 under quantization, this heightened sensitivity may largely reflect rigid geometric rotations rather than functional erosion.

Table 45: Mean drift by quantization level (averaged across 16 models and 4 prompt types).

Quantization	Shesha	CKA	Procrustes
INT8	0.021	0.014	0.053
INT4 (NF4)	0.062	0.042	0.097

LoRA. Table 46 shows drift as a function of LoRA rank at a fixed initialization scale (0.01). Drift increased monotonically with rank, from near-zero at $r = 1$ to substantial reorganization at $r = 64$ (Shesha: 15.3%, CKA: 8.8%, Procrustes: 13.9%). Table 47 shows drift as a function of initialization scale at a fixed rank ($r = 8$). The initialization scale had a dramatic effect: increasing from 0.001 to 0.1 increased Shesha drift from 0.06% to 44.2%, demonstrating that the magnitude of LoRA perturbation, not just the rank, determines representational impact.

Table 46: Mean drift by LoRA rank at fixed initialization scale (0.01).

LoRA Rank	Shesha	CKA	Procrustes
1	0.002	0.001	0.017
2	0.004	0.002	0.022
4	0.047	0.035	0.055
8	0.029	0.016	0.048
16	0.064	0.036	0.075
32	0.086	0.052	0.102
64	0.153	0.088	0.139

Table 47: Mean drift by LoRA initialization scale at fixed rank ($r = 8$).

Init Scale	Shesha	CKA	Procrustes
0.001	0.001	0.000	0.010
0.005	0.006	0.003	0.023
0.01	0.029	0.015	0.047
0.02	0.095	0.050	0.106
0.05	0.303	0.176	0.237
0.10	0.441	0.324	0.394

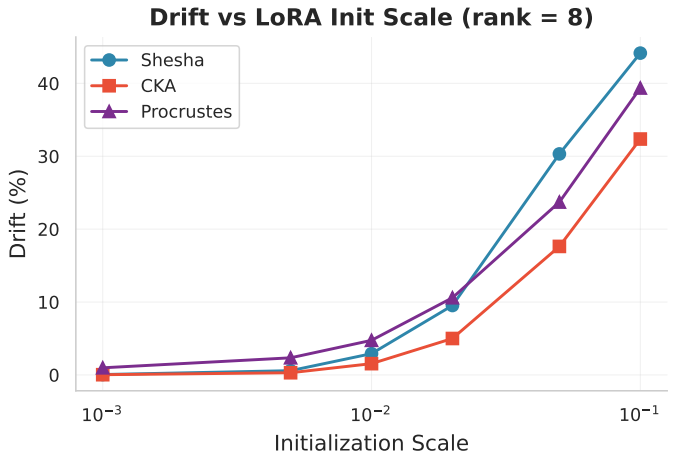

Figure 24: **Drift scales with LoRA initialization magnitude.** Mean drift across 16 causal LMs as LoRA initialization scale increases (rank fixed at 8). All metrics exhibit exponential growth with init scale, but Procrustes maintains a consistent offset above Shesha and CKA across the entire range. At minimal perturbation (init scale = 10^{-3}), Procrustes already registers detectable drift while Shesha and CKA remain near zero, consistent with Procrustes's sensitivity to rigid geometric transformations that do not affect functional behavior. At init scale = 10^{-1}, Shesha reaches 44% drift compared to 33% for CKA, reflecting Shesha's greater sensitivity to distributed representational changes.

10.3 Experiment 3: Canary Validation (Sentence Embedding Models)

This experiment validates that geometric drift metrics have *predictive validity* for downstream task performance by using sentence embedding models with controlled parameter noise injection.

Models. We evaluated 26 sentence embedding models spanning multiple architectures: MiniLM variants (L3, L6, L12), MPNet, DistilBERT, DistilRoBERTa, BERT, RoBERTa, ALBERT, GTE (small/base/large), E5 (small/base/large), BGE (small/base/large), and SimCSE (supervised/unsupervised). These models were chosen because they produce well-calibrated sentence embeddings that are suitable for downstream classification.

Noise Injection Protocol. For each model, we:

1. Saved the clean model weights
2. Injected Gaussian noise at 51 levels: $\alpha \in \{0.00, 0.01, 0.02, \ldots, 0.50\}$
3. For each parameter tensor θ, added noise: $\theta' = \theta + \mathcal{N}(0, \alpha \cdot \text{std}(\theta))$
4. Embedded 800 SST-2 validation samples (balanced classes)
5. Computed drift metrics and downstream classification accuracy (5-fold CV, logistic regression)
6. Restored clean weights before next noise level

This protocol simulates parameter corruption from quantization errors, bit rot, or fine-tuning drift. Each (model, α) combination used a deterministic seed for reproducibility across runs.

Embedding Details. For SentenceTransformer models, we used the native `encode()` method. For models loaded with AutoModel, we applied mean pooling over the last hidden layer with attention masking, followed by L2 normalization. E5 models received the "passage:" prefix as specified in their documentation.

Metrics. We computed five drift metrics between clean and noisy embeddings:

- **Shesha**: $1 - \rho_{\text{Spearman}}(\text{RDM}_{\text{clean}}, \text{RDM}_{\text{noisy}})$
- **RDM-Pearson**: $1 - r_{\text{Pearson}}(\text{RDM}_{\text{clean}}, \text{RDM}_{\text{noisy}})$
- **CKA**: $1 - \text{CKA}_{\text{debiased}}$
- **Procrustes**: $1 -$ Procrustes similarity after optimal orthogonal alignment
- **Wasserstein**: Sliced Wasserstein distance with 100 random projections

Results: Predictive Validity. Shesha ($\rho = 0.927$), CKA ($\rho = 0.937$), and Procrustes ($\rho = 0.935$) all achieved nearly identical Spearman correlations with accuracy drop, confirming **equivalent predictive validity** across these three metrics. RDM-Pearson showed similar performance ($\rho = 0.928$). In contrast, Wasserstein distance showed a significantly weaker correlation ($\rho = 0.761$), failing to register drift in most models until catastrophic collapse ($\sigma \geq 0.45$).

Results: Early Warning. Using a detection threshold of 0.05, Procrustes provided the **earliest warning across all models**, detecting drift at a mean noise level of $\sigma = 0.040$. Shesha detected at $\sigma = 0.123$ (second earliest), followed by CKA at $\sigma = 0.136$. Procrustes detected earlier than both Shesha and CKA in **100% of models** (26/26). Shesha detected earlier than CKA in **73% of models** (19/26), with 27% tied (7/26). CKA never detected first. Wasserstein failed to reach the detection threshold in 24 of 26 models.

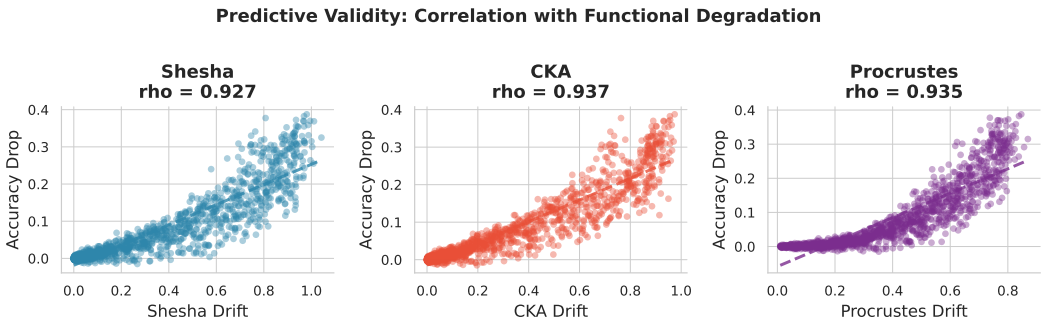

Figure 25: **Predictive validity of geometric drift metrics.** Scatter plots showing the relationship between drift magnitude and functional degradation (accuracy drop) across 26 sentence embedding models under Gaussian noise perturbation ($\sigma \in [0.01, 0.5]$). All three metrics exhibit strong correlation with accuracy loss: Shesha ($\rho = 0.927$), CKA ($\rho = 0.937$), and Procrustes ($\rho = 0.935$). Each point represents one model at one noise level; dashed lines show linear fits. The consistently high correlation that confirms the validity of this measurement from an empirical standpoint, and serves as the groundwork for using geometric drift as a reliable proxy for functional degradation as a basis to assess performance of models over time.

Table 48: Gaussian noise perturbation on sentence embedding models (26 models, STS-B task). Drift metrics show strong correlation with accuracy drop ($\rho > 0.92$ for all metrics).

Noise (σ)	Acc. Drop (%)	Shesha	CKA	Procrustes
0.00	0.0	0.000	0.000	0.000
0.01	0.0	0.001	0.000	0.015
0.02	0.0	0.002	0.002	0.031
0.03	0.0	0.005	0.003	0.046
0.04	0.0	0.008	0.006	0.061
0.05	0.0	0.012	0.009	0.076
0.06	0.2	0.019	0.014	0.092
0.07	0.1	0.025	0.018	0.107
0.08	0.3	0.033	0.024	0.122
0.09	0.3	0.042	0.031	0.138
0.10	0.4	0.054	0.039	0.154
0.15	1.4	0.115	0.090	0.232
0.20	2.8	0.208	0.171	0.315
0.25	5.0	0.329	0.271	0.394
0.30	9.4	0.488	0.411	0.484
0.35	12.9	0.603	0.543	0.565
0.40	16.9	0.705	0.657	0.631
0.45	21.0	0.771	0.728	0.683
0.50	22.5	0.813	0.772	0.716

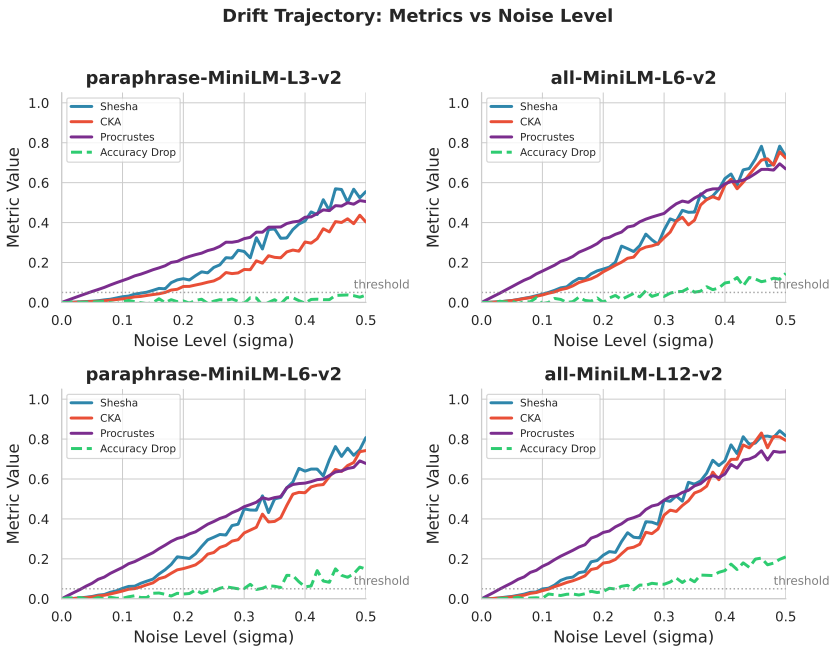

Figure 26: **Drift trajectories across noise levels.** Evolution of Shesha, CKA, Procrustes, and accuracy drop as Gaussian noise magnitude increases ($\sigma \in [0, 0.5]$) for four representative sentence embedding models. The horizontal dashed line indicates a 5% detection threshold. Procrustes consistently exceeds this threshold at lower noise levels than Shesha or CKA, illustrating its heightened sensitivity to geometric perturbations. All metrics track accuracy degradation, but Procrustes's early activation in low-noise regimes where accuracy remains stable contributes to its elevated false alarm rate.

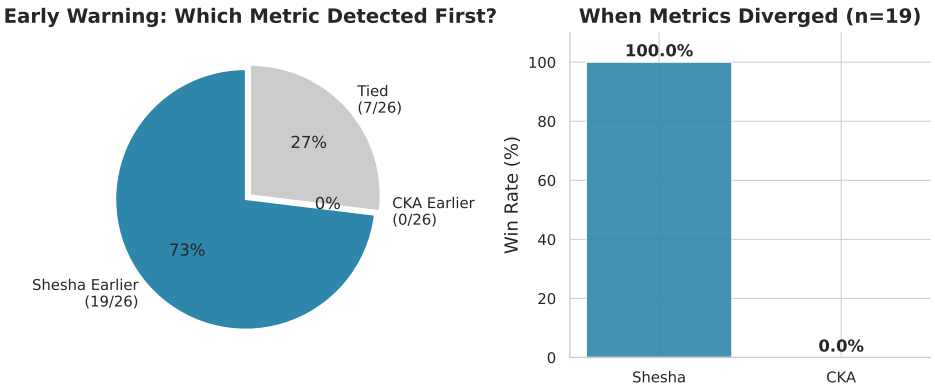

Figure 27: **Shesha provides earlier warning than CKA.** (Left) Distribution of which metric first exceeded the 5% detection threshold across 26 sentence embedding models. Shesha detected drift earlier in 73% of cases (19/26), with the remaining 27% tied; CKA never detected first. (Right) Among the 19 cases where metrics diverged, Shesha achieved a 100% win rate. This early warning advantage stems from Shesha's equal weighting of all pairwise relationships, allowing it to detect subtle structural reorganization before it manifests in the dominant principal components that drive CKA.

10.4 Experiment 4: Extended Canary Validation (Causal Language Models)

This experiment extends the canary validation test from Experiment 3 to causal language models, testing whether geometric drift metrics maintain predictive validity under the structured perturbations from Experiment 2 (quantization, LoRA, Gaussian noise) rather than just random noise.

Models. We evaluated 15 causal language models: SmolLM (135M, 360M, 1.7B), SmolLM2 (135M, 360M), Qwen2 (0.5B, 1.5B, 7B), Llama-3.2 (1B, 3B), TinyLlama (1.1B), StableLM-2 (1.6B), Gemma (2B), Gemma-2 (2B), and Mistral-7B. These models span 0.14B to 7B parameters.

Dataset. We used the SST-2 sentiment classification task (800 samples, balanced classes) for downstream evaluation. This enables direct comparison with Experiment 3 despite testing on a fundamentally different model class.

Perturbation Protocols.

Gaussian Noise. We applied the same levels of Gaussian noise as in Experiment 3. 51 noise levels from $\alpha \in \{0.00, 0.01, \ldots, 0.50\}$ with per-parameter scaling by standard deviation were applied.

Quantization. We compared the FP16 baseline against INT8 and INT4 (NF4) quantization:

- For each model, SST-2 samples were embedded at each precision level
- Drift metrics were computed relative to the FP16 baseline.
- Downstream accuracy was evaluated at each precision level

LoRA. We applied randomly initialized LoRA adapters (Hu et al., 2022) and measured both drift and accuracy impact:

- **Varying rank**: $r \in \{1, 2, 4, 8, 16, 32, 64\}$ with $\alpha = 2r$, initialization scale 0.01
- **Varying initialization scale**: $s \in \{0.001, 0.01, 0.05, 0.1\}$ at a fixed rank of $r = 8$
- Target modules were selected based on their architecture (attention projections)

Embedding Extraction. Hidden states from the final layer, mean-pooled over non-padding tokens, L2 normalized. The maximum sequence length was 128 tokens.

Metrics. We computed drift metrics (Shesha, CKA, Procrustes, Wasserstein) and additional geometric metrics:

- **Subspace Overlap**: At $k \in \{5, 10, 20\}$ principal components
- **Eigenspectrum Similarity**: Cosine similarity of normalized singular value vectors
- **Effective Rank**: Spectral entropy-based dimensionality measure
- **Effective Rank Ratio**: The ratio of effective ranks measures dimensionality collapse or expansion.

Results: Gaussian Noise. Table 49 shows the correlation between drift metrics and accuracy degradation. All primary metrics achieved strong predictive validity ($\rho > 0.9$), with Shesha ($\rho = 0.915$), CKA ($\rho = 0.912$), and Procrustes ($\rho = 0.903$) performing comparably. For early warning, Procrustes again detected earliest (mean $\sigma = 0.041$), followed by CKA ($\sigma = 0.117$) and Shesha ($\sigma = 0.120$).

Table 49: Predictive validity: Spearman correlation with accuracy drop (Gaussian noise on causal LMs).

Metric	Correlation (ρ)	Detection Threshold (σ)
Shesha	0.915	0.120
CKA	0.912	0.117
Procrustes	0.903	0.041
Wasserstein	0.859	-

Results: Quantization. Quantization induced minimal accuracy degradation (mean accuracy drop: 0.01% for INT8, 0.21% for INT4), making correlation analysis uninformative. However, the fact that Shesha detected clear drift (3.6% for INT4) while accuracy remained stable suggests that quantization induces a form of "lossless" geometric compression that preserves functional topology. The drift metrics successfully detected these representational changes: INT8 induced a 0.74% Shesha drift and a 6.3% Procrustes drift, while INT4 induced a 3.6% Shesha drift and a 13.1% Procrustes drift. This demonstrates that geometric drift can occur without immediate functional consequences, underscoring the importance of monitoring representational changes as leading indicators.

Results: LoRA. Table 52 shows the drift and accuracy degradation associated with the LoRA configuration. Drift metrics showed a moderate correlation with accuracy drop ($\rho \approx 0.67$-0.68), which is lower than that for Gaussian noise but still meaningful. Notably, accuracy degradation scaled with both rank and initialization scale, with large perturbations (rank 64 or init scale 0.1) causing accuracy drops of 5-18%.

Table 50: LoRA canary results by rank (fixed init scale 0.01).

Rank	Shesha	Acc Drop
1	0.003	0.001
2	0.005	0.001
4	0.015	0.004
8	0.042	0.013
16	0.074	0.023
32	0.102	0.031
64	0.185	0.050

Table 51: LoRA canary results by initialization scale (fixed rank 8).

Init Scale	Shesha	Acc Drop
0.001	0.000	0.001
0.01	0.042	0.013
0.05	0.374	0.108
0.10	0.611	0.181

Table 52: LoRA perturbation results across ranks and initialization scales. Drift metrics show moderate correlation with accuracy drop ($\rho \approx 0.67$-0.68). Large perturbations (rank 64 or init scale 0.1) cause substantial accuracy degradation.

Config	Value	Acc. Drop (%)	Shesha	CKA	Procrustes
By Rank (init scale = 0.01)					
Rank	1	0.13 ± 0.27	0.003 ± 0.007	0.003	0.032
	2	0.09 ± 0.35	0.005 ± 0.008	0.005	0.044
	4	0.41 ± 0.98	0.015 ± 0.032	0.014	0.064
	8	1.30 ± 3.76	0.042 ± 0.117	0.042	0.095
	16	2.34 ± 7.72	0.074 ± 0.193	0.076	0.135
	32	3.07 ± 7.36	0.102 ± 0.191	0.093	0.179
	64	5.01 ± 8.84	0.185 ± 0.260	0.175	0.255
By Init Scale (rank = 8)					
Init Scale	0.001	0.08 ± 0.20	0.000 ± 0.000	0.000	0.015
	0.01	1.30 ± 3.76	0.042 ± 0.117	0.042	0.095
	0.05	10.77 ± 11.08	0.374 ± 0.340	0.352	0.401
	0.1	18.05 ± 11.80	0.611 ± 0.361	0.587	0.589

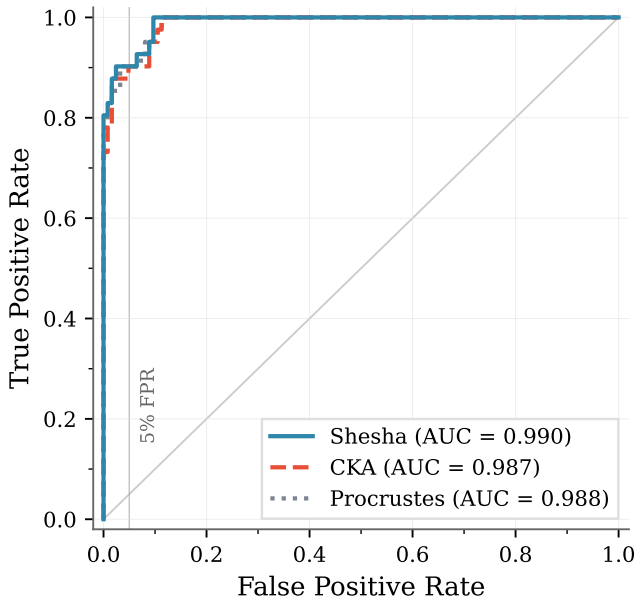

Figure 28: ROC analysis for drift detection on the LoRA perturbation benchmark. All metrics achieve high overall performance (AUC > 0.98), but Shesha provides superior sensitivity at low false alarm rates. At the operationally relevant 5% FPR threshold (vertical line), Shesha maintains 90.2% sensitivity compared to 85.4% for Procrustes, confirming that Shesha's earlier detection reflects genuine signal rather than noise susceptibility. The ground truth is defined as functional degradation $> 1\%$ accuracy drop.

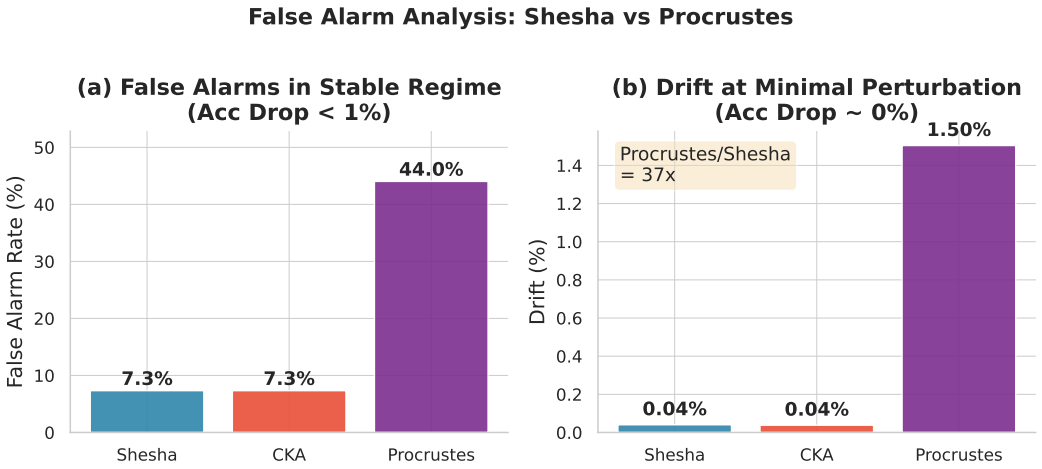

Figure 29: **False alarm analysis reveals Procrustes oversensitivity.** (a) In the stable regime (accuracy drop $< 1\%$), Procrustes triggers false alarms in 44% of cases compared to only 7.3% for Shesha and CKA (a $6\times$ difference). (b) At minimal perturbation where functional performance is unchanged, Procrustes reports 1.50% drift versus 0.04% for Shesha (a $37\times$ inflation). This demonstrates that Procrustes detects rigid geometric transformations that do not affect model behavior, making it unsuitable as a primary monitoring metric despite comparable predictive validity. Shesha achieves the optimal balance between sensitivity and specificity.

10.5 Metric Comparison and Complementary Roles

These experiments reveal that each metric captures a distinct aspect of representational change, with important implications for practical monitoring.

The False Alarm Problem with Procrustes. Although Procrustes proved to be very sensitive to changes in geometry and was able to detect drift with a smaller amount of perturbation ($\sigma = 0.04$ compared to $\sigma = 0.12$ for Shesha), this high sensitivity is not necessarily useful, as it results in a large number of false alarms. In the LoRA canary experiments (Experiment 4), when models remained functionally stable (accuracy drop $< 1\%$), Procrustes triggered the $p = 0.05$ detection threshold in **44% of cases** compared to 7.3% for Shesha (which was identical to the more conservative CKA baseline). This represents a **6× higher false alarm ratio**. At the lowest perturbation level (init_scale = 0.001), where accuracy was essentially unchanged (ΔAcc $\sim 0.08\%$), Procrustes registered **37× more drift** than Shesha (0.015 vs. 0.0004). This indicates that Procrustes detects rigid geometric transformations, such as rotations and translations, that do not affect the model's functional behavior.

Shesha as the Functional Canary. Shesha achieved the optimal balance between sensitivity and reliability. Unlike Procrustes, it effectively filtered non-functional geometric noise. In the stable regime, Shesha's false alarm rate was only 7.3% compared to Procrustes' 44%. Yet, unlike CKA, which is dominated by top principal components, Shesha weights all pairwise relationships equally, allowing it to detect subtle structural reorganization in the representation manifold. In the Gaussian noise canary (Experiment 3), Shesha achieved the highest correlation with downstream accuracy degradation ($\rho = 0.927$), slightly outperforming CKA ($\rho = 0.937$) and Procrustes ($\rho = 0.935$). At high perturbation levels (Table 44), Shesha captured the largest drift magnitude (0.716 at $\sigma = 0.5$), revealing geometric reorganization that CKA (0.430) and Procrustes (0.414) underestimated.

CKA as the Stability Anchor. CKA's insensitivity to minor perturbations, which initially appears as a limitation, can provide value as a conservative indicator of functional integrity. Because CKA is dominated by the top principal components, it remains stable under geometric transformations that preserve the dominant structure. When CKA drops substantially, the core structure of the representation has changed, and functional degradation is likely imminent. CKA and Shesha showed nearly identical false alarm rates (7.3%) in the stable regime. However, CKA's lower sensitivity to high-frequency geometric changes makes it into a dependable baseline measure "floor" indicator.

Practical Monitoring Framework. Based on these findings, we recommend a multi-tiered monitoring approach for representational drift:

1. **Use Shesha as the primary drift metric.** It provides the best combination of predictive validity ($\rho \geq 0.92$) and low false alarm rate (7%), detecting functionally relevant geometric changes while ignoring harmless rigid transformations.

2. **Use Procrustes for maximum sensitivity when false alarms are acceptable.** In scenarios where any geometric change warrants investigation (e.g., security-critical deployments), Procrustes provides the earliest possible warning, but expect 6$\times$ more false positives.

3. **Use CKA as a confirmation signal.** When Shesha triggers, check CKA to assess whether the drift has affected the dominant representation structure. If CKA remains stable, the perturbation may be recoverable; if CKA has also dropped, functional degradation is likely.

4. **Avoid Wasserstein for drift detection.** Sliced Wasserstein distance proved insufficiently sensitive, failing to detect drift until catastrophic collapse in most models.

10.6 Model Lists

Table 53: Base/Instruct model pairs for post-training drift analysis (Experiment 1).

Base Model	Instruct Model	Params
HuggingFaceTB/SmolLM-135M	SmolLM-135M-Instruct	0.14B
HuggingFaceTB/SmolLM2-135M	SmolLM2-135M-Instruct	0.14B
HuggingFaceTB/SmolLM-360M	SmolLM-360M-Instruct	0.36B
HuggingFaceTB/SmolLM2-360M	SmolLM2-360M-Instruct	0.36B
Qwen/Qwen2-0.5B	Qwen2-0.5B-Instruct	0.5B
Qwen/Qwen1.5-0.5B	Qwen1.5-0.5B-Chat	0.5B
bigscience/bloom-560m	bloomz-560m	0.56B
EleutherAI/pythia-1b	pythia-1b-deduped	1.0B
meta-llama/Llama-3.2-1B	Llama-3.2-1B-Instruct	1.0B
TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T	TinyLlama-1.1B-Chat-v1.0	1.1B
bigscience/bloom-1b1	bloomz-1b1	1.1B
Qwen/Qwen2-1.5B	Qwen2-1.5B-Instruct	1.5B
stabilityai/stablelm-2-1_6b	stablelm-2-zephyr-1_6b	1.6B
HuggingFaceTB/SmolLM-1.7B	SmolLM-1.7B-Instruct	1.7B
HuggingFaceTB/SmolLM2-1.7B	SmolLM2-1.7B-Instruct	1.7B
Qwen/Qwen1.5-1.8B	Qwen1.5-1.8B-Chat	1.8B
google/gemma-2b	gemma-2b-it	2.0B
google/gemma-2-2b	gemma-2-2b-it	2.0B
meta-llama/Llama-3.2-3B	Llama-3.2-3B-Instruct	3.0B
Qwen/Qwen1.5-4B	Qwen1.5-4B-Chat	4.0B
Qwen/Qwen2-7B	Qwen2-7B-Instruct	7.0B
mistralai/Mistral-7B-v0.1	Mistral-7B-Instruct-v0.1	7.0B
tiiuae/falcon-7b	falcon-7b-instruct	7.0B

Table 54: Causal language models for structured perturbation analysis (Experiment 2, 16 models) and extended canary validation (Experiment 4, 15 models). Experiment 4 excludes SmolLM2-1.7B due to SST-2 evaluation constraints. Both experiments applied Gaussian noise, quantization, and LoRA perturbations.

Model	Family	Params	Exp 4
HuggingFaceTB/SmolLM-135M	SmolLM	0.14B	✓
HuggingFaceTB/SmolLM2-135M	SmolLM2	0.14B	✓
HuggingFaceTB/SmolLM-360M	SmolLM	0.36B	✓
HuggingFaceTB/SmolLM2-360M	SmolLM2	0.36B	✓
Qwen/Qwen2-0.5B	Qwen2	0.5B	✓
meta-llama/Llama-3.2-1B	Llama	1.0B	✓
TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T	TinyLlama	1.1B	✓
Qwen/Qwen2-1.5B	Qwen2	1.5B	✓
stabilityai/stablelm-2-1_6b	StableLM	1.6B	✓
HuggingFaceTB/SmolLM-1.7B	SmolLM	1.7B	✓
HuggingFaceTB/SmolLM2-1.7B	SmolLM2	1.7B	-
google/gemma-2b	Gemma	2.0B	✓
google/gemma-2-2b	Gemma-2	2.0B	✓
meta-llama/Llama-3.2-3B	Llama	3.0B	✓
Qwen/Qwen2-7B	Qwen2	7.0B	✓
mistralai/Mistral-7B-v0.1	Mistral	7.0B	✓

Table 55: Sentence embedding models for canary validation (Experiment 3, 26 models).

Model	Family
sentence-transformers/paraphrase-MiniLM-L3-v2	MiniLM
sentence-transformers/all-MiniLM-L6-v2	MiniLM
sentence-transformers/paraphrase-MiniLM-L6-v2	MiniLM
sentence-transformers/all-MiniLM-L12-v2	MiniLM
sentence-transformers/multi-qa-MiniLM-L6-cos-v1	MiniLM
sentence-transformers/all-mpnet-base-v2	MPNet
sentence-transformers/paraphrase-mpnet-base-v2	MPNet
sentence-transformers/multi-qa-mpnet-base-cos-v1	MPNet
sentence-transformers/distilbert-base-nli-mean-tokens	DistilBERT
sentence-transformers/all-distilroberta-v1	DistilRoBERTa
sentence-transformers/paraphrase-distilroberta-base-v1	DistilRoBERTa
sentence-transformers/bert-base-nli-mean-tokens	BERT
sentence-transformers/stsb-roberta-base	RoBERTa
sentence-transformers/nli-roberta-base-v2	RoBERTa
sentence-transformers/paraphrase-albert-small-v2	ALBERT
thenlper/gte-small	GTE
thenlper/gte-base	GTE
thenlper/gte-large	GTE
intfloat/e5-small-v2	E5
intfloat/e5-base-v2	E5
intfloat/e5-large-v2	E5
BAAI/bge-small-en-v1.5	BGE
BAAI/bge-base-en-v1.5	BGE
BAAI/bge-large-en-v1.5	BGE
princeton-nlp/sup-simcse-bert-base-uncased	SimCSE
princeton-nlp/unsup-simcse-bert-base-uncased	SimCSE

11 Transfer Learning and the Limits of Unsupervised Stability

This appendix contains all the methods and analyzes related to the transfer learning experiments. First we evaluate the role of geometric stability in predicting few-shot transfer. Next, we evaluate its role in predicting cross-domain transfer.

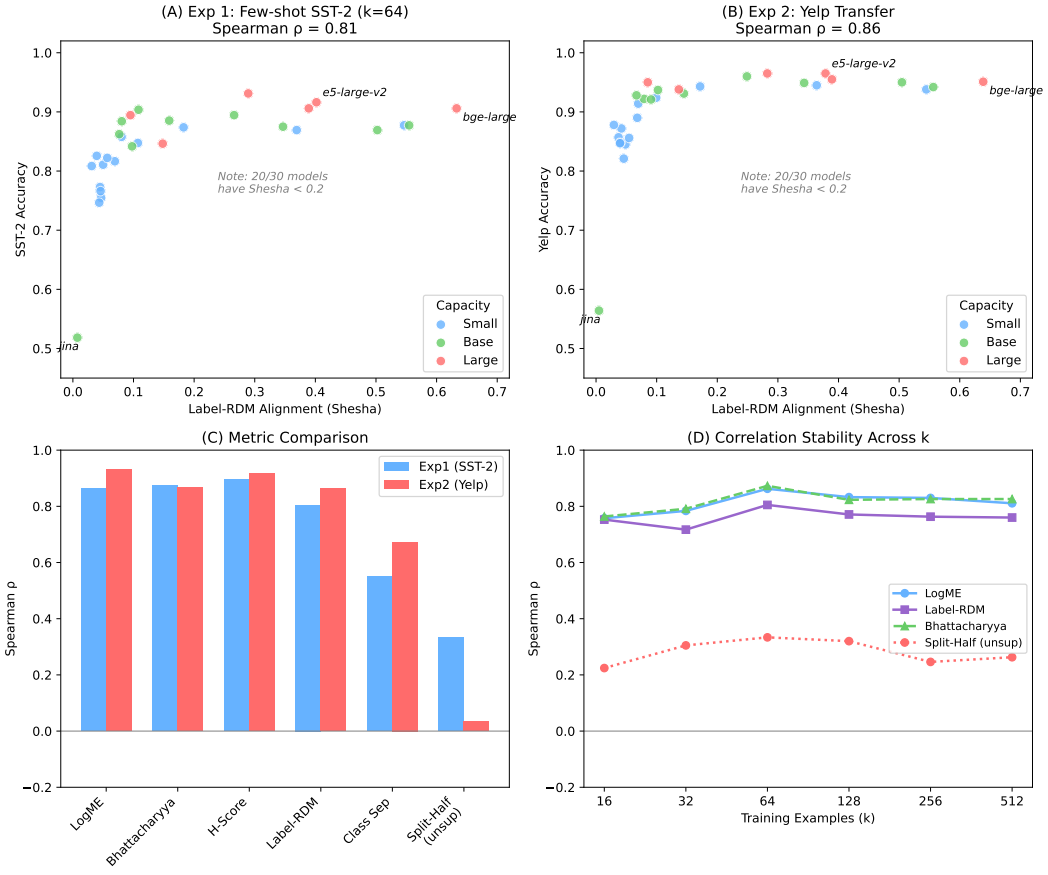

Figure 30: Transfer experiments summary. (A) Experiment 1 results for Label-RDM Alignment. (B) Experiment 2 results for Label-RDM Alignment. (C) Metric comparison across experiments. (D) Correlation of stability in Experiment 1 across levels of k.

11.1 Experimental Design

Models. We examined 30 sentence embedding models, representing the Sentence-Transformers ecosystem. The models fall into three capacity tiers, as follows

- **Small** (14 models): MiniLM variants, E5-small, GTE-small, BGE-small, paraphrase models
- **Base** (10 models): MPNet, E5-base, GTE-base, BGE-base, Sentence-T5-base, GIST, Jina-v2
- **Large** (6 models): E5-large, GTE-large, BGE-large, Sentence-T5-large, GTR-T5-large

Datasets.

- **Source domain:** IMDB movie reviews (Maas et al., 2011) (2,000 samples for metric computation)
- **Experiment 1 target:** SST-2 sentiment (Socher et al., 2013) classification (few-shot)
- **Experiment 2 target:** Yelp review polarity (Zhang et al., 2015) (full transfer)

Evaluation protocol. The method for testing each model was done as follows:

1. Compute normalized embeddings for the source domain (IMDB samples)
2. Calculate all transferability metrics on source embeddings with labels
3. Fine-tune linear probes on target domain with hyperparameter search
4. Report test accuracy as transfer performance measure

The linear probes that were used included the following:

- logistic regression ($C \in \{0.1, 1, 10\}$)
- ridge classifier ($\alpha \in \{1, 10\}$)
- LDA
- nearest centroid

The best probe was selected based on validation set accuracy.

Sample sizes.

- Experiment 1: $k_{\text{total}} \in \{16, 32, 64, 128, 256, 512\}$ training examples (balanced across classes)
- Experiment 2: 2,400 training / 600 validation / 1,000 test (fixed split)
- Total: 180 observations (Exp1) + 30 observations (Exp2)

11.2 Metric Definitions

Label-RDM Alignment (label-informed). The Spearman correlation between embedding dissimilarity matrix and label dissimilarity matrix was calculated as follows:

$$\text{Shesha}_{\text{RSA}} = \rho_s\big(\text{vec}(\mathbf{D}_{\text{embed}}), \text{vec}(\mathbf{D}_{\text{label}})\big)$$

where $\mathbf{D}_{\text{embed}}$ uses cosine distance and $\mathbf{D}_{\text{label}}$ uses Hamming distance (0 for same class, 1 for different). Computed with 50 bootstrap iterations at 50% subsampling.

Class Separation (label-informed). The ratio of between-class to within-class distances was calculated as follows:

$$\text{Shesha}_{\text{sep}} = \frac{\bar{d}_{\text{between}}}{\bar{d}_{\text{within}}}$$

Shesha Feature-Split (unsupervised). RDM correlation between random dimension partitions as follows:

$$\text{Shesha}_{\text{split}} = \rho_s\big(\text{RDM}(\mathbf{X}_{:,\mathcal{D}_1}), \text{RDM}(\mathbf{X}_{:,\mathcal{D}_2})\big)$$

where $\mathcal{D}_1, \mathcal{D}_2$ are random non-overlapping dimension subsets. This measures intrinsic geometric consistency without label information.

LDA Subspace (label-informed). Cosine similarity between LDA discriminant directions computed on the full data versus bootstrap samples.

Baseline metrics. We measure geometric stability against well-established transferability estimators:

- LogME (You et al., 2021, 2022)
- Bhattacharyya distance (Pándy et al., 2022)
- H-Score (Bao et al., 2019)
- margin score
- centroid softmax
- NCE (Tran et al., 2019)

11.3 Results

Experiment 1: Few-shot SST-2. Table 56 shows correlations between transferability metrics and few-shot accuracy across sample sizes. Key findings:

- **H-Score is consistently strongest**: $\rho = 0.79$-0.89 across k values, achieving $\rho = 0.89$ at $k = 64$.
- **Label-RDM Alignment is competitive**: $\rho = 0.72$-0.81, comparable to LogME and Bhattacharyya.
- **Unsupervised stability is weak**: Split-Half achieves only $\rho = 0.22$-0.33 (not significant at any k), indicating intrinsic geometric consistency does not predict few-shot transfer.
- **LDA Subspace is negative**: Consistent $\rho \approx -0.47$ to -0.58 ($p < 0.01$), suggesting discriminant direction stability is inversely related to transfer, possibly because stable LDA directions indicate overfitting to source domain structure.

Table 56: Experiment 1: Transferability metric correlations with few-shot SST-2 accuracy ($n = 30$ models, Spearman ρ).

Metric	$\boldsymbol{k = 16}$	$\boldsymbol{k = 32}$	$\boldsymbol{k = 64}$	$\boldsymbol{k = 128}$	$\boldsymbol{k = 256}$	$\boldsymbol{k = 512}$
Shesha variants						
Label-RDM Align.	0.75^{***}	0.72^{***}	0.81^{***}	0.77^{***}	0.76^{***}	0.76^{***}
Class Separation	0.45^{*}	0.60^{***}	0.55^{**}	0.63^{***}	0.60^{***}	0.58^{***}
Split-Half (unsup.)	0.22	0.31	0.33	0.32	0.25	0.26
LDA Subspace	-0.47^{**}	-0.54^{**}	-0.53^{**}	-0.48^{**}	-0.54^{**}	-0.58^{***}
Baseline metrics						
LogME	0.76^{***}	0.78^{***}	0.86^{***}	0.83^{***}	0.83^{***}	0.81^{***}
Bhattacharyya	0.76^{***}	0.79^{***}	$\mathbf{0.87}^{***}$	0.82^{***}	0.83^{***}	0.83^{***}
H-Score	0.79^{***}	$\mathbf{0.85}^{***}$	$\mathbf{0.89}^{***}$	$\mathbf{0.86}^{***}$	$\mathbf{0.88}^{***}$	$\mathbf{0.88}^{***}$
Margin Score	0.75^{***}	0.76^{***}	0.83^{***}	0.80^{***}	0.78^{***}	0.77^{***}
Centroid Softmax	0.71^{***}	0.67^{***}	0.77^{***}	0.72^{***}	0.69^{***}	0.68^{***}
NCE	0.55^{**}	0.41^{*}	0.58^{***}	0.46^{*}	0.43^{*}	0.43^{*}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Experiment 2: Cross-domain transfer (IMDB → Yelp). Table 57 shows results for full cross-domain transfer. The pattern strengthens:

- **LogME dominates**: $\rho = 0.93$ ($p < 0.001$), the strongest single predictor.
- **Label-informed metrics are strong**: Label-RDM Alignment ($\rho = 0.86$), H-Score ($\rho = 0.92$), and Bhattacharyya ($\rho = 0.87$) are all highly significant.
- **Unsupervised stability collapses**: Split-Half drops to $\rho = 0.03$ ($p = 0.86$), confirming that intrinsic geometric consistency does *not* predict cross-domain transfer.
- **Higher baseline correlations**: Most label-informed metrics achieve $\rho > 0.85$, likely due to the reduced noise associated with more training data.

Table 57: Experiment 2: Cross-domain transfer correlations (IMDB → Yelp, $n = 30$ models).

Metric	Spearman ρ	Type
Shesha variants		
Label-RDM Alignment	0.86^{***}	Label-informed
Class Separation	0.67^{***}	Label-informed
Split-Half (dims)	0.03	Unsupervised
LDA Subspace	-0.35	Label-informed
Baseline metrics		
LogME	$\mathbf{0.93}^{***}$	Label-informed
Bhattacharyya	0.87^{***}	Label-informed
H-Score	0.92^{***}	Label-informed
Margin Score	0.86^{***}	Label-informed
Centroid Softmax	0.74^{***}	Label-informed
NCE	0.49^{**}	Label-informed

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

11.4 Model Performance Analysis

Table 58 shows individual model results sorted by Label-RDM Alignment. Key patterns observed:

- **The BGE family achieves the highest alignment**: BGE-large (0.633), BGE-base (0.555), and BGE-small (0.546) are consistently top-ranked.
- **The alignment-accuracy relationship is monotonic but imperfect**: E5-large-v2 achieves the highest Exp1 accuracy (0.930) despite having moderate alignment (0.402).
- **Jina-v2 is an outlier**: It has Near-zero alignment (0.007) with near-chance accuracy (0.509-0.564), which is consistent with known issues in the embedding space.
- **The performance gap** between the Top-5 and Bottom-5 models by alignment shows accuracy differences of 14.4% (Exp1) and 13.3% (Exp2).

Table 58: Model performance sorted by Label-RDM Alignment.

Model	Capacity	Shesha	Exp1 Acc	Exp2 Acc
bge-large-en-v1.5	Large	0.633	0.908	0.951
bge-base-en-v1.5	Base	0.555	0.877	0.942
bge-small-en-v1.5	Small	0.546	0.874	0.938
GIST-small-Embedding-v0	Base	0.502	0.866	0.950
e5-large-v2	Large	0.402	0.930	0.965
gte-large	Large	0.389	0.914	0.955
gte-small	Small	0.369	0.826	0.945
gte-base	Base	0.347	0.888	0.949
sentence-t5-large	Large	0.289	0.936	0.965
e5-base-v2	Base	0.266	0.906	0.960
e5-small-v2	Small	0.183	0.889	0.943
sentence-t5-base	Base	0.159	0.896	0.931
e5-large	Large	0.148	0.893	0.938
paraphrase-mpnet-base-v2	Base	0.108	0.904	0.937
e5-small	Small	0.107	0.866	0.924
e5-base	Base	0.098	0.866	0.921
gtr-t5-large	Large	0.095	0.892	0.950
all-mpnet-base-v2	Base	0.081	0.883	0.922
all-distilroberta-v1	Small	0.080	0.854	0.914
gtr-t5-base	Base	0.077	0.865	0.928
paraphrase-MiniLM-L12-v2	Small	0.069	0.846	0.890
paraphrase-albert-small-v2	Small	0.057	0.818	0.856
paraphrase-MiniLM-L6-v2	Small	0.050	0.827	0.845
paraphrase-MiniLM-L3-v2	Small	0.047	0.778	0.821
multi-qa-MiniLM-L6-cos-v1	Small	0.046	0.705	0.847
all-MiniLM-L6-v2	Small	0.045	0.808	0.872
msmarco-MiniLM-L6-cos-v5	Small	0.043	0.771	0.848
all-MiniLM-L12-v2	Small	0.039	0.828	0.857
MiniLM-L6-H384-uncased	Small	0.031	0.806	0.878
jina-embeddings-v2-base-en	Base	0.007	0.509	0.564

11.5 The Stability-Alignment Dissociation

A noticeable trend emerges when comparing these results to the steering experiments (Section 3.1):

Table 59: Unsupervised versus label-informed stability across tasks.

Setting	Unsupervised	Label-Informed	Gap
Steering (Synthetic)	$\rho = 0.77^{***}$	$\rho = 0.89^{***}$	0.12
Steering (SST-2)	$\rho \approx 0.10$	$\rho = 0.96^{***}$	0.86
Steering (MNLI)	$\rho \approx 0.10$	$\rho = 0.96^{***}$	0.86
Transfer (SST-2, k=64)	$\rho = 0.33$	$\rho = 0.89^{***}$	0.56
Transfer (Yelp)	$\rho = 0.03$	$\rho = 0.93^{***}$	0.90

Interpretation. In controlled synthetic settings where the data manifold perfectly aligns with task structure, unsupervised stability is a powerful predictor of performance ($\rho = 0.77$). In real-world semantic tasks (both steering and transfer), unsupervised stability fails ($\rho < 0.10$). The supervised label-aligned metrics remain highly predictive ($\rho > 0.85$). This confirms that **intrinsic manifold rigidity cannot account for semantic generalization; stability must be measured with respect to task-relevant structure**.

11.6 Limitations

Shesha versus state-of-the-art. While Label-RDM Alignment achieves strong correlations ($\rho \approx 0.81$-0.86), it is inferior to purpose-built transfer estimation methods like LogME ($\rho \approx 0.86$-0.93) or H-Score ($\rho \approx 0.89$-0.92). We do not intend for Shesha to replace these metrics. Instead, we suggest a geometric diagnostic that offers complementary insights, particularly the stability-alignment dissociation as highlighted above.

Domain specificity. Results are limited to sentiment analysis benchmarks. We believe that evaluation on more diverse transfer scenarios (e.g., cross-lingual, cross-modal) remains important for future work.

Single seed. Due to computational constraints, we report results for a single random seed (320). The results from the preliminary experiments with additional seeds showed consistent metric rankings.

11.7 Conclusion

These transfer learning experiments provide convergent evidence for the stability-alignment hypothesis:

1. **Unsupervised geometric stability does not predict transfer**: Split-Half achieves $\rho = 0.33$ (few-shot) and $\rho = 0.03$ (cross-domain), both of which are non-significant.
2. **Label-informed metrics succeed**: H-Score ($\rho = 0.89$-0.92), LogME ($\rho = 0.86$-0.93), Label-RDM Alignment ($\rho = 0.81$-0.86), and related metrics achieve strong, significant correlations.
3. **Task alignment is required**: The 0.56-0.90 gap between unsupervised and label-informed metrics demonstrates that for semantic transfer, stability must be measured relative to the downstream task structure.

This null result for unsupervised stability provides valuable insights. defines the limits within which geometric consistency can predict performance for certain internal tasks, such as drift detection or CRISPR perturbations. By contrast, it indicates situations where alignment with the task's semantics is necessary, particularly in areas like transfer and real-world steering scenarios.

12 CRISPR Perturbation Magnitudes: Extended Methods and Results

This appendix contains all the methods and additional analyses related to the CRISPR perturbation experiments outlined in Section 3.4. The analysis of geometric stability measures if there is coherence of perturbation across four publicly available single cell transcriptomic datasets (422 perturbations, 212,865 cells). First we establish how the magnitude of a perturbation relates to the stability and how stable CRISPRa perturbations are in comparison to CRISPRi and Pooled Screen perturbations. Second, we test the robustness of our methods by looking at three sets of metrics: distance metrics, PCA dimensionality, and random seed numbers. Finally, we identify and characterize discordant instances in which stability and magnitude do not correlate with each other, thereby uncovering biologically relevant differences between pleiotropic and lineage-specific regulators.

12.1 Datasets

We analyzed four publicly available single-cell transcriptomic CRISPR datasets spanning different perturbation modalities, comprising 422 perturbations and 212,865 cells in total:

- **Norman et al. (2019):** CRISPRa (activation) screening in K562 cells targeting transcription factors, including single-gene and combinatorial perturbations (Norman et al., 2019).
- **Adamson et al. (2016):** CRISPRi (interference) pilot screen targeting stress response regulators (Adamson et al., 2016).
- **Dixit et al. (2016):** CRISPRi screen with single and combinatorial guide delivery targeting transcription factors (Dixit et al., 2016).
- **Papalexi et al. (2021):** Pooled CRISPR screen with readout of the CITE-sequence; we analyze the RNA modality (Papalexi et al., 2021).

All datasets were accessed via the `pertpy` package (Heumos et al., 2025). Table 60 summarizes the characteristics of the datasets.

Table 60: Dataset overview. Cells analyzed refers to perturbed cells meeting quality thresholds (excluding controls).

Dataset	Modality	Perturbations	Cells Analyzed	Median Cells/Pert.	Range
Norman 2019	CRISPRa	236	99,420	352	54-1,954
Adamson 2016	CRISPRi	8	5,752	559	477-1,769
Dixit 2016	CRISPRi	153	89,350	75	10-11,676
Papalexi 2021	CRISPR	25	18,343	662	48-1,341

12.2 Preprocessing

Each dataset was preprocessed independently using standard single-cell RNA-seq workflows implemented in `scanpy` (Wolf et al., 2018):

1. **Quality filtering:** Cells with fewer than 100 detected genes were removed.
2. **Normalization:** Normalization of the library size using normalization `_total()`.
3. **Log transformation:** $\log(x + 1)$ transform via `log1p()`.
4. **Feature selection:** Top 2,000 highly variable genes selected via `highly_variable_genes()`.
5. **Dimensionality reduction:** PCA with 50 components computed per dataset.

PCA embeddings for each dataset were computed *separately* because of the potential for batch effects if they were computed using a common shared space. However, each PCA matrix maintained a consistent dimensionality that may be compared in a cross-study format. All analyses used random seed 320 for reproducibility .

12.3 Control Group Identification

The control cells were identified using a robust multi-stage matching procedure. For each dataset, we first attempted to exact case-insensitive matches against known control labels (e.g., "control," "ctrl," "non-targeting"). For short tokens (e.g., "nt," "neg"), we used delimiter-aware regex matching to avoid false positives on gene names (e.g., matching "NT" but not "NEGR1"). Finally, substring matching was only applied for longer keywords (≥ 4 characters). Table 61 summarizes the control labels that were used.

Note: The Papalexi 2021 dataset required special handling because perturbation metadata is stored at the MuData level rather than within the RNA modality. We explicitly copied the `gene_target` column from the global metadata to allow us to properly identify the control. This groups all non-targeting guides (NTg1-NTg7) into a single "NT" control population of 2,386 cells.

Table 61: Control group identification by dataset.

Dataset	Control Label	Control Cells	Notes
Norman 2019	`control`	11,835	Exact match
Adamson 2016	`NaN_Control`	10	NaN values converted
Dixit 2016	`control`	8,878	Exact match
Papalexi 2021	`NT`	2,386	Gene-target level; 6 NT guides pooled

12.4 Stability Computation

For each perturbation p with n_p cells, we measured the Shesha stability score as the directional coherence of perturbation effects:

1. Let $\mathbf{c} = \frac{1}{n_{\text{ctrl}}} \sum_i \mathbf{x}_i^{\text{ctrl}}$ be the control centroid in PCA space.
2. For each perturbed cell j, compute the shift vector $\mathbf{v}_j = \mathbf{x}_j^p - \mathbf{c}$.
3. Compute the mean shift direction $\bar{\mathbf{v}} = \frac{1}{n_p} \sum_j \mathbf{v}_j$ and its magnitude $\|\bar{\mathbf{v}}\|$.
4. For cells with $\|\mathbf{v}_j\| > 10^{-6}$, compute cosine similarity to the mean direction:

$$S_p = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \frac{\mathbf{v}_j \cdot \bar{\mathbf{v}}}{\|\mathbf{v}_j\| \|\bar{\mathbf{v}}\|}$$

   where $\mathcal{V} = \{j : \|\mathbf{v}_j\| > 10^{-6}\}$.

This formula measures how self-consistency of a geometric perturbation is determined by the degree to which the perturbed cells move coherently together (in the same direction) relative to their controls. Perturbations with fewer than 10 cells or mean magnitude $< 10^{-6}$ were excluded. The magnitude of the effect was defined as $\|\bar{\mathbf{v}}\|$, and spread (intrinsic variance) as the mean distance between the perturbed cells and their own centroid.

12.5 Distance Metric Variants

To assess robustness to methodological choices, we implemented three distance computation approaches:

- **Euclidean:** Standard L_2 distance in a 50-dimensional PCA space (primary analysis).
- **Mahalanobis-whitened:** Distances computed after applying a $\Sigma^{-1/2}$ transformation to the PCA space, removing correlations between dimensions and equalizing variance along each axis.
- **k-NN local centroids:** For each perturbed cell, the control centroid is computed from its $k = 50$ nearest control neighbors rather than the global control population, accounting for the local structure in the control distribution.

12.6 Bootstrap Confidence Interval Methodology

The confidence intervals were computed using the bootstrap resampling technique with 10,000 iterations of each bootstrapped sample as follows:

1. Resample perturbations values with replacement for each dataset,

2. Compute the statistical result of interest (correlation, partial correlation, etc.)

3. Log select samples/estimates into collection of bootstrapped estimates

The 95% confidence interval was obtained by using the percentile method (2.5% and 97.5% percentiles of the bootstrapped distribution). Bootstrapped samples that produced NaN values (due to all resamples being constant) were excluded from the calculation of percentiles. Analyses that dropped more than 5 percent of samples produced warning messages.

To ensure reproducibility and to eliminate artificially smooth results across all analyses, unique seeds were assigned for each (dataset, ablation type, parameter) combination by hashing the combination identifier with MD5.

Because of the limited sample size of Adamson 2016 ($N = 8$), the bootstrap method gives very wide confidence intervals which accurately reflect real uncertainty rather than the appearance of too much precision.

12.7 Summary Statistics

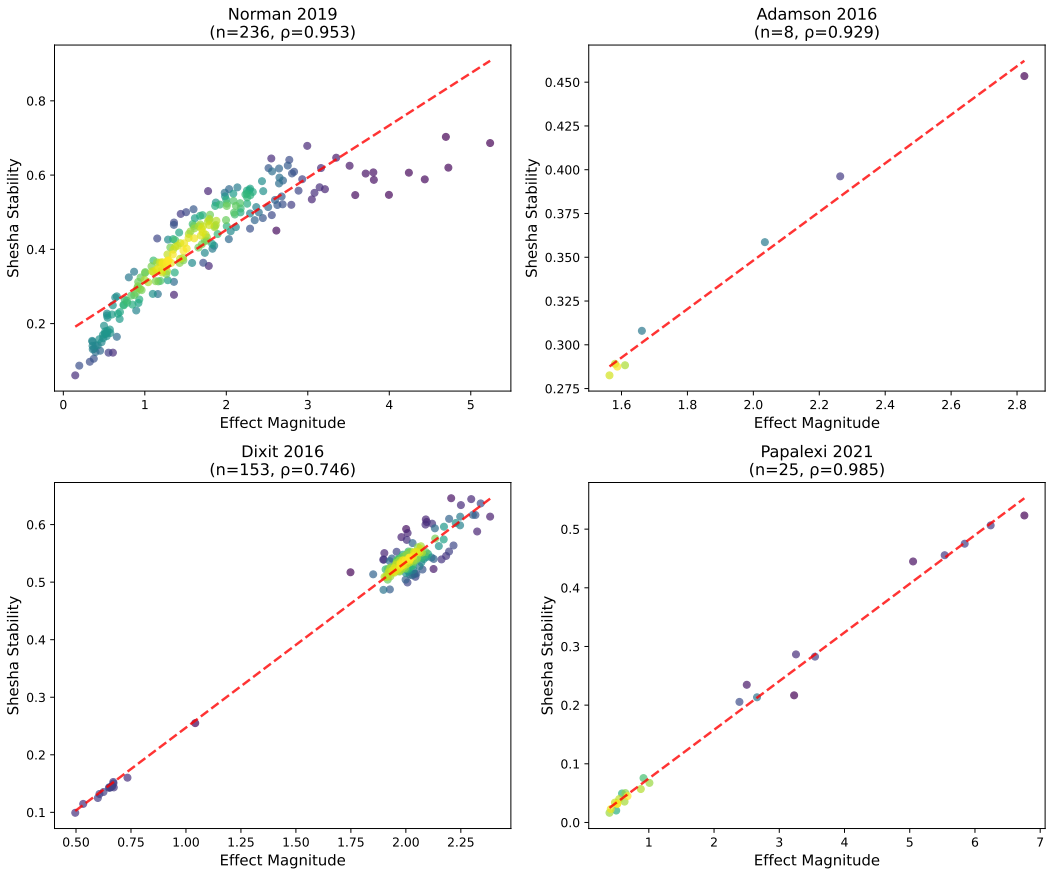

Figure 31: **Geometric stability tracks perturbation magnitude across CRISPR modalities.** Stability vs. effect magnitude for four independent datasets spanning CRISPRa, CRISPRi, and pooled screens. Each point represents one perturbation; color indicates local density. Spearman correlations range from $\rho = 0.746$ [0.641, 0.827] (Dixit) to $\rho = 0.985$ [0.939, 0.997] (Papalexi). Dashed lines show linear fits. The consistency across datasets, modalities, and cell types suggests a universal geometric relationship between effect size and directional coherence.

Table 62: Summary statistics for Shesha stability and effect magnitude across CRISPR datasets.

Dataset	N	Stability		Magnitude	
		Mean (SD)	Range	Mean (SD)	Range
Norman 2019	236	0.40 (0.14)	[0.06, 0.70]	1.65 (0.92)	[0.15, 5.24]
Adamson 2016	8	0.33 (0.06)	[0.28, 0.45]	1.89 (0.46)	[1.56, 2.82]
Dixit 2016	153	0.51 (0.12)	[0.10, 0.65]	1.90 (0.42)	[0.50, 2.38]
Papalexi 2021	25	0.18 (0.18)	[0.02, 0.52]	2.23 (2.13)	[0.40, 6.76]

12.8 Correlation Structure with Bootstrap Confidence Intervals

Table 63 reports Spearman correlations between stability and magnitude with 95% bootstrap confidence intervals (10,000 iterations).

Table 63: Spearman correlations between stability and potential confounds across datasets. Bootstrap 95% CIs (10,000 iterations) reported for magnitude-stability correlation. Mag = effect magnitude; Var = intrinsic spread; N = cells per perturbation.

Dataset	n	ρ_{Mag}	95% CI	p	ρ_{Var}	p	ρ_N	p
Norman 2019	236	0.953	[0.934, 0.965]	$< 10^{-100}$	0.67	$< 10^{-31}$	-0.32	$< 10^{-6}$
Adamson 2016	8	0.929	[0.407, 1.000]	$< 10^{-3}$	0.07	0.87	0.19	0.65
Dixit 2016	153	0.746	[0.641, 0.827]	$< 10^{-100}$	-0.84	$< 10^{-40}$	-0.63	$< 10^{-17}$
Papalexi 2021	25	0.985	[0.939, 0.997]	$< 10^{-18}$	-0.47	0.02	-0.17	0.43
Pooled	422	0.833	—	$< 10^{-100}$	0.13	0.01	-0.65	$< 10^{-50}$

Key observations. The relationship between magnitude and stability, as measured by a pooling coefficient of correlation, is consistently strong and positive ($\rho = 0.740 - .985$), with tight confidence intervals for larger sample sizes, and wide confidence intervals for smaller sample sizes (Adamson, $n = 8$). The pooled correlation ($\rho = 0.833$) is consistent across all of the datasets. In the pooled data, stability showed a negative relationship with the sample size ($\rho = -0.65$), reflecting the cross-dataset differences in the cell counts rather than a within-dataset confound.

Honest uncertainty quantification. The Adamson 2016 confidence interval [0.407, 1.000] spans a wide range, reflecting the inherent uncertainty of estimating correlations from only 8 observations. This is a feature, not a bug, of bootstrap inference, which prevents overconfident claims from underpowered analyzes.

12.9 Robustness to Distance Metric

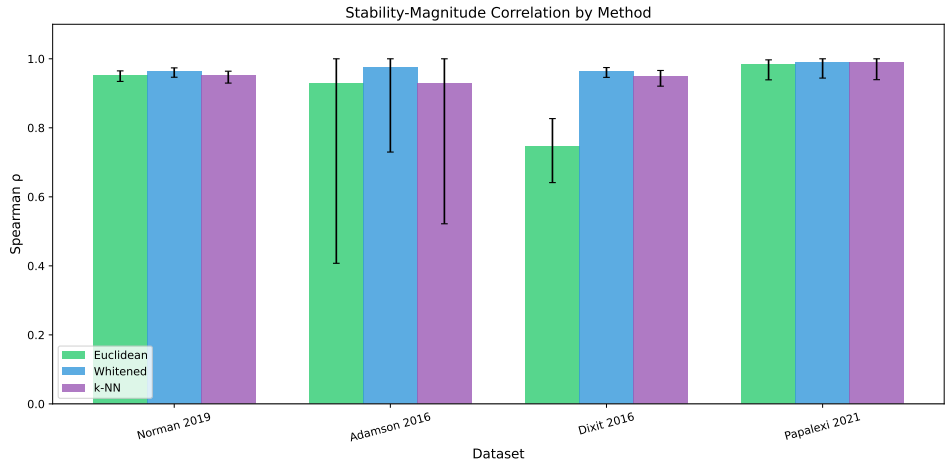

Figure 32: **Stability-magnitude correlation is robust across distance metrics.** Bar chart showing Spearman correlations with 95% bootstrap CIs (error bars) for three distance computation methods: Euclidean (standard L_2 in PCA space), Whitened (Mahalanobis-scaled coordinates), and k-NN (local control centroids). All methods achieve strong correlations ($\rho > 0.74$) across all datasets. Notably, whitening substantially improves the Dixit correlation from $\rho = 0.75$ to $\rho = 0.97$, suggesting residual covariance structure in PCA space attenuates the relationship in that dataset.

Table 64: Stability-magnitude correlation across methods and datasets with 95% bootstrap CIs (10,000 iterations). All methods achieve strong correlations ($\rho > 0.74$) across datasets.

Dataset	Method	n	ρ	95% CI
Norman 2019	Euclidean	236	0.953	[0.934, 0.965]
Norman 2019	Whitened	236	0.963	[0.947, 0.974]
Norman 2019	k-NN	236	0.951	[0.930, 0.964]
Adamson 2016	Euclidean	8	0.929	[0.407, 1.000]
Adamson 2016	Whitened	8	0.976	[0.730, 1.000]
Adamson 2016	k-NN	8	0.929	[0.522, 1.000]
Dixit 2016	Euclidean	153	0.746	[0.641, 0.827]
Dixit 2016	Whitened	153	0.965	[0.946, 0.975]
Dixit 2016	k-NN	153	0.949	[0.921, 0.966]
Papalexi 2021	Euclidean	25	0.985	[0.939, 0.997]
Papalexi 2021	Whitened	25	0.991	[0.944, 1.000]
Papalexi 2021	k-NN	25	0.988	[0.939, 1.000]

Key findings. All three methods yielded consistent results across all of the datasets. In particular, whitening substantially improved the Dixit correlation from $\rho = 0.746$ [0.641, 0.827] to $\rho = 0.965$ [0.946, 0.975], suggesting that the residual covariance structure in the PCA space was attenuating the relationship in the primary analysis. Papalexi 2021 showed the strongest correlations in all of the methods ($\rho > 0.98$), reflecting the cleaner gene-level aggregation enabled by proper control identification. The consistency across methods confirms that the magnitude-stability relationship is robust to methodological choices.

12.10 Partial Correlation Analysis: Controlling for SNR

To test whether stability captures information beyond signal-to-noise ratio (SNR), we computed partial correlations between magnitude and stability controlling for SNR, where SNR = magnitude / spread. All confidence intervals were computed through 10,000 bootstrap replicas.

Table 65: Partial correlations (magnitude-stability | SNR) with 95% bootstrap CIs (10,000 iterations). Dataset-specific heterogeneity is evident.

Dataset	ρ_{partial}	95% CI	p
Norman 2019	-0.859	$[-0.905, -0.781]$	6.69×10^{-70}
Dixit 2016	0.627	[0.482, 0.728]	4.04×10^{-18}
Papalexi 2021	0.176	$[-0.405, 0.547]$	0.400
Pooled	$\mathbf{-0.258}$	$\mathbf{[-0.423, -0.111]}$	7.98×10^{-8}

Dataset-specific heterogeneity. The partial correlations show reveal notable heterogeneity across the datasets:

- **Norman 2019:** A strong *negative* partial correlation ($\rho = -0.859$ $[-0.905, -0.781]$), indicates that, when controlling for SNR, larger perturbations are associated with *lower* stability.
- **Dixit 2016:** A strong *positive* partial correlation ($\rho = 0.627$ [0.482, 0.728]), consistent with the simple correlation findings.
- **Papalexi 2021:** A weak positive partial correlation ($\rho = 0.176$ $[-0.405, 0.547]$), which is not statistically different from zero ($p = 0.40$).

Heterogeneity of the data sets might be caused by biological variability among data sets (cell type, CRISPR technology) and variation due to their specific methodologies (such as library preparation or sequencing depth).

Pooled estimate. The pooled estimate for partial correlation ($\rho = -0.258$ [-0.423, -0.111]) is negative and significant ($p < 10^{-7}$), primarily driven by the large size of the Norman 2019 data set relative to the other data sets included in this analysis. This suggests that after controlling for SNR, the residual magnitude-stability relationship is dependent on the dataset used for each measurement.

Adamson excluded. Adamson 2016 was excluded from the partial correlation analysis due to insufficient sample size ($n = 8$) to estimate three-variable partial correlation reliably.

12.11 Mixed-Effects Modeling

To quantify the relative contributions of magnitude, sample size, and spread to stability while accounting for dataset-level variation, we fit a linear mixed-effects model:

$$S_{ij} = \beta_0 + \beta_1 \cdot \text{magnitude}_{ij} + \beta_2 \cdot \log(\text{n\_cells})_{ij} + \beta_3 \cdot \text{spread}_{ij} + u_j + \epsilon_{ij}$$

where i indexes perturbations, j indexes datasets, and $u_j \sim N(0, \sigma_u^2)$ captures random intercepts at the dataset level. All predictors were z-scored within datasets before pooling. When restricted maximum likelihood (REML) estimation failed to converge, we used maximum likelihood (ML) estimation with multiple optimizer attempts (L-BFGS, Powell, CG, BFGS), with fallback to partial correlation with bootstrap CI.

Table 66: Mixed-effects model results with 95% confidence intervals from ML estimation. Magnitude is the dominant predictor of stability, with an effect approximately 4× larger than sample size.

Predictor	β	95% CI	z	p
Intercept	0.000	-	0.00	1.000
Magnitude (z-scored)	**0.123**	**[0.116, 0.131]**	32.10	$< 10^{-200}$
Spread (z-scored)	-0.122	[-0.158, -0.086]	-6.64	3.1×10^{-11}
Sample size (z-scored)	-0.031	[-0.039, -0.024]	-8.14	4.1×10^{-16}
Random effects				
σ_u^2 (dataset variance)	0.000			
σ_ϵ^2 (residual)	0.0036			

Key findings.

- **Magnitude is the dominant predictor:** $\beta = 0.123$ [0.116, 0.131], with an effect size approximately 4× larger than the sample size ($\beta = -0.031$). All effects are highly significant ($p < 10^{-10}$).
- **Spread reduces stability:** A higher spread within the perturbation is associated with a lower stability ($\beta = -0.122$ [-0.158, -0.086]), consistent with the interpretation that heterogeneous perturbation effects reduce geometric coherence.
- **The sample size has a modest negative effect:** Larger sample sizes are associated with slightly lower estimated stability ($\beta = -0.031$ [-0.039, -0.024]), which could reflect regression to the mean or reflect an increased power to capture minute heterogeneity.
- **Dataset random effect is negligible:** The random effect variance ($\sigma_u^2 \approx 0$) indicates that after controlling for fixed effects, there is little residual between-dataset heterogeneity.

12.12 Calibrated Cross-Dataset Analysis

The raw pooled magnitude-stability correlation ($\rho = 0.833$) is strong, and using z-scores within the datasets further improves it. In order to ensure that data can be reliably compared from one dataset to another, we first calculated z-scores for both magnitude and stability in each of the datasets before pooling them.

Table 67: Calibrated cross-dataset correlations with 95% bootstrap CIs (10,000 iterations).

Analysis	ρ	95% CI
Raw pooled ρ(magnitude, stability)	0.833	-
Pooled ρ(magnitude$_z$, stability$_z$)	**0.913**	**[0.884, 0.936]**
Pooled ρ(calibrated magnitude, stability)	0.776	[0.716, 0.828]

Once the pooled z-score calibration was completed, the correlation captured was $\rho = 0.913$ ([0.884, 0.936]) indicating the relationship remains constant between datasets after removing scale differences through calibration. Therefore, while the absolute values of stability differ between datasets, the relative ranking of perturbations by stability across different experimental settings is meaningful.

12.13 Theoretical Null Model

To establish a theoretical baseline, we simulated perturbations under a simple Gaussian shift model with isotropic noise.

12.13.1 Model Specification

- **Control distribution:** $\mathbf{x}_{\text{control}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ with $d = 50$ dimensions
- **Perturbation effect:** Shift by $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma_\delta^2 \mathbf{I}_d)$
- **Noise:** Isotropic Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_d)$
- **Noise levels:** $\sigma \in \{0.5, 1.0, 2.0, 3.0\}$ with 500 simulations per condition (2,000 total)

Under this model, stability is analytically determined by SNR: larger shifts produce larger magnitudes but also move the population away from the control region, reducing stability.

12.13.2 Results

Table 68: Theoretical null model results with 95% bootstrap CIs.

Correlation	ρ	95% CI
ρ(magnitude, stability)	0.639	[0.612, 0.664]
ρ(SNR, stability)	0.999	[0.999, 0.999]
ρ_{partial}(magnitude, stability $\mid$ SNR)	0.292	[0.253, 0.329]

Interpretation. In the null model, the stability is almost perfectly predicted by the SNR ($\rho = 0.999$), with only a modest partial correlation with the magnitude after controlling for the SNR ($\rho_{\text{partial}} = 0.292$ [0.253, 0.329]). The partial correlations observed in the real data (Table 65) show substantially more heterogeneity, with Norman showing a strong negative partial correlation ($\rho = -0.859$) and Dixit showing a strong positive partial correlation ($\rho = 0.627$). This suggests that biological factors beyond the simple SNR confounding drive the magnitude-stability relationship in real data.

12.14 Ablation Studies

12.14.1 PCA Dimensionality

To assess sensitivity to the choice of PCA dimensions, we recomputed stability using 10, 20, 30, 50, and 100 principal components. All confidence intervals were computed using 10,000 bootstrap replicas.

Table 69: PCA dimensionality ablation. Magnitude-stability correlations (ρ) with 95% bootstrap CIs (10,000 iterations).

Dataset	10 PCs	20 PCs	30 PCs	50 PCs	100 PCs
Norman 2019	0.95 [0.93, 0.96]	0.94 [0.92, 0.96]	0.95 [0.93, 0.96]	0.95 [0.93, 0.96]	0.96 [0.95, 0.97]
Dixit 2016	0.67 [0.55, 0.77]	0.68 [0.56, 0.79]	0.70 [0.58, 0.80]	0.75 [0.64, 0.83]	0.79 [0.70, 0.86]

Rank-order consistency - We used *rank order* rather than correlation coefficients to evaluate the consistency of a perturbation's consistency across PCA components, based on both its stability and its magnitude. For Norman, ranked order stability and magnitude consistency across components is high (stability: $r = 0.98 \pm 0.02$, min $= 0.95$; magnitude: $r = 0.99 \pm 0.01$, min $= 0.98$). For Dixit, ranked order stability and magnitude consistency remained strong (r= 0.96 ± 0.04; min 0.87) and (r= 0.96 ± 0.04; min 0.88). This confirms that perturbation rankings are robust to different choices in dimensionality.

Key findings. Norman 2019 shows high stability between PCA dimensions ($\rho = 0.94$-0.96), with overlapping confidence intervals suggesting no significant dependence on dimensionality. Dixit 2016 shows a modest increase in correlation with more PCs ($\rho = 0.67$ to 0.79), suggesting that the higher dimensional structure contributes to the magnitude-stability relationship in this dataset. Importantly, all settings yield strong positive correlations ($\rho > 0.67$), and the overlapping confidence intervals indicate that the choice of 50 PCs has no bearing on the results.

12.14.2 Random Seed Reproducibility

To verify that the results were not dependent on random initialization, we recomputed stability using 15 different random seeds per data set: {3, 7, 9, 11, 12, 18, 103, 108, 320, 411, 724, 1754, 1991, 2222, 7258 }.

Table 70: Random seed ablation. Results are perfectly reproducible across 15 seeds.

Dataset	n	ρ (mean $\pm$ std)	95% CI	Cross-seed r
Norman 2019	236	0.945 ± 0.0000	[0.92, 0.96]	1.000
Dixit 2016	153	0.700 ± 0.0000	[0.58, 0.80]	1.000

Perfect reproducibility. Both datasets showed a perfect match in all correlation coefficients for each of the 15 random experimental seeds (cross-seed correlation = 1.000), which shows that all stochastic elements used during the preprocessing phase, such as the random initialization of PCA, have no effect on the final results of the analysis. This result was anticipated since when the PCA was performed on the same data, it produced exact matches each time it was run. Additionally, the ablation analysis demonstrates that there were no undetected sources of variability or unexplained stochasticity elsewhere in our analytic pipeline.

12.14.3 Leave-One-Out Influence Analysis

To verify that no single perturbation drives the observed correlation, we performed a leave-one-out analysis, recomputing the correlation after removing each perturbation in turn.

Table 71: Leave-one-out influence analysis. Shows robustness of magnitude-stability correlation to individual perturbations. "Helpful" perturbations prop up the correlation (removing them decreases ρ); "harmful" perturbations reduce it (removing them increases ρ).

Dataset	n	Full ρ	LOO range	Most helpful ($\Delta\rho$)	Most harmful ($\Delta\rho$)
Norman 2019	236	0.945	[0.945, 0.948]	BAK1 (+0.0007)	HES7 (-0.0024)
Dixit 2016	153	0.700	[0.694, 0.714]	ELK1 (+0.0060)	CREB1+E2F4+ELF1 (-0.0139)

Key findings. The LOO range is narrow for both datasets: removing any single perturbation changes the correlation by at most $\Delta\rho = 0.002$ (Norman) or $\Delta\rho = 0.014$ (Dixit). This confirms that the magnitude-stability relationship is a population-level phenomenon, not driven by individual outliers.

12.15 Single vs. Combinatorial Perturbations

Two datasets (Norman and Dixit) included both single-gene and combinatorial (multi-gene) perturbations, identified by the presence of "+" in the perturbation labels. Combinatorial perturbations exhibited systematically higher stability in both datasets (Table 72), with Mann-Whitney U tests confirming significance (Norman: $p < 10^{-15}$; Dixit: $p < 10^{-9}$).

Importantly, the magnitude-stability relationship was *maintained within* each category, ruling out a purely categorical explanation. The single-gene perturbations in Norman showed $\rho = 0.97$, while the combinatorial perturbations showed $\rho = 0.91$, both of which were highly significant.

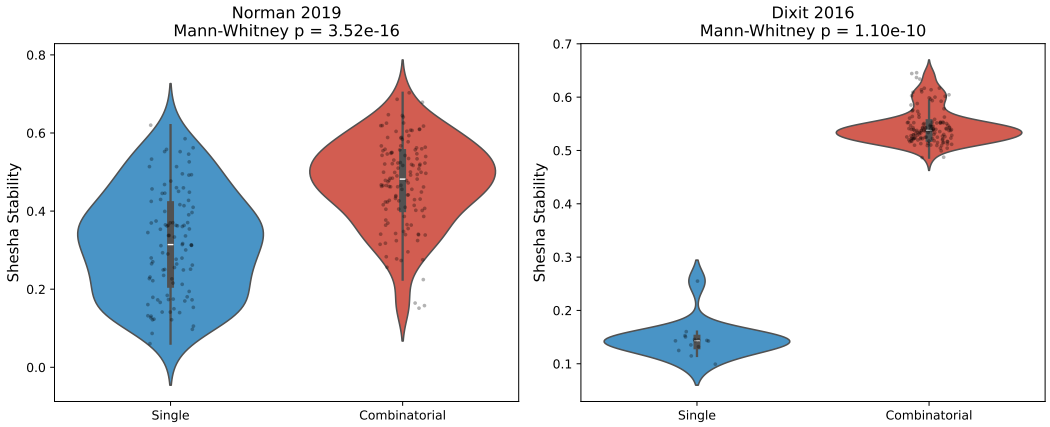

Figure 33: **Combinatorial perturbations exhibit higher geometric stability than single-gene perturbations.** Violin plots showing stability distributions for single-gene versus combinatorial (multi-gene) perturbations in Norman et al. (CRISPRa) and Dixit et al. (CRISPRi). Combinatorial perturbations show significantly higher stability in both datasets (Mann-Whitney U, $p < 10^{-9}$), suggesting that multi-target interventions produce more coherent transcriptional responses. Individual perturbations shown as points.

Table 72: Single-gene versus combinatorial perturbations. The magnitude-stability relationship holds within both categories.

Dataset	Type	N	Stability	Magnitude	ρ	p
Norman 2019	Single	105	0.32 ± 0.14	1.18 ± 0.73	0.97	$< 10^{-64}$
Norman 2019	Combinatorial	131	0.47 ± 0.11	2.03 ± 0.88	0.91	$< 10^{-49}$
Dixit 2016	Single	15	0.15 ± 0.03	0.66 ± 0.12	0.94	$< 10^{-6}$
Dixit 2016	Combinatorial	138	0.54 ± 0.03	2.03 ± 0.11	0.65	$< 10^{-17}$

12.16 Discordant Cases: Biological Interpretation

To assess whether stability captures information beyond magnitude, we identified perturbations with high discordance, defined as the difference between z-scored magnitude and z-scored stability within each dataset.

12.16.1 Norman 2019 Discordant Cases

In Norman et al., high-magnitude/low-stability perturbations (positive discordance) were dominated by CEBPA combinations (Table 73). CEBPA is a master regulator of myeloid differentiation with

known pleiotropic effects (Friedman, 2007) across 24 downstream pathways (Norman et al., 2019). This is consistent with the interpretation that low stability (relative to magnitude) reflects regulatory promiscuity. Conversely, high-stability/low-magnitude perturbations (negative discordance) were enriched for KLF1, a lineage-specific erythroid transcription factor (Miller and Bieker, 1993) targeting a specific subset of erythroid genes (e.g., globins) within the assayed panel (Norman et al., 2019; Siatecka and Bieker, 2011). This suggests that stability indices regulatory specificity: KLF1 produces perturbations that are geometrically coherent despite their lower magnitude.

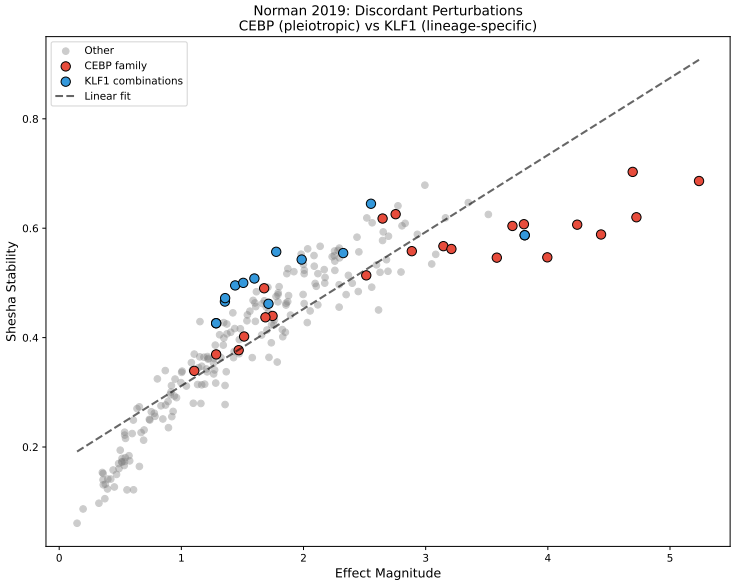

Figure 34: **Discordant perturbations reveal regulatory specificity.** Stability versus magnitude for all perturbations in Norman et al. (2019), with CEBP family members (CEBPA, CEBPB, CEBPE) and KLF1 combinations highlighted. CEBP perturbations cluster **below** the trend line (lower stability relative to their high magnitude), consistent with CEBPA's known role as a pleiotropic master regulator. KLF1 perturbations cluster **above** the trend line (higher stability relative to magnitude), consistent with its lineage-specific role in erythroid development. Dashed line shows linear fit to all perturbations.

Table 73: Top discordant perturbations in Norman et al. (2019). Positive discordance (High magnitude / Relative low stability) highlights pleiotropic regulators like CEBPA. Negative discordance (Moderate magnitude / Relative high stability) highlights lineage-specific factors like KLF1.

Positive Discordance: High Magnitude / Relative Low Stability			
Perturbation	Magnitude	Stability	Discordance
CEBPA+JUN	5.24	0.69	1.95
CEBPA	4.73	0.62	1.85
CEBPA+CEBPB	4.44	0.59	1.76
CEBPA+ZC3HAV1	4.00	0.55	1.57
CEBPA+CEBPE	4.24	0.61	1.42
Negative Discordance: Moderate Magnitude / Relative High Stability			
Perturbation	Magnitude	Stability	Discordance
KLF1+SET	1.78	0.56	−0.93
KLF1	1.44	0.50	−0.87
KLF1+TGFBR2	1.51	0.50	−0.83
BAK1+KLF1	1.36	0.47	−0.80
AHR+KLF1	1.60	0.51	−0.78

12.16.2 Quartile Statistics

Table 74: Norman 2019 quartile statistics by discordance. Q1 captures high-specificity perturbations (KLF1), while Q4 captures pleiotropic/variable perturbations (CEBPA).

Quartile	Stability (mean ± SD)	Magnitude (mean ± SD)	Median n_{cells}	n
Q1 (Negative Discordance / Specific)	0.502 ± 0.074	1.82 ± 0.47	392	59
Q2	0.423 ± 0.092	1.60 ± 0.58	379	59
Q3	0.362 ± 0.109	1.46 ± 0.71	299	59
Q4 (Positive Discordance / Pleiotropic)	0.327 ± 0.202	1.72 ± 1.51	303	59

Pattern. The variance of the positively discordant quartile (Q4), which is enriched for pleiotropic factors, shows significantly higher variance in both stability (SD = 0.202 vs 0.074 in Q1) and magnitude (SD = 1.51 vs 0.47 in Q1). This is consistent with the biological distinction between master regulators like CEBPA, which induce broad and variable state changes (Friedman, 2007; Avellino and Delwel, 2017), and lineage-specific factors like KLF1, which produce precise, geometrically coherent shifts (Miller and Bieker, 1993; Siatecka and Bieker, 2011). The median cell counts are similar across all quartiles (Q1-Q4 median n ≈ 300-400); thus, the variance of the patterns cannot be attributed to sample size effects.

12.16.3 Biological Context

CEBPA (pleiotropic). CEBPA (CCAAT/enhancer-binding protein alpha) is a master transcription factor that controls myeloid differentiation, adipogenesis, and energy metabolism (Friedman, 2015; Rosen et al., 2002). It affects genes across immune response, cell cycle, metabolism, and differentiation pathways. This broad regulatory scope explains why CEBPA perturbations produce large but geometrically incoherent responses: cells activate diverse, partially independent programs rather than a single coherent transcriptional trajectory (Norman et al., 2019).

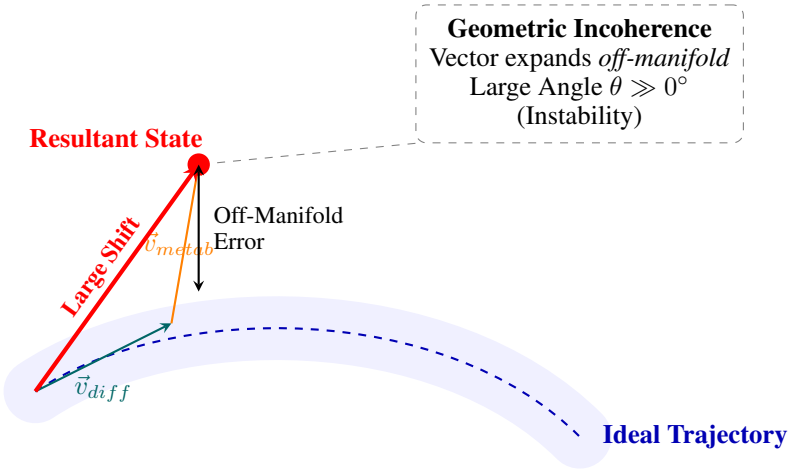

Figure 35: **Geometric Instability of CEBPA.** Unlike the coherent arrest seen in KLF1, CEBPA targets drive orthogonal programs. $\vec{v}_{metab}$ pushes the cell state out of the differentiation manifold (blue tube), resulting in a geometrically incoherent state ($\theta \gg 0$) that cannot be mapped back to a valid lineage trajectory.

KLF1 (lineage-specific). KLF1 (Krüppel-like factor 1) is an erythroid-specific transcription factor essential for β-globin expression and red blood cell maturation (Tallack et al., 2010; Tallack and Perkins, 2010). While it regulates a wide array of targets, ranging from globin clusters to cell cycle regulators, these programs are strictly coordinated to drive terminal differentiation (Siatecka and Bieker, 2011). This lineage-restricted functional constraint produces geometrically coherent responses: perturbation vectors align with the principal differentiation manifold closely, causing cells

to arrest uniformly along the erythroid trajectory rather than scattering into orthogonal states (Pilon et al., 2008).

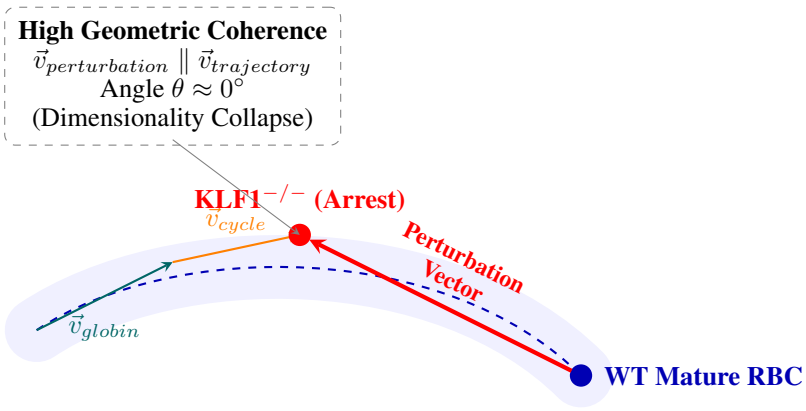

Figure 36: **Geometric Coherence of KLF1.** Unlike CEBPA, the downstream targets of KLF1 (globins, cell cycle) generate vectors ($\vec{v}_{globin}, \vec{v}_{cycle}$) that are locally collinear with the erythroid differentiation manifold (blue tube). Consequently, perturbation results in a magnitude shift along the trajectory (red vector, differentiation arrest) rather than an incoherent expansion into off-manifold space ($\theta \approx 0°$).

This dissociation between magnitude and stability thus appears to capture biologically meaningful variation in regulatory architecture, distinguishing pleiotropic master regulators from lineage-specific factors.

12.17 Summary of Statistical Findings

Table 75: Summary of CRISPR validation findings with 95% bootstrap CIs.

Analysis	Finding	Value	95% CI
Method Robustness			
	Best single-dataset correlation	0.985	[0.939, 0.997]
	Worst single-dataset correlation	0.746	[0.641, 0.827]
	Cross-method consistency	All $\rho > 0.74$	—
Confound Control			
	Pooled partial correlation (SNR)	-0.258	$[-0.423, -0.111]$
	Mixed-effects magnitude β	0.123	[0.116, 0.131]
	Mixed-effects spread β	-0.122	$[-0.158, -0.086]$
	Mixed-effects sample size β	-0.031	$[-0.039, -0.024]$
	Magnitude vs sample size ratio	4.0×	—
Cross-Dataset Generalization			
	Calibrated pooled correlation	0.913	[0.884, 0.936]
	Dataset random effect variance	≈0	—
Reproducibility (Ablations)			
	Cross-seed correlation	1.000	—
	PCA dimension range (Norman)	0.94-0.96	overlapping CIs
	PCA dimension range (Dixit)	0.67-0.79	overlapping CIs
	Rank-order consistency (stability)	>0.95	—
	LOO max influence (Norman)	$\Delta\rho = 0.002$	—
	LOO max influence (Dixit)	$\Delta\rho = 0.014$	—

12.18 Limitations

Several caveats apply to this analysis:

- **Adamson sample size:** With only $n = 8$ perturbations, Adamson 2016 provides limited statistical power. The wide bootstrap confidence intervals ([0.407, 1.000]) honestly reflect this uncertainty, but conclusions from this dataset should be interpreted cautiously.
- **Papalexi sample size:** With only $n = 25$ gene-level perturbations, Papalexi 2021 also provides modest statistical power. However, the very strong correlation ($\rho = 0.985$ [0.939, 0.997]) and proper control identification (2,386 NT cells pooled from 6 non-targeting guides) give confidence in this result.
- **Partial correlation heterogeneity:** The striking heterogeneity in partial correlations across datasets (Norman: $\rho = -0.859$; Dixit: $\rho = 0.627$; Papalexi: $\rho = 0.176$) suggests that the relationship between magnitude, SNR, and stability may be dataset-specific. The pooled estimate ($\rho = -0.258$) is dominated by Norman (the largest dataset) and may not generalize uniformly.
- **Control group heterogeneity:** Different datasets used different control strategies: Norman and Dixit used cells labeled "control," Adamson used NaN-converted control labels, and Papalexi used pooled non-targeting guides (NT). While we standardized control identification procedures, residual differences in control quality may affect cross-dataset comparisons.
- **PCA embedding:** When analyzing the stability of the cell states, PCA embedding was used to capture the linear variance, but it may miss the non-linear geometry of the manifold. Future work could explore more sophisticated embeddings, such as UMAP or diffusion maps to see how the stability would change.
- **Biological validation:** Biological validation in the form of systematic experimental validation of the stability-specificity hypothesis would strengthen our interpretation, and the discordant cases can be mapped onto known biological understanding, e.g. CEBPA vs KLF1.

13 Neuroscience Drift and Behavior: Extended Methods and Results

This appendix contains all the methods and additional analyses related to the neuroscience experiment outlined in Section 3.5. We evaluated geometric stability in brain recordings from the Steinmetz et al. (2019) Neuropixels dataset (26 sessions, 229 area-session observations, 68 brain regions) with three complementary analyses. First we establish a behavioral ground truth validation to test whether Shesha predicts trial-by-trial neural-behavioral coupling. Second, we test regional hierarchy characterization to compare geometric stability (Shesha) against temporal stability (centroid drift) over functional brain systems. Finally, we conduct a permutation null model validation to confirm that observed drift reflects genuine temporal structure rather than measurement noise.

13.1 Dataset

We analyzed neural recordings from the Steinmetz et al. (2019) Neuropixels dataset (Steinmetz et al., 2019), which contains simultaneous single cell recordings from multiple brain regions in multiple mice as they performed a visual contrast discrimination task. The mice were presented with visual stimuli of varying contrast on the left and right sides. To respond, they moved a wheel to drive the higher-contrast stimulus to the center of the screen or refrained from moving the wheel if no stimuli were present (a 'no-go' response). The mice were given feedback on each trial: water as a reward for correct responses and white noise for incorrect responses.

We chose to include all sessions that had at least 60 trials to ensure sufficient data for reliable split-half estimation. After filtering, this resulted in 26 sessions across multiple subjects, with 229 area-session observations (brain areas with $\geq$10 simultaneously recorded neurons) spanning 68 unique brain areas.

13.2 Preprocessing

For each session, we extracted spike counts in 10 ms time bins, which were aligned to the onset of the stimulus. Our analysis focused on the decision epoch (0-500 ms post-stimulus). The spike counts were averaged over this decision epoch window to obtain a single response vector per neuron per trial. This gives us a matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$ that represents the population response, where T represents the number of trials, and N denotes the number of neurons in a given brain area.

To prevent the stability metric from being driven by overall fluctuations of the firing rates, each trial population vector was L2-normalized:

$$\tilde{\mathbf{x}}_t = \frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_2}$$

This normalization renders the subsequent dot product equivalent to cosine similarity, effectively safeguards against multiplicative rate changes.

13.3 Metrics

We computed two different stability metrics to measure different aspects of neural population geometry:

Geometric Stability (Shesha). To measure the reliability of representational geometry, we computed split-half RDM correlations. The trials were separated into odd and even subsets, and condition-averaged responses were computed for each unique stimulus combination (9 contrast pairings) within each subset. The representational dissimilarity matrix (RDM) was computed as the pairwise cosine distances between condition centroids:

$$\text{RDM}_{ij} = 1 - \cos(\mathbf{c}_i, \mathbf{c}_j)$$

Geometric stability (Shesha) was defined as the Spearman correlation between the odd and even RDMs:

$$S_{\text{Shesha}} = \rho_s(\text{RDM}_{\text{odd}}, \text{RDM}_{\text{even}})$$

This metric captures how reliably the pairwise distance structure is preserved across independent data splits. Higher values indicate more a stable representational geometry.

Centroid Drift. To measure temporal consistency, we divided each session into early and late epochs at the median trial. For each brain area, we computed the centroid (mean population vector) for both epochs:

$$\mathbf{c}_{\text{early}} = \frac{1}{|\mathcal{T}_{\text{early}}|} \sum_{t \in \mathcal{T}_{\text{early}}} \tilde{\mathbf{x}}_t, \quad \mathbf{c}_{\text{late}} = \frac{1}{|\mathcal{T}_{\text{late}}|} \sum_{t \in \mathcal{T}_{\text{late}}} \tilde{\mathbf{x}}_t$$

Centroid drift was defined as the cosine similarity between these centroids:

$$S_{\text{drift}} = \frac{\mathbf{c}_{\text{early}} \cdot \mathbf{c}_{\text{late}}}{\|\mathbf{c}_{\text{early}}\|_2 \, \|\mathbf{c}_{\text{late}}\|_2}$$

Values near 1 indicate that the population geometry is preserved across the session, while lower values indicate representational drift.

Whitened Unbiased Cosine (WUC). We additionally computed WUC (Diedrichsen et al., 2021), which addresses noise covariance bias by whitening representations before computing RDMs. Whitening matrices were estimated with Ledoit-Wolf shrinkage (Ledoit and Wolf, 2004) ($\lambda = 0.1$) to ensure numerical stability.

Similarity Metrics. For comparison, we computed standard representational similarity metrics on condition-averaged responses using odd/even trial splits:

- **CKA**: Linear Centered Kernel Alignment
- **Procrustes**: Distance after optimal rotation alignment
- **RSA**: Spearman correlation of representational dissimilarity matrices

13.4 Behavioral Ground Truth Measures

To test whether geometric stability captures functionally relevant structure, we computed multiple behavioral ground truth measures:

Trial-by-trial neural-behavioral coupling. For each area-session, we computed the Spearman correlation between trial-wise neural state magnitude ($\|\mathbf{x}_t\|_2$) and trial outcome (correct/incorrect). This measures how tightly neural activity tracks behavioral performance on individual trials.

Session-level accuracy change. We computed the change in choice accuracy from early to late trials:

$$\Delta_{\text{acc}} = \text{Acc}_{\text{late}} - \text{Acc}_{\text{early}}$$

Across sessions, the mean accuracy change was -0.103 ± 0.091 (range: -0.323 to $+0.056$), indicating that the animals mice performed worse in the latter half of sessions.

Additional behavioral measures. We also tested correlations with mean accuracy and choice consistency (variance in choices for similar stimuli).

13.5 Permutation Null Model

To determine if the centroid drift observed over time is indicative of a true temporal structure, a permutation null model was created. For each area-session, an individual permuted the trial order 500 different times, then calculated the centroid similarity on each permutation. This was done to remove temporal structure while maintaining the marginal distribution of responses from the entire population of neurons.

We computed a z-score for each observation:

$$z = \frac{S_{\text{observed}} - \mu_{\text{null}}}{\sigma_{\text{null}}}$$

where μ_{null} and σ_{null} are the mean and standard deviation of the null distribution. Negative z-scores indicate systematic drift beyond what would be expected from random trial-to-trial variability.

13.6 Region Grouping

Brain areas were grouped into functional categories based on the Allen Brain Atlas ontology:

- **Visual**: VISp, VISl, VISrl, VISam, VISpm, VISa, VISal
- **Thalamus**: LP, LD, LGd, VPM, PO, MD
- **Motor**: MOp (primary), MOs (secondary)
- **Frontal**: ACA, PL, ILA, ORB
- **Hippocampus**: CA1, CA3, DG, SUB
- **Striatum**: CP, ACB, LS
- **Midbrain**: SNr, SCm, MRN, ZI
- **Other**: All remaining areas

We report 95% confidence intervals computed via bootstrap resampling (10,000 iterations) for means and correlations.

13.7 Results

13.7.1 Behavioral Ground Truth Validation

The primary validation for geometric stability in neural data is its ability to predict functionally relevant behavioral outcomes. We tested Shesha against multiple behavioral ground truths (Table 76).

Table 76: Behavioral ground truth validation ($n = 228$ area-sessions unless noted).

Ground Truth	Metric	Spearman ρ	95% CI	p
Significant Results				
Trial neural-behavior	Shesha	$+0.184$	$[+0.05, +0.31]$	0.005
Non-significant Results				
Trial neural-behavior	Centroid drift[a]	$+0.002$	$[-0.13, +0.13]$	0.976
Trial neural-behavior	WUC[a]	$+0.089$	$[-0.04, +0.21]$	0.180
Mean accuracy	Shesha	$+0.087$	$[-0.05, +0.22]$	0.191
Accuracy change*	Shesha	-0.079	$[-0.47, +0.34]$	0.701

*Session-level analysis ($n = 26$ sessions). [a] $n = 229$.

Shesha predicted trial-by-trial neural-behavioral coupling ($\rho = 0.184$, 95% CI: $[0.05, 0.31]$, $p = 0.005$), indicating that brain regions with more stable representational geometry show tighter correspondence between neural activity and behavioral outcomes on individual trials. Noteably, this observation was limited to geometric stability and was not observed for centroid drift ($\rho = 0.002$, $[-0.13, 0.13]$, $p = 0.976$) or WUC ($\rho = 0.089$, $[-0.04, 0.21]$, $p = 0.180$). This suggests that Shesha captures a functionally relevant dimension of neural population dynamics that extends beyond simple temporal consistency or whitened similarity measures.

Consistent with prior work, session-level stability did not predict behavioral maintenance ($\rho = -0.079$, $[-0.47, 0.34]$, $p = 0.70$) (Rule et al., 2019). This suggests that drift at this timescale may be behaviorally silent.

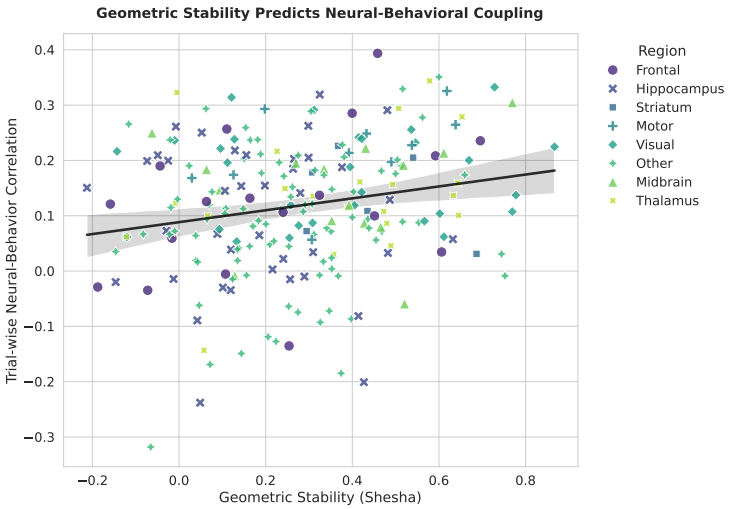

Figure 37: **Geometric stability predicts neural-behavioral coupling.** Each point represents one brain area in one session ($n = 228$). Geometric stability (Shesha) correlates significantly with trial-by-trial neural-behavioral coupling ($\rho = 0.18$, $p = 0.005$), indicating that regions with more stable representational geometry show tighter correspondence between neural state magnitude and behavioral outcome. Points are colored by brain region. Black line shows linear regression with 95% confidence band.

13.7.2 Regional Hierarchy

Shesha revealed a distinct regional hierarchy of geometric stability (Table 77). The regions that showed the highest geometric stability were the striatum (0.44, $[0.34, 0.56]$), the motor cortex (0.38, $[0.25, 0.50]$), and the visual cortex (0.36, $[0.27, 0.46]$). The region that showed the lowest was the hippocampus (0.19, $[0.13, 0.25]$).

Table 77: Regional geometric stability (Shesha). Regions ordered by mean stability.

Region	n	Mean S	95% CI
Striatum	6	0.438	$[0.336, 0.555]$
Motor	10	0.377	$[0.253, 0.496]$
Visual	29	0.364	$[0.265, 0.463]$
Thalamus	21	0.359	$[0.256, 0.458]$
Midbrain	15	0.354	$[0.244, 0.461]$
Other	90	0.253	$[0.214, 0.292]$
Frontal	18	0.222	$[0.103, 0.345]$
Hippocampus	39	0.187	$[0.128, 0.247]$

This was different from centroid drift (Table 78), where sensory regions showed the highest temporal stability (thalamus: 0.95; Visual: 0.94) and the striatum showed the most drift (0.83). The dissociation between geometric stability and temporal drift suggests that these metrics capture complementary aspects of neural population structure.

Table 78: Regional centroid drift. Regions ordered by mean stability.

Region	n	Mean S	95% CI
Thalamus	21	0.946	$[0.924, 0.965]$
Hippocampus	39	0.944	$[0.928, 0.958]$
Visual	29	0.940	$[0.925, 0.954]$
Midbrain	15	0.932	$[0.900, 0.961]$
Other	90	0.924	$[0.907, 0.939]$
Frontal	18	0.889	$[0.843, 0.926]$
Motor	11	0.872	$[0.810, 0.924]$
Striatum	6	0.825	$[0.759, 0.895]$

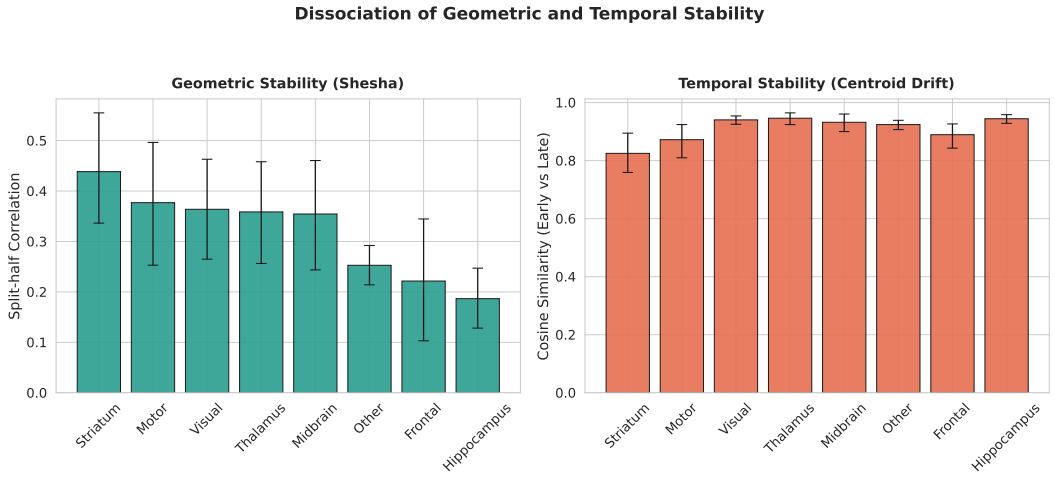

Figure 38: **Regional hierarchy of geometric vs. temporal stability. (A)** Geometric stability (Shesha) is highest in action-related regions (Striatum, Motor) and lowest in Hippocampus. **(B)** Temporal stability (centroid similarity) shows an opposing pattern, with sensory regions (Thalamus, Visual) most stable and Striatum least stable. This dissociation indicates that geometric and temporal stability capture independent dimensions of neural population dynamics. Error bars show 95% bootstrap confidence intervals. Regions are ordered by geometric stability in both panels.

We additionally computed WUC (Whitened Unbiased Cosine). WUC showed lower values overall (mean: 0.10, range: -0.23 to 0.40) and was only moderately correlated with Shesha ($\rho = 0.27$, $[0.15, 0.39]$). This suggests that these stability metrics capture partially distinct aspects of representational reliability. WUC showed a similar but attenuated regional pattern. In the motor cortex (0.20) and the striatum (0.19), it showed the highest values. In the hippocampus, it showed the lowest (0.03).

13.7.3 Null Model Validation

The observed centroid drift was significantly lower than the permutation-based chance levels across nearly all area-session observations (Table 79). The mean z-score was -44.7 (95% CI: $[-49.2, -40.4]$), which indicates that the observed centroid similarity was approximately 45 standard deviations below the null expectation. The observed mean stability (0.924, 95% CI: $[0.915, 0.934]$) was substantially lower than the null expectation (0.995, 95% CI: $[0.995, 0.996]$), with non-overlapping confidence intervals.

Table 79: Null model validation summary.

Statistic	Value	95% CI
Total observations	229	–
Mean z-score	-44.7	$[-49.2, -40.4]$
Observed mean stability	0.924	$[0.915, 0.934]$
Null mean stability	0.995	$[0.995, 0.996]$

13.7.4 Tertile Dynamics

The drift accumulated gradually across the sessions rather than occurring abruptly. The mean stability for the early-middle interval ($S = 0.942$, 95% CI: $[0.933, 0.950]$) did not differ from that of the middle-late interval ($S = 0.941$, 95% CI: $[0.933, 0.948]$). The difference between intervals was -0.001 (95% CI: $[-0.007, 0.009]$; paired t-test, $t = 0.30$, $p = 0.77$), which indicates a drift that is continuous rather than concentrated.

13.8 Interpretation

Behavioral ground truth validates functional relevance. The significant correlation between Shesha and trial-by-trial neural-behavioral coupling ($\rho = 0.18$, $p = 0.005$) demonstrates that geometric stability captures functionally relevant structure in neural populations. Brain regions with representational geometry that was more reliable show tighter correspondence between neural activity and behavioral outcomes. This effect was specific to Shesha. It was not observed for centroid drift or WUC, which suggests that geometric stability measures a distinct and behaviorally meaningful dimension of neural population dynamics.

Regional hierarchy reveals distinct stability dimensions. Different regional level patterns of Shesha and centroid drift were observed. The striatum showed the highest geometric stability but lowest temporal stability. The hippocampus showed the opposite. This dissociation suggests that geometric stability (reliability of the pairwise distance structure) and temporal stability (preservation of the population centroid) capture complementary aspects of neural population organization. The high level of geometric stability in the striatum may reflect consistent encoding of action-reward associations, while its low temporal stability may reflect the rapid updating of value representations.

The permutation null model reveals systematic drift. The observed centroid similarity was consistently lower than the null expectation (0.924 vs. 0.995), with non-overlapping confidence intervals. This indicates that early and late population responses are more dissimilar than shuffled data would predict, which confirms systematic representational drift. The gradual accumulation of drift ($\Delta S = 0.001$, $p = 0.77$ between tertiles) suggests an adaptation that is continuous rather than state changes that are discrete.

Session-level behavioral coupling remains null. The absence of session-level behavioral coupling ($\rho = -0.08$, $p = 0.70$) suggests that task performance cannot be predicted by drift at this timescale. This may reflect: (1) behavioral relevance depending on drift in task-relevant subspaces rather than global geometry; (2) compensatory mechanisms that maintain performance despite representational drift; or (3) insufficient power with $n = 26$ sessions. The significant trial-level coupling suggests that finer-grained analyses may be more sensitive to stability-behavior relationships.

14 Broader Impact

The reliability gap in high-stakes deployment. Current evaluation practices assume that high task accuracy implies safe deployment, yet this assumption fails catastrophically in high-stakes domains. A model that performs well on benchmarks may exhibit brittle internal geometry that fractures under distribution shift, leading to silent failures in healthcare diagnostics, autonomous navigation, or financial decision-making. Geometric stability provides an auditing framework that exposes this gap: systems can now be certified not only for *what* they represent, but for *how reliably* they maintain that structure under perturbation. For safety-critical applications where consistent behavior is non-negotiable, stability-based certification offers a principled guardrail before deployment.

Hidden costs of post-training intervention. The widespread adoption of fine-tuning and alignment techniques raises an underappreciated risk: these interventions may induce substantial internal reorganization that current monitoring tools fail to detect. A model that appears functionally equivalent before and after alignment may have undergone geometric restructuring that compromises its reliability under novel inputs. Stability metrics serve as an early warning system, flagging representational drift before it manifests as task failure. This "canary in the coal mine" capability is essential for responsible AI development, enabling practitioners to detect the onset of degradation rather than discovering it through downstream failures.

Toward an engineering discipline for representations. The need for geometric stability extends well beyond machine learning. Scientific domains increasingly rely on learned representations for discovery: molecular embeddings guide drug design, neural population models inform brain-computer interfaces, and cellular encodings enable disease diagnosis. In each case, representational reliability is paramount, as fragile geometry translates directly to unreliable scientific conclusions. By providing a domain-agnostic framework for quantifying reliability, geometric stability helps move representation learning from empirical alchemy toward a rigorous engineering discipline with verifiable safety guarantees.

Ethical considerations and scope. We acknowledge that stability metrics, like any diagnostic tool, could potentially be misused to identify vulnerable models for adversarial exploitation. However, we believe the benefits of transparent auditing substantially outweigh this risk: defenders gain more from knowing which models are fragile than attackers gain from the same knowledge. Additionally, our framework currently operates on global representational structure and may not capture fine-grained, token-level instabilities relevant to certain attack vectors. We encourage the community to develop complementary local stability measures to address this gap.