COMPARING AND INTEGRATING DIFFERENT NOTIONS OF REPRESENTATIONAL CORRESPONDENCE IN NEURAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representational similarity metrics emphasize different properties, so their agreement with one scientific criterion need not extend to another. We compare seven widely used metrics using hypothesis-driven tests of within-group correspondence and between-group separation in artificial and biological neural systems. The benchmarks include 35 vision models evaluated on ImageNet, Ecoset, CIFAR10, and CIFAR100, human fMRI responses from the Natural Scenes Dataset, and mouse Neuropixels recordings from the Allen Brain Observatory. Geometry- and tuning-sensitive metrics generally recover stronger model-family and anatomical-group separation than more flexible mappings, but the leading metric changes with dataset and separation criterion. In a population of independently trained seeds, the same metrics separate training objectives beyond seed-to-seed variation, while analyses at six network depths show that intermediate-layer rankings differ from final-layer rankings. Integrating the metrics with Similarity Network Fusion improves most group-separation scores and yields data-driven typologies in which training paradigm can cut across architecture, but the fused similarity does not uniformly outperform individual metrics. These results demonstrate metrics' different abilities in separating neural systems, while also showing that two different neural systems could still share great similarity in some representational aspects.

1 INTRODUCTION

Understanding when two neural systems, biological or artificial, rely on equivalent internal representations is a central challenge in neuroscience and machine learning. Researchers typically compare representations using a single similarity metric, such as Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008), Centered Kernel Alignment (CKA) (Kornblith et al., 2019), or linear predictivity (Yamins et al., 2014). Crucially, these metrics are not interchangeable. Each metric captures a distinct notion of correspondence, preserving representational geometry, aligning unit-level tuning, or measuring alignment in linearly accessible information, and therefore makes distinct assumptions about which representational transformations should be considered irrelevant. As a result, different metrics can yield qualitatively different conclusions about which systems are similar or distinct (Soni et al., 2024; Cloos et al., 2024).

This observation points to a deeper issue: representational similarity is not a single notion of correspondence. Some properties may be shared across systems, while others retain signatures of architecture, training objective, or biological organization. A high similarity score alone does not establish that a metric captures the structure relevant to a scientific question. For example, a comparison intended to detect the effects of a training objective should distinguish those effects from variation across random seeds. Similarly, a comparison across brains should preserve correspondence within a region without erasing differences between regions. We therefore ask which metrics retain these expected relative relationships.

Separation provides a test of this question by contrasting within-group with between-group similarities, rather than rewarding large distances alone. In artificial neural networks, we evaluate *procedural correspondence*: whether models with related architectures or training paradigms are more similar than models from different families. In biological data, we evaluate correspondence within and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

between anatomically defined groups. These are hypothesis-driven criteria, not ground truth for representational equivalence: different procedures can converge, and systems in the same group can differ. Separation is useful because it tests sensitivity to a specified source of variation together with consistency within a group; it does not establish behavioral equivalence or make a metric universally preferable (Ding et al., 2021; Bo et al., 2024).

Using vision models and biological recordings as a testbed, our work makes three contributions. First, we compare representational similarity metrics across four image benchmarks, human fMRI responses from the Natural Scenes Dataset (NSD), and mouse Neuropixels recordings from the Allen Brain Observatory (ABO). Metrics that preserve representational geometry or unit-level tuning often recover stronger group structure than more flexible mappings, but their ordering depends on the dataset and evaluation criterion. Second, random-seed and network-depth analyses test the limits of this pattern: in a population of independently trained seeds, these metrics separate training objectives beyond seed-to-seed variation, whereas penultimate-layer rankings need not hold at intermediate layers.

Finally, we use the existing Similarity Network Fusion (SNF) method (Wang et al., 2014) to integrate the similarity graphs of all metrics and build data-driven typologies of models and brain regions. Fusion improves most group-separation scores, and clustering the fused matrix shows that training paradigm can override architecture (for example, self-supervised ResNets group with a self-supervised ViT rather than with supervised ResNets), while in NSD it recovers the ventral-stream hierarchy most closely. Its advantage is not uniform, however, so fusion complements rather than replaces the comparison of individual metrics. Our main aim is to identify which notions of correspondence support which scientific comparisons across models and biological data.

2 METHODS

2.1 MODEL SELECTION AND DATASET

To evaluate how different representational similarity measures recover known procedural structure, we curated a controlled population of vision models in which architecture and training objective vary systematically while other factors are held fixed. We define a model's procedure as the combination of its architecture and training paradigm, and treat models sharing the same procedure as belonging to the same model family. We analyze 35 vision models across four primary categories: supervised Convolutional Neural Networks (CNNs), self-supervised CNNs, supervised Transformers, and self-supervised Transformers. We treat ConvNeXt (Liu et al., 2022) and Swin (Liu et al., 2021) as distinct families due to their hybrid nature: ConvNeXt incorporates Transformer-inspired design principles within a convolutional architecture, while Swin introduces CNN-like inductive biases into the Transformer framework. We deliberately restricted the model set to encoders pre-trained on ImageNet-1k. This design isolates the effects of architecture and training objective within a shared data and label space so that differences in training data distribution do not impact the results. For datasets, we use the ImageNet-1k (Deng et al., 2009) and Ecoset (Mehrer et al., 2021) validation sets and CIFAR10 (Krizhevsky, 2009) and CIFAR100 (Krizhevsky, 2009) test sets. More details are provided in Appendix A and B. For brain data, we consider fMRI responses to the 1,000 shared natural images from NSD (Allen et al., 2022), spanning 10 visual regions across 4 subjects, and mouse Neuropixels responses during natural-movie viewing from ABO (Siegle et al., 2021). The ABO separation benchmark uses 41 subdivisions in nine groups defined by CCFv3 ancestry (Wang et al., 2020). NSD tests cross-subject region correspondence, whereas ABO tests correspondence within anatomical groups (Appendix J). Additional model analyses use independently trained seeds and six normalized depths (Section 3.2).

2.2 REPRESENTATIONAL METRICS

We evaluate widely used similarity metrics that differ in the flexibility of the mappings they permit. Each metric compares two networks along a different representational facet (Table C.1). For every network pair, representations are denoted by $\mathbf{X}_i \in \mathbb{R}^{M \times N_i}$ and $\mathbf{X}_j \in \mathbb{R}^{M \times N_j}$, where M denotes the number of stimuli and N_i, N_j denote the number of units. All representations are mean-centered along the sample dimension. For metrics requiring a fitting procedure (e.g., Soft Matching, Linear Predictivity, Procrustes), similarity values reflect the mean 5-fold cross-validation score.

Singular Vector Canonical Correlation Analysis (Raghu et al., 2017). SVCCA first applies singular value decomposition (SVD) to the representation matrices from two models or layers to isolate their most informative directions: $\mathbf{X}_i = \mathbf{U}_i \Sigma_i \mathbf{V}_i^\top$ and $\mathbf{X}_j = \mathbf{U}_j \Sigma_j \mathbf{V}_j^\top$. Retaining the right singular vectors that explain 99% of the variance gives the reduced representations $\mathbf{X}_i' = \mathbf{X}_i \mathbf{V}_{i,99}$ and $\mathbf{X}_j' = \mathbf{X}_j \mathbf{V}_{j,99}$, which preserve the dominant variance structure. Canonical correlation analysis (CCA) (Hardoon et al., 2004) is then applied to these reduced matrices to find linear projections $\mathbf{A}$ and $\mathbf{B}$ that maximize their correlation $\max_{\mathbf{A},\mathbf{B}} \mathbf{corr}\big(\mathbf{X}_i'\mathbf{A},\ \mathbf{X}_j'\mathbf{B}\big)$ subject to unit-variance constraints, providing a final similarity score that reflects how closely the subspaces of two representations align.

Projection-Weighted Canonical Correlation Analysis (Morcos et al., 2018). Compared to SVCCA, PWCCA did not apply SVD before CCA and it re-weights the canonical correlations. After CCA, we obtain canonical variates for the two representations and their corresponding correlations q_r for $r = 1, 2, \ldots, k$, where $k = \min\{N_i, N_j\}$. Let $\mathbf{Z} = \mathbf{X}_i$ if $N_i \leq N_j$, and $\mathbf{Z} = \mathbf{X}_j$ otherwise. For the selected representation, let $\mathbf{p}_r$ denote the rth QR-orthonormalized canonical variate. The projection weight is $\alpha_r = \frac{\|\mathbf{p}_r^\top \mathbf{Z}\|_1}{\sum_{s=1}^{k} \|\mathbf{p}_s^\top \mathbf{Z}\|_1}$. The final PWCCA similarity is then the weighted sum of canonical correlations: $\sum_{r=1}^{k} \alpha_r q_r$, which emphasizes canonical variates with larger projection magnitude onto $\mathbf{Z}$ and down-weights noisy components.

Linear Centered Kernel Alignment (Kornblith et al., 2019). Linear CKA provides a scalar measure of how similarly two sets of representations capture the relationships among the same collection of samples. It is defined as $\frac{\|\mathbf{X}_i^T \mathbf{X}_j\|_F^2}{\|\mathbf{X}_i^T \mathbf{X}_i\|_F \|\mathbf{X}_j^T \mathbf{X}_j\|_F}$, where $\|.\|_F$ is the Frobenius norm, and $\mathbf{X}_i$ and $\mathbf{X}_j$ can be assumed to be normalized. Linear CKA is invariant to orthogonal transformations (rotations or reflections), isotropic scaling, and translations of the representations, so it captures only the relational structure shared between the two spaces.

Representational Similarity Analysis (Kriegeskorte et al., 2008). Representational Similarity Analysis (RSA) compares the geometry of representations via their Representational Dissimilarity Matrices (RDMs). For each representation, we compute pairwise dissimilarities between stimuli using $1 -$ Pearson correlation, yielding an $M \times M$ RDM that encodes the relational structure. Model similarity is then quantified as the Pearson correlation between their RDMs. RSA is invariant to orthogonal transformations and reflects how models structure their representational spaces.

Soft Matching (Khosla & Williams, 2024). Soft Matching (SoftMatch) generalizes permutation distance (Ding et al., 2021) to representations with different numbers of units by relaxing permutations to "soft permutations." Specifically, consider a non-negative matrix $\mathbf{P} \in \mathbb{R}^{N_i \times N_j}$ whose rows each sum to $1/N_i$ and columns to $1/N_j$. The set of such matrices defines a transportation polytope (De Loera & Kim, 2013), $\mathcal{T}(N_i, N_j)$. The optimization problem is

$$d_T(\mathbf{X}_i, \mathbf{X}_j) = \min_{\mathbf{P} \in \mathcal{T}(N_i, N_j)} \sum_{k,l} \mathbf{P}_{kl} \, \|x_i^{(k)} - x_j^{(l)}\|^2,$$

where $x_i^{(k)}$ and $x_j^{(l)}$ are the k-th and l-th columns (units) of $\mathbf{X}_i$ and $\mathbf{X}_j$. The optimal transport plan $\mathbf{P}^\star$ is found via the network simplex algorithm. When $N_i = N_j$, this reduces to an optimal permutation. The final similarity score is the mean unit-wise correlation between $\mathbf{X}_j$ and $\mathbf{X}_i \mathbf{P}^\star$.

Procrustes Alignment (Ding et al., 2021). Procrustes analysis finds the orthogonal transformation that best aligns two representations while preserving geometry. For unequal dimensions, the smaller representation is zero-padded. The optimization problem is $\min_{\mathbf{R} \in \mathcal{O}(N)} \|\mathbf{X}_j - \mathbf{X}_i \mathbf{R}\|_2^2$, where $\mathcal{O}(N) = \{\mathbf{R} \in \mathbb{R}^{N \times N} : \mathbf{R}^\top \mathbf{R} = \mathbf{I}\}$. The optimal transformation $\mathbf{R}^\star$ is obtained via singular value decomposition. The similarity score is the mean unit-wise correlation between $\mathbf{X}_j$ and $\mathbf{X}_i \mathbf{R}^\star$.

Linear Predictivity (Yamins et al., 2014). Linear predictivity seeks an unconstrained linear transformation that best predicts one representation from another: $\min_{\mathbf{L}} \|\mathbf{X}_j - \mathbf{X}_i \mathbf{L}\|_2^2$. The optimal mapping $\mathbf{L}^\star$ is estimated via ordinary least squares. The final similarity score is the mean unit-wise correlation between $\mathbf{X}_j$ and $\mathbf{X}_i \mathbf{L}^\star$.

Average Baseline. To provide a baseline that naively uses all metrics' information, we symmetrized and min-max rescaled all metrics' matrices and simply averaged them.

2.3 SEPARABILITY MEASURES

We assess how well each representational similarity metric recovers specified group structure by comparing similarities within a model family or anatomical group to similarities across groups. Labels define the evaluation target, not a claim of representational identity. For groups A and B, let $\mathcal{W}_A$ be the set of similarities between distinct members of A and $\mathcal{X}_{AB}$ the set of similarities between members of A and members of B, with means μ and variances σ^2. The contrastive ratio measures the relative gap between the two means, and d' (Bo et al., 2024) scales the gap by their spread: $CR_{A|B} = (\mu(\mathcal{W}_A) - \mu(\mathcal{X}_{AB}))/(\mu(\mathcal{W}_A) + \mu(\mathcal{X}_{AB}))$ and $d'_{A|B} = (\mu(\mathcal{W}_A) - \mu(\mathcal{X}_{AB}))/\sqrt{(\sigma^2(\mathcal{W}_A) + \sigma^2(\mathcal{X}_{AB}))/2}$. The silhouette score (Rousseeuw, 1987) works per item: $s_{A|B} = \frac{1}{|A|}\sum_{i \in A}(a_i - b_i)/\max(a_i, b_i)$, where a_i and b_i are the mean similarities of item i to the other members of A and to the members of B. Each score for a pair $\{A, B\}$ averages the two directions $A|B$ and $B|A$. Heatmaps show these pair scores, and bar plots average them over all unique pairs of groups. Higher values indicate stronger recovery of known procedural structure in models or anatomical–functional structure in neural data.

2.4 INTEGRATION BASELINES

We include arithmetic averaging and the existing SNF algorithm (Wang et al., 2014; Markello, 2020) as secondary summaries of multiple metric-specific similarity matrices. SNF iteratively combines local affinity graphs without using group labels in the fusion updates. We use $K = 5$ neighbors for the model population and NSD and $K = 3$ for ABO, whose smallest group has three regions, with $\mu = 0.5$ and $T = 20$ iterations. The construction, update equations, and controls are given in Appendix O. Fusion scores are evaluated against the same criteria as individual metrics, not treated as ground truth. For typologies, we cluster the fused matrix hierarchically with SciPy's linkage function (Virtanen et al., 2020).

3 RESULTS

3.1 WHICH SIMILARITY METRICS RECOVER PROCEDURAL DIFFERENCES BETWEEN MODELS?

To evaluate which dimensions of representational correspondence recover meaningful structure among models, we systematically compared a suite of representational similarity measures using the procedural criteria introduced above. Our goal is to assess how effectively each metric recovers known procedural structure, distinguishing models that differ in architecture or training paradigm while grouping procedurally matched models.

We quantify recovery of procedural structure using the separability measures described in Section 2, which contrast within-family and across-family similarities. Results reported in this section are based on ImageNet representations; the same 35 models are also evaluated on Ecoset, CIFAR10, and CIFAR100 (Appendix G). Across metrics, we observe substantial variation in the degree to which procedural structure is recovered (Figures 1, G.1, G.2, G.3), indicating that different representational facets carry markedly different information about model families.

Metrics that preserve representational geometry or unit-level tuning properties demonstrate the strongest ability to discriminate between model families. As in Figure 1, RSA, which captures representational geometry, achieves the highest separability with a d' of 3.95 and silhouette coefficient of 0.56, indicating that geometric organization or relational structure, i.e. how models arrange points in representation space, constitutes a strong family-specific signature. Linear CKA also shows strong discrimination (d' = 3.91), which aligns with recent theoretical work showing that centered RSA and linear CKA are mathematically equivalent when appropriate centering is applied (Williams, 2024). Similarly, SoftMatch, which preserves individual unit tuning while mapping two representations, shows robust discrimination (d' = 3.59, silhouette = 0.30). Even supervised and self-supervised variants within the same architecture family (particularly CNNs) are reliably separated by these metrics, demonstrating that the training paradigm fundamentally shapes the geometry and tunings of individual neurons, such that metrics diagnostic of these representational facets achieve good separation. These properties thus constitute the unique representational "fingerprints" of model families. Interestingly, Procrustes alignment, which allows orthogonal transformations, shows intermediate discrimination (d' = 2.96), falling between SoftMatch and Linear Predictivity. This

reveals a clear pattern among mapping-based metrics: discriminability decreases monotonically as the transformations become more flexible (SoftMatch $>$ Procrustes $>$ Linear Predictivity). The constraints imposed by less flexible mappings appear to preserve family-specific signatures.

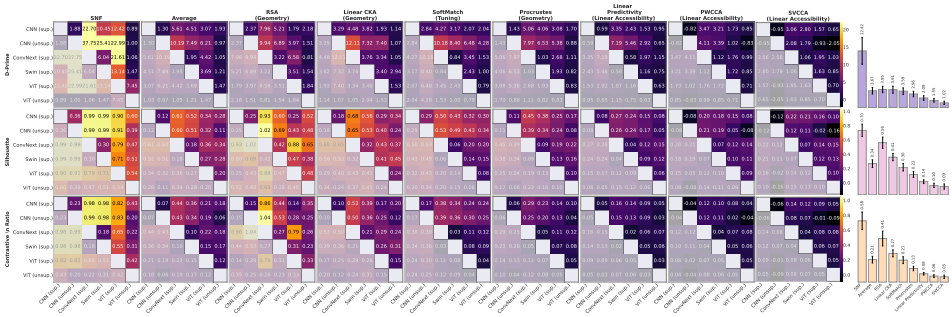

Figure 1: Model-family separability on ImageNet under d', silhouette score and contrastive ratio. Columns correspond to seven commonly used representational metrics with their emphasized aspects and two integration baselines (SNF, average). The barplot on the right side of each row shows the mean model-family separability with standard error. Scores are shown in their native scales and are not directly comparable across measures.

In contrast, metrics capturing linearly accessible information show substantially weaker discrimination between families. Linear Predictivity demonstrates the lowest separability among direct mapping-based metrics (d' = 2.09, silhouette = 0.14), while CCA-based metrics (PWCCA: d' = 1.55; SVCCA: d' = 1.02) show even weaker family separation. The weak discrimination of CCA-based metrics is particularly revealing. CCA identifies maximally correlated linear projections between representations, finding shared subspaces that are invariant to invertible linear transformations. CCA loads on the linear-accessibility facet: it detects shared linearly decodable subspaces but, unlike RSA/CKA or Procrustes/SoftMatch, it does not constrain or preserve representational geometry or tuning. The invariance of CCA to linear transformations, which makes it powerful for finding shared structure across superficially dissimilar representations, also makes it insensitive to the geometric and topological features that distinguish families.

The theoretical relationships among these metrics help explain the discrimination hierarchy. RSA and Linear CKA are mathematically equivalent under appropriate centering (Williams, 2024) and both preserve the geometric structure of representations. They compare how similarly models organize their representation spaces without fitting any transformation. In contrast, the mapping-based metrics show decreasing discrimination as they allow increasingly flexible transformations: SoftMatch permits only permutations that preserve individual unit correspondences; Procrustes allows orthogonal transformations, while CCA searches for optimal linear projections that maximize correlation. Linear Predictivity provides the most flexibility, allowing any linear transform that minimizes prediction error. The weak discrimination of metrics that assess linearly accessible information might suggest that this aspect of representation is more consistent across model families than geometric organization. Whether this reflects convergent computational strategies, methodological limitations of these metrics, or task-imposed constraints remains open.

The advantage of geometry- and tuning-sensitive metrics recurs on the other image benchmarks, but the leading metric changes. Linear CKA has the highest single-metric d' on Ecoset (3.76), whereas SoftMatch leads on CIFAR10 (3.22) and CIFAR100 (3.43); RSA keeps the highest single-metric silhouette score and contrastive ratio on all four benchmarks. Linear Predictivity and the CCA variants remain the weakest on every benchmark.

3.2 Do these comparisons hold across random seeds and network depth?

Family separation is only informative if differences between procedures exceed the variation produced by training alone. We therefore use two seeds of each of seven configurations (supervised AlexNet, VGG16, VGG19, ResNet18, ResNet50, and MoCo ResNet18 and ResNet50), grouped into supervised AlexNet, VGG, and ResNet, and self-supervised ResNet (Figure 2; Appendix H). Within-group similarity now includes seed-to-seed variation.

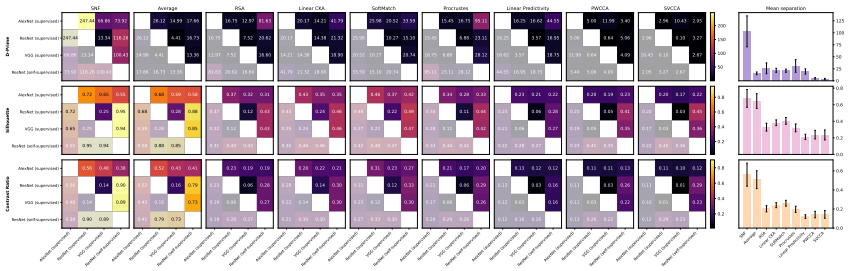

Figure 2: Group separation among independently trained seeds on ImageNet. The 14 instances (two seeds of seven configurations) form four groups. Layout as in Figure 1; bars summarize the six group pairs with standard errors across pairs.

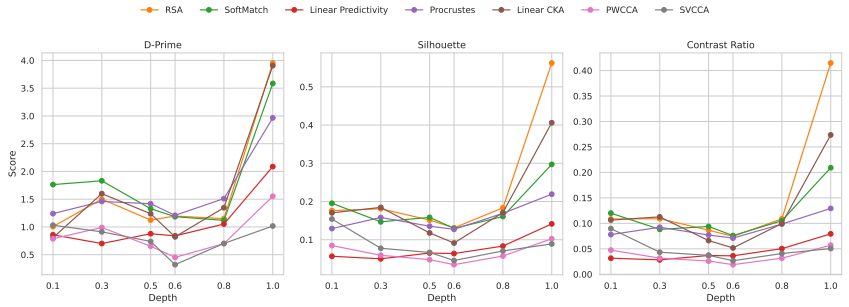

Figure 3: Model-family separation across normalized network depth on ImageNet. Each panel compares individual metrics and integration baselines under one separation criterion; depth 1.0 is the final extracted representation. Other benchmarks are in Appendix I.

On ImageNet, the ordering broadly follows the 35-model benchmark: SoftMatch has the highest single-metric silhouette score (0.40) and contrastive ratio (0.26), followed by Linear CKA (0.38, 0.24), while Procrustes (30.90) and RSA (26.01) have the highest d'. Linear Predictivity has the lowest silhouette score (0.21) and contrastive ratio (0.12), and the CCA variants the lowest d' (PWCCA 5.03, SVCCA 3.58). Under the silhouette score, supervised ResNets and supervised VGGs, which differ only in architecture, form the hardest pair for every metric (e.g., 0.12 for RSA and 0.06 for Linear Predictivity), whereas supervised and self-supervised ResNets, which share an architecture but differ in objective, are separated more clearly (0.43 and 0.28). On the other three benchmarks, SoftMatch again leads silhouette score and contrastive ratio, and Procrustes (Ecoset) or SoftMatch (CIFAR10, CIFAR100) leads d'. Within-group variance is low here, so these d' values are not comparable in magnitude with the 35-model benchmark.

Network depth introduces another limit on penultimate-layer conclusions. We evaluate family separation at normalized depths $d \in \{0.1, 0.3, 0.5, 0.6, 0.8, 1.0\}$ (Figure 3; Appendix I). On ImageNet, SoftMatch leads single-metric d' at depths 0.1 and 0.3, Procrustes leads at 0.5, 0.6, and 0.8, and RSA leads at 1.0. RSA rises from $d' = 1.15$ at depth 0.8 to 3.95 at the final representation. Separation is not monotonic with depth, and a metric's final-layer ranking should not be assumed to hold throughout a network.

3.3 Which similarity metrics recover known anatomical–functional structure in the brain?

We next ask whether the patterns observed in models generalize to biological systems. In NSD, we compare cortical regions across subjects using responses to a shared set of natural images. Comparisons across different regions within the same subject are excluded due to the spatial blurring inherent to fMRI data, which can artificially inflate within-subject similarity across regions. We exclude SVCCA and PWCCA from the NSD analyses, as their alignment values were systematically inflated by the large voxel dimensionality relative to the number of stimuli.

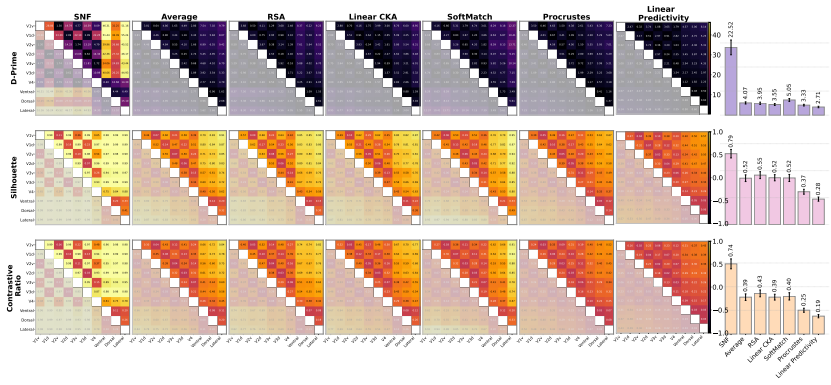

Figure 4: Similar to Figure 1 but on brain responses (NSD) and excluding SVCCA and PWCCA. Bars summarize 45 unique region pairs.

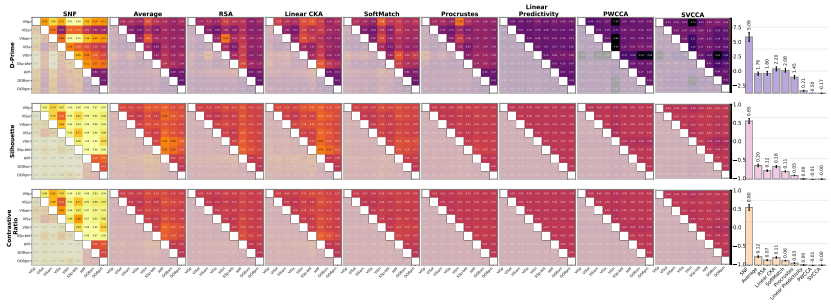

Figure 5: Anatomical-group separation in ABO. The nine groups are VISp, VISal, VISam, VISa, VISrl, SSp-bfd, hippocampal formation (HPF), sensory–motor thalamus (DORsm), and polymodal association thalamus (DORpm). Bars summarize 36 unique group pairs with standard errors across pairs. Unlike NSD, this analysis includes PWCCA and SVCCA.

Consistent with the model results, geometry- and tuning-sensitive metrics differentiate cortical regions better than more permissive mappings. SoftMatch has the highest single-metric mean d' (5.05), followed by RSA (3.95), Linear CKA (3.55), Procrustes (3.33), and Linear Predictivity (2.71; Figure 4). RSA instead leads single-metric silhouette score (0.55) and contrastive ratio (0.43).

ABO extends the comparison to mouse Neuropixels responses during natural-movie viewing (de Vries et al., 2020; Siegle et al., 2021). Following the CCFv3 anatomical ancestry (Wang et al., 2020), the analysis groups 41 region or layer subdivisions into nine anatomical groups spanning visual cortex, hippocampal formation, and thalamus (Appendix J). Unlike the NSD cross-subject test, it asks whether subdivisions within the same anatomical group respond more similarly than subdivisions in different groups. Linear CKA has the strongest single-metric separation under all three criteria ($d' = 2.20$, silhouette $= 0.18$, contrastive ratio $= 0.11$; Figure 5). SoftMatch follows on d' (2.08), ahead of RSA (1.80), Procrustes (1.45), and Linear Predictivity (0.21), while PWCCA and SVCCA have negative mean d'. The advantage of geometry- and tuning-sensitive metrics therefore extends from human fMRI to mouse electrophysiology, although the leading metric again depends on the dataset and criterion.

3.4 Integrating metrics as a summary of shared structure

Because each metric emphasizes a different facet, we also examine whether combining them recovers structure that is consistently supported across metrics. SNF, which integrates the similarity graphs of all metrics, attains the highest mean d' in the main model benchmark on every image set (12.42 on ImageNet; 10.51, 5.28, and 5.24 on Ecoset, CIFAR10, and CIFAR100), as well as in NSD (22.52) and ABO (5.09). As shown in Figure 1, it maintains high discrimination across nearly all family pairs, whereas individual metrics often separate some families while failing for others. The advantage is

not uniform, however. At depth 0.1 SoftMatch and averaging exceed it (Figure 3), and in the seeded population on CIFAR10 and CIFAR100, simple averaging exceeds SNF on silhouette score (0.52 and 0.54 versus 0.46 and 0.47) and contrastive ratio (Appendix H).

To test whether SNF replicates any single metric, we correlate the vectorized off-diagonal entries of the symmetrized similarity matrices (Figure K.1). Geometry-preserving metrics (RSA, SoftMatch, CKA) agree strongly, mapping-based metrics are moderately aligned, and the CCA variants agree mainly with each other; SNF aligns only moderately with any single metric and is distinct from averaging. Leave-one-out fusion shows that its separation does not hinge on any one input: removing a single metric keeps the ImageNet d' between 11.70 and 13.20 (Appendix P). Permutation controls further indicate that SNF performance depends on cross-metric signal rather than graph sharpening alone (Figure R.1). We therefore treat the fused matrix as a summary of relations shared across metrics, not as a more accurate metric; individual metrics remain necessary to identify which representational properties drive a relation.

3.5 Data-driven typologies in models and brains

We next ask what population-level organization this shared structure reveals, clustering the fused similarity matrices of artificial models and biological brain regions without using architecture, training paradigm, or anatomy labels.

3.5.1 Population structure in vision models

As a baseline, clustering based on individual similarity metrics produces fragmented and inconsistent organization (Figure L.1, L.2, L.3, L.4, L.6, L.7, L.9, L.10). Metrics such as Linear Predictivity and Procrustes yield relatively uniform similarity values across models, resulting in diffuse clusters with weak separation. SoftMatch, which emphasizes unit-level geometric alignment, struggles to clearly organize models that fall outside standard CNN or supervised regimes. CCA-based metrics (PWCCA, SVCCA) produce noisy similarity matrices in which no stable clustering structure is apparent. RSA reveals partial organization (e.g. a coarse separation between CNNs and Transformers) but the resulting clusters remain diffuse. Each metric emphasizes a particular representational facet, so the population structure it reveals is partial.

Hierarchical clustering of the fused similarity matrix instead yields well-defined groupings that both recover expected relationships and reveal additional organizational structure (Figure 6, L.5, L.8, L.11). The cophenetic correlation of the SNF dendrogram (0.98 on ImageNet) is comparable to the best single metric (SoftMatch, 0.99), indicating that the clustering faithfully reflects the fused pairwise similarities (Figure M.1). Examining these clusters reveals a nuanced interplay between architecture and training paradigm. While some groupings align with architectural expectations, such as supervised CNNs and supervised ViTs forming distinct clusters (Figure 6), training paradigms can override architectural boundaries. Self-supervised ResNets group with the MoCo ViT rather than with their supervised architectural counterparts, suggesting that self-supervised objectives may induce shared computational strategies that dominate over architectural differences. Hybrid architectures (ConvNeXt and Swin) cluster with MAE models despite their different design philosophies. ConvNeXts modernize CNNs with Transformer-inspired components, Swins introduce CNN-like biases into Transformers, and MAE employs masked reconstruction, yet they converge on similar representational structures. These groupings describe the chosen metrics and model population; they are hypotheses about shared computational strategies rather than definitive categories.

3.5.2 Population structure in visual cortex

The same analysis of visual cortical regions reveals cross-subject organization consistent with known anatomy. In the PCA layouts, individual metrics recover broad trends, while the fused matrix gives the most compact layout, with early visual areas (V1–V4) forming a smooth progression and V4 occupying a transitional position between early and higher-level visual regions (Figure 7).

To quantify how well each measure preserves the expected ventral-stream hierarchy, following Thobani et al. (2025), we compute Kendall's τ between representational dissimilarity and the hierarchical level difference of five ventral-stream regions (Appendix J.1). Linear CKA and RSA align most strongly with the hierarchy ($\tau = 0.64$ and 0.62), and among mapping-based metrics

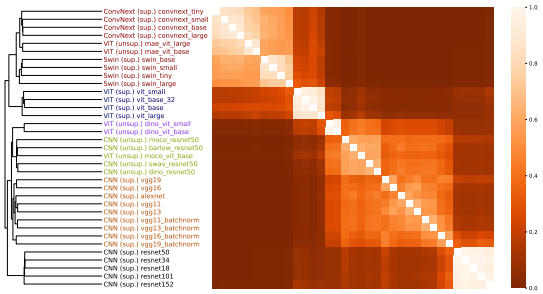

Figure 6: SNF-based clustering groups models by architecture and supervision regime. The heatmap shows the SNF-fused similarity matrix reordered by leaf ordering. Rows and columns follow the same ordering. Leaf labels are colored by the SNF-derived cluster they belong to.

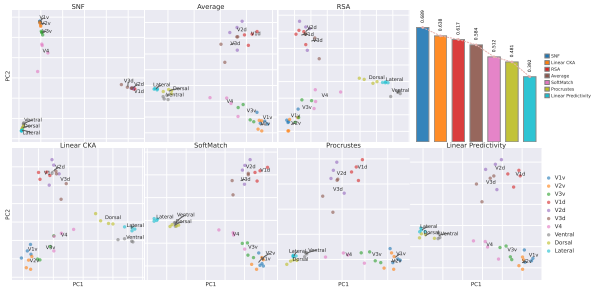

Figure 7: PCA visualization and hierarchical score comparison of cross-brain-region representational relationships in NSD. Each panel corresponds to one metric, points represent brain region instances, and text labels indicate centroid positions. The top-right bar chart displays the raw hierarchical scores for each representational similarity metric.

alignment decreases as the mapping becomes more flexible (SoftMatch 0.51, Procrustes 0.48, Linear Predictivity 0.39). The fused matrix reaches $\tau = 0.69$ (Figure 7). This evaluates a coarse ordering of five regions, not detailed circuit or laminar organization.

4 Discussion

This work reframes representational comparison as a question of *which relationships a metric should preserve for a given scientific purpose*. Across four image benchmarks, NSD, and ABO, geometry- and tuning-sensitive metrics generally separate model families or anatomical groups more strongly than flexible mappings, but the leading metric depends on stimulus set, recording modality, network depth, and separation criterion.

Separation is useful when the question concerns distinctions associated with architecture, training, or anatomy, but it is not a universal measure of representational quality. The seed experiment shows why the criterion must control for training variability: differences in objective are separated beyond seed-to-seed variation, whereas architectural differences among supervised CNNs are much harder to separate. Flexible mappings may correctly identify shared linearly accessible information that geometric comparisons treat as different; behavioral or task-specific tests are needed to decide which notion is appropriate.

SNF complements this comparison rather than replacing it. It summarizes neighborhood structure shared across metrics, improves most separation scores, and yields typologies in which training paradigm can cut across architecture. Its outputs, however, depend on the metric set, population, and fusion parameters. The main result is therefore not a universally preferred measure, but evidence for matching the evaluation criterion to the correspondence of interest.

AI USE STATEMENT

LLMs are mainly used in two ways. For aiding or polishing writing, they are primarily used to identify typos and make the language more aligned with conventions of academic writing. For retrieval and discovery, LLMs with internet access are used to search for related work.

ETHICS STATEMENT

This work uses existing public data, pretrained models, and software. The brain data comprise fMRI responses from NSD (Allen et al., 2022), used under the NSD Data Access Agreement, and existing mouse Neuropixels recordings from ABO (Siegle et al., 2021). We collected no new human or animal data. The licenses or terms of use of all assets are listed in Appendix D and were properly respected. As discussed in Appendix F, the work is foundational, and we therefore do not identify a direct path to societal harm.

REPRODUCIBILITY STATEMENT

The evaluation is described in the main text and appendix: Section 2.2 defines the representational similarity measures together with the mean-centering and 5-fold cross-validation protocol used to score them, Section 2.3 defines the separability measures, and Section O describes the fusion procedure and fixes its hyperparameters. Appendix A gives the dataset splits, the sources of the pretrained weights, the feature extraction convention, and the compute environment, and Appendix B lists the models in each family. The seed analysis additionally uses paired ImageNet-1k checkpoints trained with the same architecture and recipe but different random seeds. The analysis code will be made publicly available upon publication.

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A Massive 7T fMRI Dataset to Bridge Cognitive Neuroscience and Artificial Intelligence. *Nature neuroscience*, 25(1):116–126, January 2022. ISSN 1097-6256. doi: 10.1038/s41593-021-00962-x.

Yiqing Bo, Ansh Soni, Sudhanshu Srivastava, and Meenakshi Khosla. Evaluating representational similarity measures from the lens of functional correspondence. *arXiv:2411.14633 [q-bio.NC]*, 2024. doi: 10.48550/arXiv.2411.14633. URL https://arxiv.org/abs/2411.14633.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Nathan Cloos, Moufan Li, Markus Siegel, Scott L Brincat, Earl K Miller, Guangyu Robert Yang, and Christopher J Cueva. Differentiable optimization of similarity scores between models and brains. *arXiv preprint arXiv:2407.07059*, 2024.

Jesús A De Loera and Edward D Kim. Combinatorics and geometry of transportation polytopes: An update. *Discrete geometry and algebraic combinatorics*, 625:37–76, 2013.

Saskia E. J. de Vries, Jerome A. Lecoq, Michael A. Buice, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23:138–151, 2020. doi: 10.1038/s41593-019-0550-9.

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

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

Meenakshi Khosla and Alex H Williams. Soft Matching Distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pp. 326–341. PMLR, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, Volume 2 - 2008, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s, 2022. URL https://arxiv.org/abs/2201.03545.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016.

Ross Markello. snfpy: Similarity Network Fusion in Python, 2020. URL https://github.com/rmarkello/snfpy.

Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Joshua H. Siegle, Xiaoxuan Jia, Séverine Durand, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592:86–92, 2021. doi: 10.1038/s41586-020-03171-x.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Ansh Soni, Sudhanshu Srivastava, Konrad Kording, and Meenakshi Khosla. Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, 2024. doi: 10.1101/2024.08.07.607035. URL https://www.biorxiv.org/content/early/2024/08/09/2024.08.07.607035.

Imran Thobani, Javier Sagastuy-Brena, Aran Nayebi, Jacob Prince, Rosa Cao, and Daniel Yamins. Model-brain comparison using inter-animal transforms, 2025. URL https://arxiv.org/abs/2510.02523.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3):261–272, 2020.

Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.

Quanxin Wang, Song-Lin Ding, Yang Li, et al. The allen mouse brain common coordinate framework: A 3d reference atlas. *Cell*, 181:936–953.e20, 2020. doi: 10.1016/j.cell.2020.04.007.

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

Alex Williams, Sarah Harvey, T. Toos, Heiko Schütt, and Nikolaus Kriegeskorte. Quantifying similarity between neural population codes. Keynote and tutorial, Cognitive Computational Neuroscience (CCN), 2024.

Alex H Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. *bioRxiv*, pp. 2024–10, 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6/`.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction, 2021. URL `https://arxiv.org/abs/2103.03230`.

A Experiment Settings

Datasets. The chosen datasets are balanced across different classes as shown in Table A.1.

Table A.1: Per-class and total sample counts for standard evaluation splits.

Dataset	Number of Classes	Samples / Class	Total Samples
ImageNet-1k (valid)	1,000	50	50,000
Ecoset (valid)	565	50	28,250
CIFAR-10 (test)	10	1,000	10,000
CIFAR-100 (test)	100	100	10,000

Models. All models are trained on the ImageNet-1K training set. We obtain pretrained weights from torchvision (maintainers & contributors, 2016), Torch Hub (Paszke et al., 2019), timm (Wightman, 2019), or the official repositories. Unless noted otherwise, we extract activations from each model's penultimate layer. For CNNs, which commonly include global average pooling, we use that pooled feature. For ViT-style models, we average non-CLS token embeddings to form the final representation for consistency across architectures.

Compute Resources. Experiments were run on an internal server with 48 AMD EPYC 9224 CPU cores and 768 GB RAM. Two NVIDIA A40 GPUs with 48 GB of memory and 4 CPU data-loading workers are used in experiments. CPU thread usage outside data loading was not explicitly constrained in the code. Cached pretrained model checkpoints, embeddings, similarity matrices, and RSA RDMs used approximately 1.4 TB of storage.

B Model Family and Architecture Choices

We evaluate multiple architectures within each family to capture variation in depth, width, parameter amounts and design choices.

Convolutional Neural Network (supervised; CNN (sup.)). Bottom-up hierarchies with convolutions and pooling that impose strong local inductive biases. We include AlexNet (Krizhevsky et al., 2012), VGG-11/13/16/19 (with/without batch normalization) (Simonyan & Zisserman, 2015), and ResNet-18/34/50/101/152 (He et al., 2016).

Transformer (supervised; Trans (sup.)). Vision Transformers partition images into fixed-size patches and use multi-head self-attention for global interactions (Dosovitskiy et al., 2021). We include ViT-S/16, ViT-B/16, ViT-L/16, and ViT-B/32.

ConvNeXt (Liu et al., 2022). A convolutional family inspired by Transformer design (e.g., large-kernel depthwise convolutions, patchified stems, inverted bottlenecks). We use ConvNeXt-Tiny/Small/Base/Large.

Swin Transformer (Liu et al., 2021). A hierarchical Transformer with shifted window attention for efficient locality while retaining global context. We use Swin-Tiny/Small/Base/Large.

Convolutional Neural Network (self-supervised; CNN (selfsup.)). Methods trained without labels using CNN backbones (ResNet-50). We include MoCo (Chen* et al., 2021), DINO (Caron et al., 2021), SwAV (Caron et al., 2020), and Barlow Twins (Zbontar et al., 2021), spanning contrastive and non-contrastive paradigms (momentum contrast, self-distillation, online clustering, and redundancy reduction).

Transformer (self-supervised; Trans (selfsup.)). Label-free training with Transformer backbones. We include DINO-ViT-Small/16 and DINO-ViT-Base/16 (Caron et al., 2021), MoCo-ViT-Base/16 (Chen* et al., 2021), and MAE-ViT-Base/16 and MAE-ViT-Large/16 (He et al., 2021).

C Representational Facet of Different Metrics

Table C.1: Different representational similarity metrics emphasize different representational facets. A high similarity between two models under a certain metric means they are similar in that facet.

Metric	Representational Facet
RSA	Focuses on whether two representations preserve the same relational geometry across stimuli, i.e., the pattern of pairwise dissimilarities in the chosen representational dissimilarity matrixes (Kriegeskorte et al., 2008).
Linear CKA	Focuses on the preservation of global inter-example similarity structure, equivalently the centered kernel structure. It is invariant to orthogonal transformations and isotropic scaling but remains sensitive to more general linear reparameterizations such as anisotropic scaling or shear (Kornblith et al., 2019).
SoftMatch	Focuses on unit-level correspondence: high similarity means that individual units can be matched across networks and generalized to soft assignments when layer widths differ. It is invariant to permutation or relabeling of units but sensitive to rotations or arbitrary linear mixing that destroys neuron-wise correspondence (Khosla & Williams, 2024).
Procrustes	Focuses on whether one representation can be aligned to the other by a single orthogonal transformation after centering. It is invariant to rotation/reflection, and centering removes global translation, but it remains sensitive to anisotropic scaling, shear, and more general linear reparameterizations. Thus, it also preserves rigid geometric alignment rather than arbitrary linear equivalence (Ding et al., 2021).
Linear Predictivity	Focuses on shared linearly accessible information: high similarity means that one representation can linearly predict the other with low residual error. It does not require neuron-wise correspondence or preservation of geometry and is therefore comparatively tolerant to flexible linear reparameterizations of the predictor representation (Yamins et al., 2014).
PWCCA	Focuses on shared linear subspaces with canonical directions weighted by how strongly they explain the original representation. Like CCA more generally, it is comparatively insensitive to invertible linear reparameterizations of the compared representations and therefore emphasizes shared decodable structure rather than exact geometry or neuron identity. The projection weighting further makes strong, representation-relevant directions count more than weak ones (Morcos et al., 2018).
SVCCA	Focuses on shared linearly accessible structure after first restricting each representation to its dominant singular-vector subspace. It inherits CCA-style insensitivity to linear reparameterizations within the retained subspaces but additionally assumes that low-variance directions are less important and can be discarded before comparison. Thus, it emphasizes dominant shared subspace structure rather than full-space geometry or exact unit alignment (Raghu et al., 2017).

D Existing Assets and Licenses

This work uses existing datasets, pretrained model weights, and software packages. We use all assets for non-commercial academic research, credit the original creators, and do not redistribute third-party datasets or pretrained weights. Table D.1 summarizes the sources and license or usage terms of the main assets used in this work. Unless otherwise noted, the pretrained weights used in this work are pre-trained on ImageNet-1k and are subject to ImageNet usage terms; to avoid repetition, these terms are listed once under the data category and are not repeated for each pretrained-weight entry.

Table D.1: Main existing assets used in this work.

Category	Item	License
Data	ImageNet (Deng et al., 2009)	ImageNet terms of access, non-commercial research and educational use
	NSD (Allen et al., 2022)	NSD Data Access Agreement
	ABO (Siegle et al., 2021)	Allen Institute terms of use
	EcoSet (Mehrer et al., 2021)	CC BY-NC-SA 2.0
	CIFAR-10 (Krizhevsky, 2009)	License not specified, citation requested
	CIFAR-100 (Krizhevsky, 2009)	License not specified, citation requested
Pretrained Weights	AlexNet (Krizhevsky et al., 2012)	torchvision, BSD-3-Clause
	VGG-11/13/16/19, with and without batch normalization (Simonyan & Zisserman, 2015)	torchvision, BSD-3-Clause
	ResNet-18/34/50/101/152 (He et al., 2016)	torchvision, BSD-3-Clause
	ViT-S/16, ViT-B/16, and ViT-B/32 (Dosovitskiy et al., 2021)	timm, Apache-2.0
	ViT-L/16 (Dosovitskiy et al., 2021)	torchvision, BSD-3-Clause
	ConvNeXt-Tiny/Small/Base/Large (Liu et al., 2022)	timm, Apache-2.0; Facebook Research, MIT
	Swin-Tiny/Small/Base/Large (Liu et al., 2021)	Microsoft, MIT
	DINO ResNet-50 and DINO ViT-Small/Base-16 (Caron et al., 2021)	Facebook Research, Apache-2.0
	MoCo ResNet-50 and MoCo ViT-Base-16 (Chen* et al., 2021)	Facebook Research, CC BY-NC 4.0
	SwAV ResNet-50 (Caron et al., 2020)	Facebook Research, CC BY-NC 4.0
	Barlow Twins ResNet-50 (Zbontar et al., 2021)	Facebook Research, MIT
	MAE ViT-Base-16 and ViT-Large-16 (He et al., 2021)	Facebook Research, CC BY-NC 4.0
Libraries	torchvision 0.25.0 (maintainers & contributors, 2016)	BSD-3-Clause
	timm 1.0.20 (Wightman, 2019)	Apache-2.0
	Transformers 4.51.3 (Wolf et al., 2020)	Apache-2.0
Other Software	snfpy 0.2.2 (Markello, 2020)	LGPL-3.0; GPL-3.0

E LIMITATIONS

Our analysis has several important limitations. First, we focus primarily on natural image databases (ImageNet, Ecoset, CIFAR) and testing the stability of the discovered typology to more out-of-distribution domains would test the generalizability of these groupings. Second, the depth analysis samples six normalized positions rather than every layer, and the seed experiment covers only seven CNN configurations with two seeds each. Neither establishes robustness across all architectures or training runs. Third, a limitation is that all models used in this study were trained on ImageNet-1k. Exploring how the typology changes for web-scale SSL or image–text models would be an important direction for future work. Finally, procedural and anatomical groups are hypothesis-driven evaluation targets, not definitive labels of representational or behavioral equivalence. NSD and ABO differ in species, stimulus format, recording modality, and grouping scheme, so differences in rankings cannot be attributed to modality alone. Fusion-based typologies also depend on the chosen input metrics and population.

F BROADER IMPACTS

The societal impact of this work is indirect, since it is foundational rather than a deployed AI system. Its main positive impact is to improve the reliability of scientific claims about model similarity, brain-model alignment, and representation learning. The work may also support downstream interpretability research by providing a clearer way to compare internal representations across neural systems.

G SEPARATION PERFORMANCE ON OTHER DATASETS

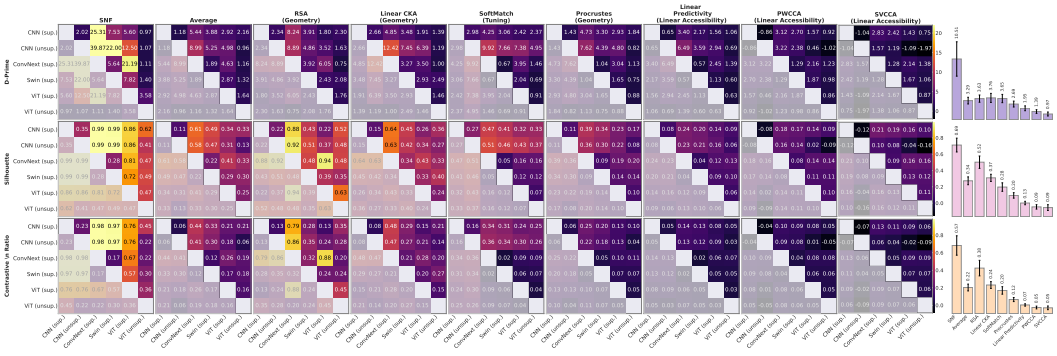

Figure G.1: Same as Figure 1, but using Ecoset instead of ImageNet.

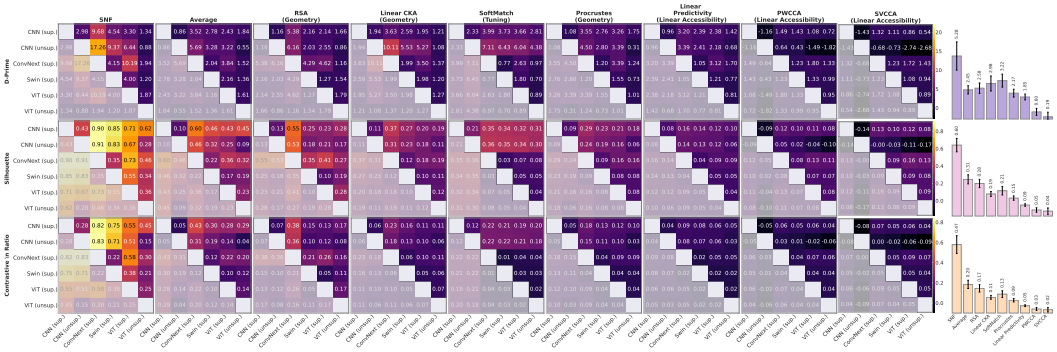

Figure G.2: Same as Figure 1, but using CIFAR10 instead of ImageNet.

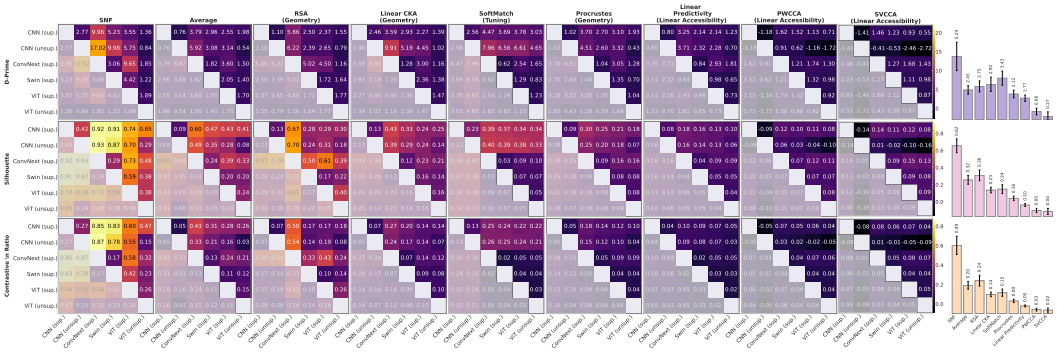

Figure G.3: Same as Figure 1, but using CIFAR100 instead of ImageNet.

H RANDOM-SEED SEPARATION

The seeded population contains two instances of each of seven configurations. Separation uses four groups: supervised AlexNet, supervised ResNet, supervised VGG, and self-supervised ResNet. The six group pairs differ from the 35-model benchmark; a large d' in this small population can reflect a small within-group variance rather than a proportional gain in scientific validity.

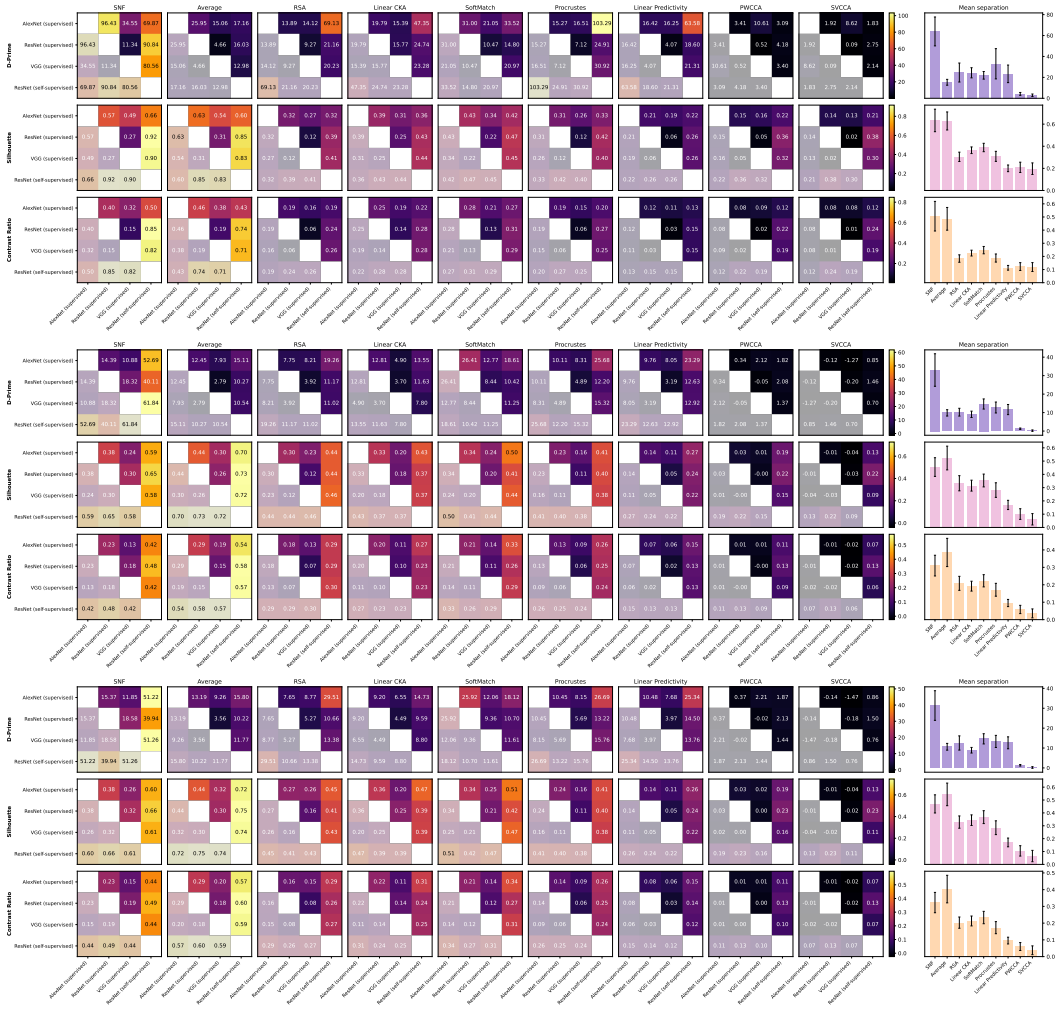

Figure H.1: Seeded-model separation on Ecoset, CIFAR10, and CIFAR100 (top to bottom); ImageNet is shown in Figure 2. Each panel reports d', silhouette score, and contrastive ratio; bars summarize six group pairs with standard errors across pairs.

I SEPARATION ACROSS NETWORK DEPTH

For CNNs, layers are counted using batch normalization units when present and ReLU units otherwise. For ViTs, one feature-extraction unit consists of the first layer normalization, attention block, and second layer normalization; features are taken after the second normalization. For L units, the layer at normalized depth d is selected as $\ell = \max(1, \lfloor dL \rfloor)$. We evaluate $d \in \{0.1, 0.3, 0.5, 0.6, 0.8, 1.0\}$, with 1.0 denoting the final extracted representation. These curves evaluate family separation at each depth, rather than correlations between similarity matrices from different depths.

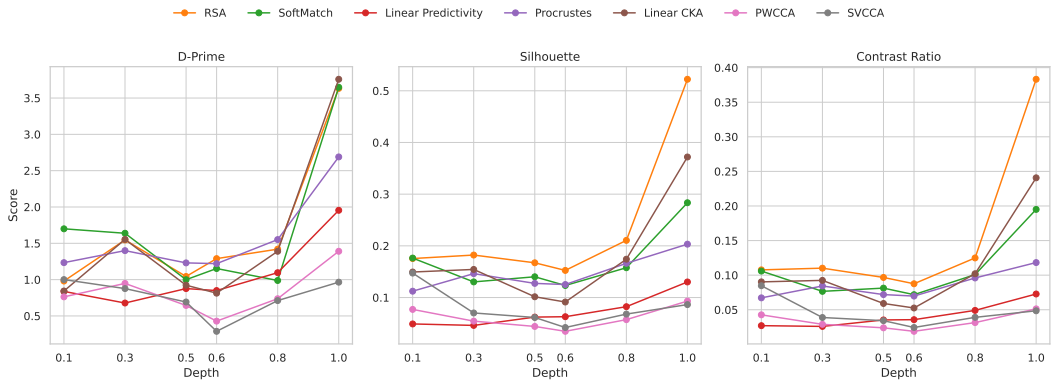

Figure I.1: Model-family separation across depth on Ecoset, using the criteria in Figure 3.

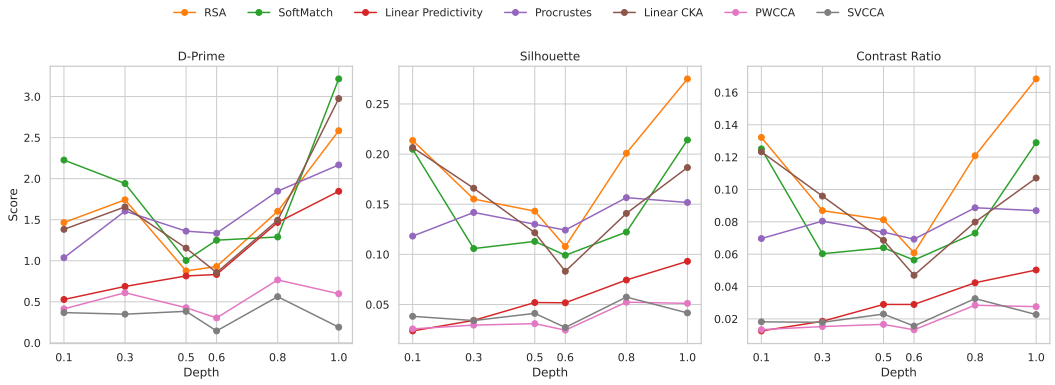

Figure I.2: Model-family separation across depth on CIFAR10.

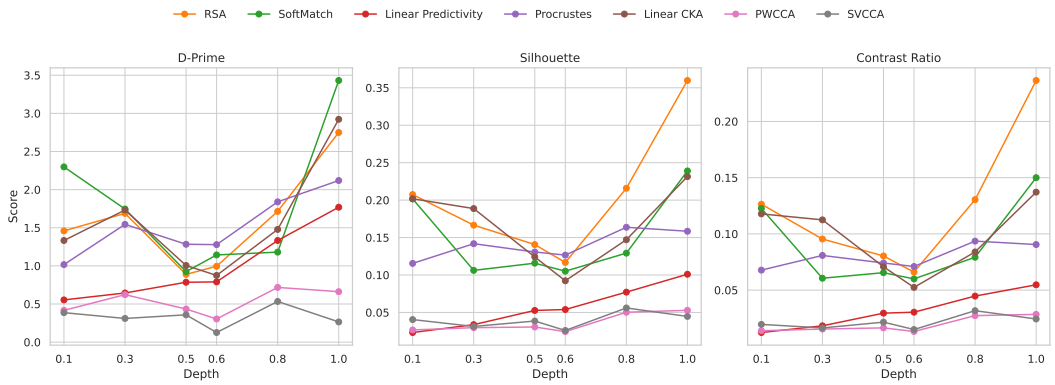

Figure I.3: Model-family separation across depth on CIFAR100.

J BIOLOGICAL BENCHMARK DETAILS

Data. We use the Allen Brain Observatory Visual Coding Neuropixels recordings (de Vries et al., 2020; Siegle et al., 2021), restricted to the `brain_observatory_1.1` sessions, the only session type that presents natural movie three (120 s at 30 Hz, giving 3,600 time bins). We use the preprocessed responses from the tutorial of Williams et al. (2024), in which units pass quality criteria at least as strict as the AllenSDK defaults (amplitude cutoff < 0.1, presence ratio > 0.95, ISI violations < 0.5) and each unit is assigned to a region by looking up its channel's CCFv3 coordinate in the annotation volume (Wang et al., 2020). This yields 55 regions with at least 50 units each; cortical regions are resolved by layer and subcortical regions by nucleus or subfield. Each region is represented by its time-bin-by-unit response matrix, normalized to unit Frobenius norm, and every similarity metric compares these matrices over the shared 3,600 time bins.

Grouping. A group plays the role of a model family. We assign each region to an ancestor in the CCFv3 structure tree, so that every group is a single CCF structure, every member is a descendant of it, and no group contains another (Table J.1). Cortical layers are grouped by area and subcortical nuclei or subfields one or two levels higher, which gives nine groups of three to seven regions. Fourteen regions are excluded. The AUDd and AUDpo layers lie mostly in visual area AL under the release's own labels and would duplicate VISal; VISl and VISli each have only one layer above the unit threshold; TEa has only deep layers and no parent below isocortex; LGv belongs to the ventral thalamus rather than the dorsal groups; APN and NOT would form a two-region group; and TH is recorded as whole thalamus and would overlap every thalamic group. Coarser groupings are not usable for separation: grouping by the four CCF systems puts 33 of 52 regions in one group, and grouping by area leaves many singleton groups with no within-group variance.

Table J.1: ABO anatomical groups used for separation (41 regions). Cortical group names are CCFv3 labels of the recording sites; because the parcellation comes from the annotation volume rather than the release's curated area identity, units in SSp-bfd lie mostly in visual area RL, units in VISam mostly in PM, and units in VISa mostly in AM.

Group	CCF structure	Regions
VISp	primary visual area	VISp2/3, 4, 5, 6a
VISal	anterolateral visual area	VISal2/3, 4, 5, 6a
VISam	anteromedial visual area	VISam2/3, 4, 5, 6a
VISa	anterior area	VISa2/3, 4, 5, 6a
VISrl	rostrolateral visual area	VISrl2/3, 4, 5
SSp-bfd	primary somatosensory area, barrel field	SSp-bfd2/3, 4, 5, 6a
HPF	hippocampal formation	CA1, CA3, DG-mo, DG-po, DG-sg, SUB, ProS
DORsm	thalamus, sensory–motor cortex related	LGd-co, LGd-ip, LGd-sh, MGd, MGv, VPM
DORpm	thalamus, polymodal association cortex related	LP, PO, POL, SGN, Eth

Within-group similarities are those between different regions of the same group, and between-group similarities those between regions of two groups. Unlike NSD, ABO has no cross-subject pairing, so this benchmark tests anatomical grouping rather than cross-individual correspondence. The same three separation measures are used in both datasets, but their magnitudes are not a controlled comparison of recording modalities. Because the groups are small (three to seven regions), SNF on ABO uses $K = 3$.

J.1 VENTRAL-STREAM HIERARCHY ALIGNMENT

Following Thobani et al. (2025), we evaluate a coarse hierarchy in the five NSD ventral-stream regions. We assign $L(\mathrm{V1v}) = 1$, $L(\mathrm{V2v}) = 2$, $L(\mathrm{V3v}) = 3$, $L(\mathrm{V4}) = 4$, and $L(\mathrm{ventral_visual}) = 5$. The analysis uses the corresponding 20 region instances across four subjects, not all ten NSD regions. For similarity matrix $\mathbf{S}$, define $H_{ij} = |L(r_i) - L(r_j)|$ and compute Kendall's τ-b between off-diagonal hierarchy distances and negative similarities:

$$\tau_{\mathrm{hier}} = \tau_b(\{H_{ij} : i \neq j\}, \{-S_{ij} : i \neq j\}) .$$

The scores are computed from the similarity matrices before any dimensionality reduction, so they do not depend on the PCA layouts in Figure 7. Region-pair observations share region instances, so nominal p-values for these scores should not be read as independent-subject replication.

K METRICS' SIMILARITY SCORES CONSISTENCY

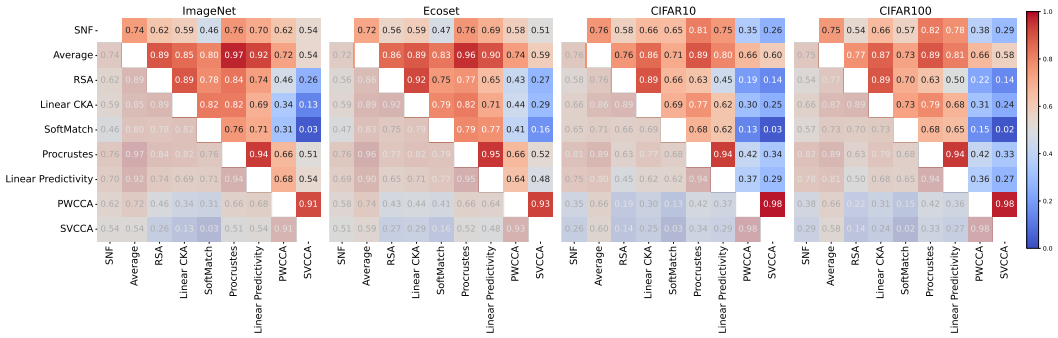

Figure K.1: Each subplot (ImageNet, Ecoset, CIFAR10, CIFAR100) shows pairwise Pearson correlations between the vectorized upper-triangle entries of the symmetrized model–model similarity matrices produced by nine metrics. Higher values indicate that two metrics have more similar relational geometry among models.

L More Clustering Performance

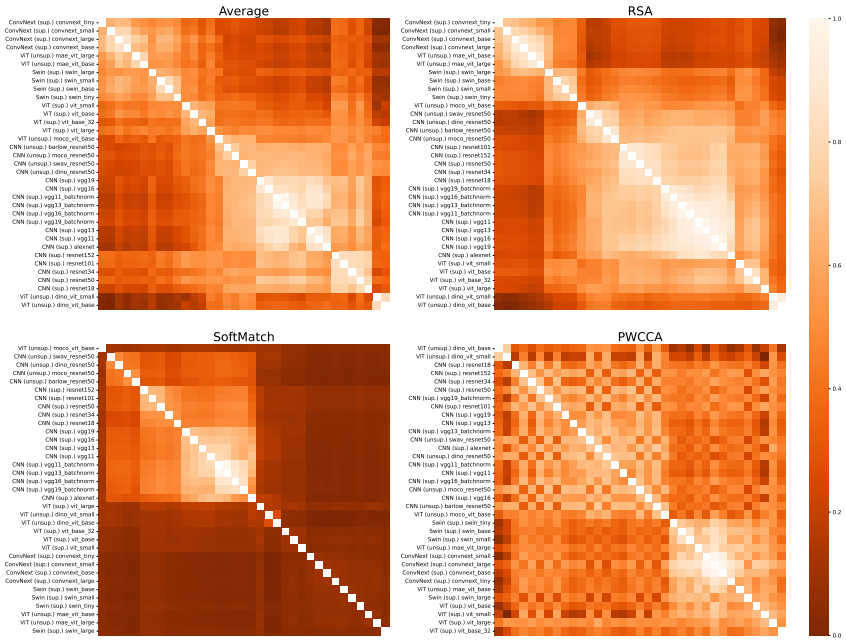

Figure L.1: Hierarchical clustering of models using three functionally distinct representational similarity metrics and a similarity-metric–averaging baseline (see Fig. L.2 for additional metrics). Clustering is performed with average linkage and optimal leaf ordering, based on induced distances ($1-$ similarity score). Rows and columns are reordered to match the leaf ordering produced by the clustering algorithm. Lighter colors indicate higher similarity; diagonal entries (self-comparisons) are omitted.

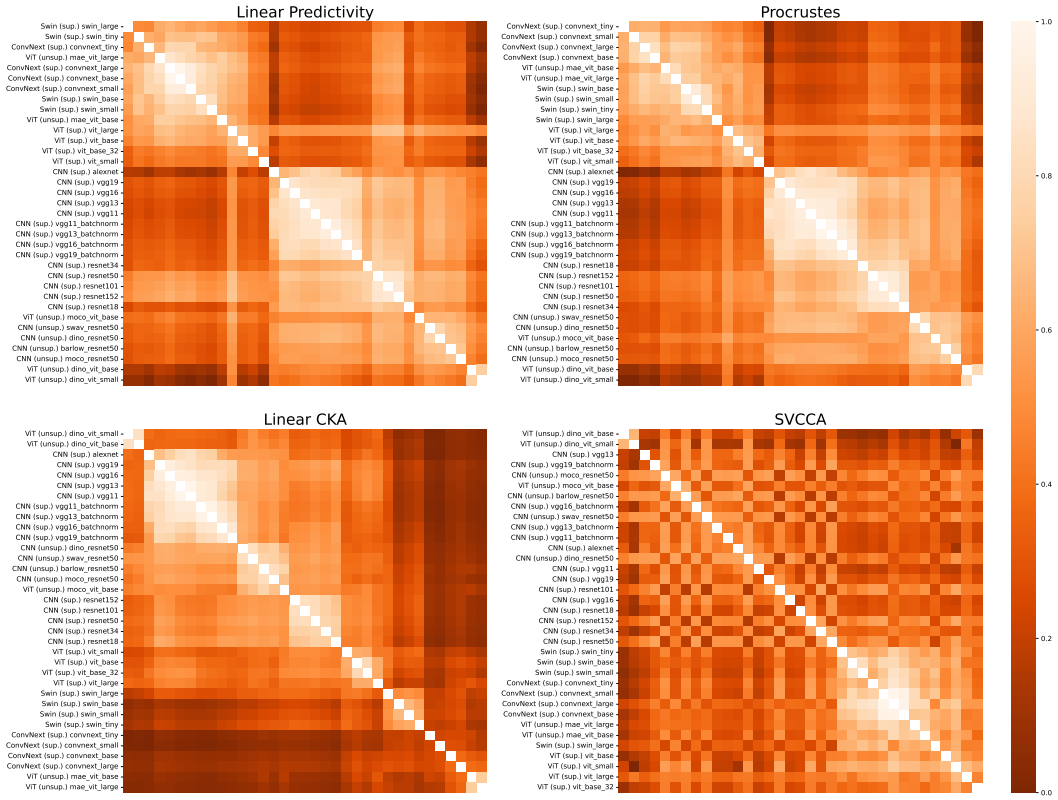

Figure L.2: Same as Figure L.1, but for the other 4 metrics.

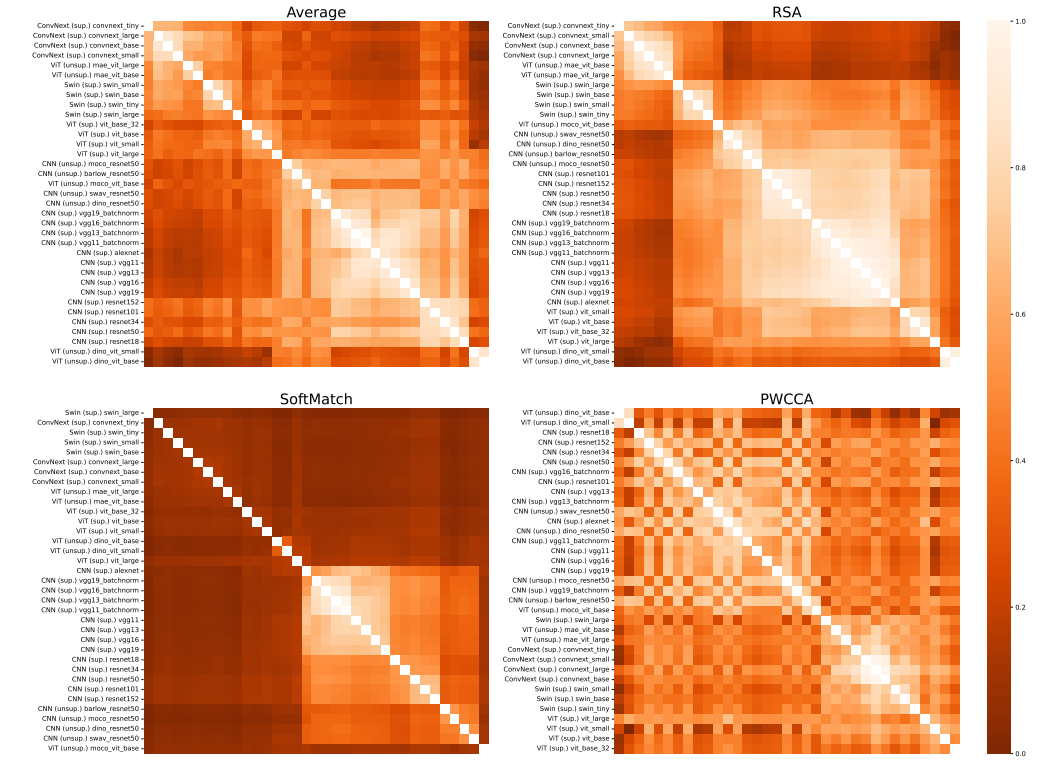

Figure L.3: Same as Figure L.1, but using Ecoset instead of ImageNet.

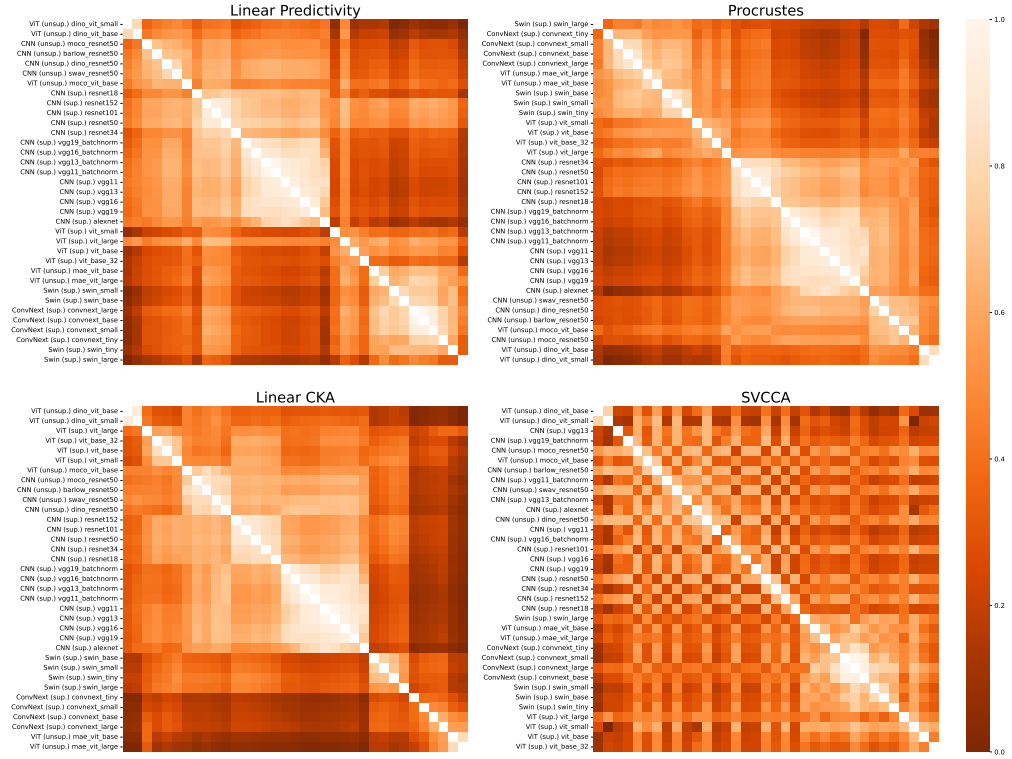

Figure L.4: Same as Figure L.2, but using Ecoset instead of ImageNet.

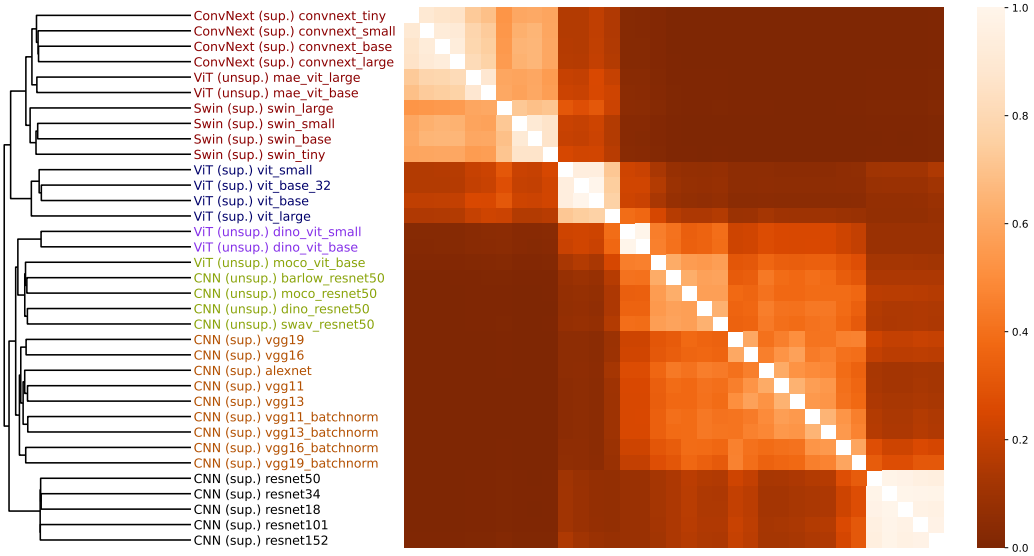

Figure L.5: Similar to Figure 6, but using Ecoset instead of ImageNet.

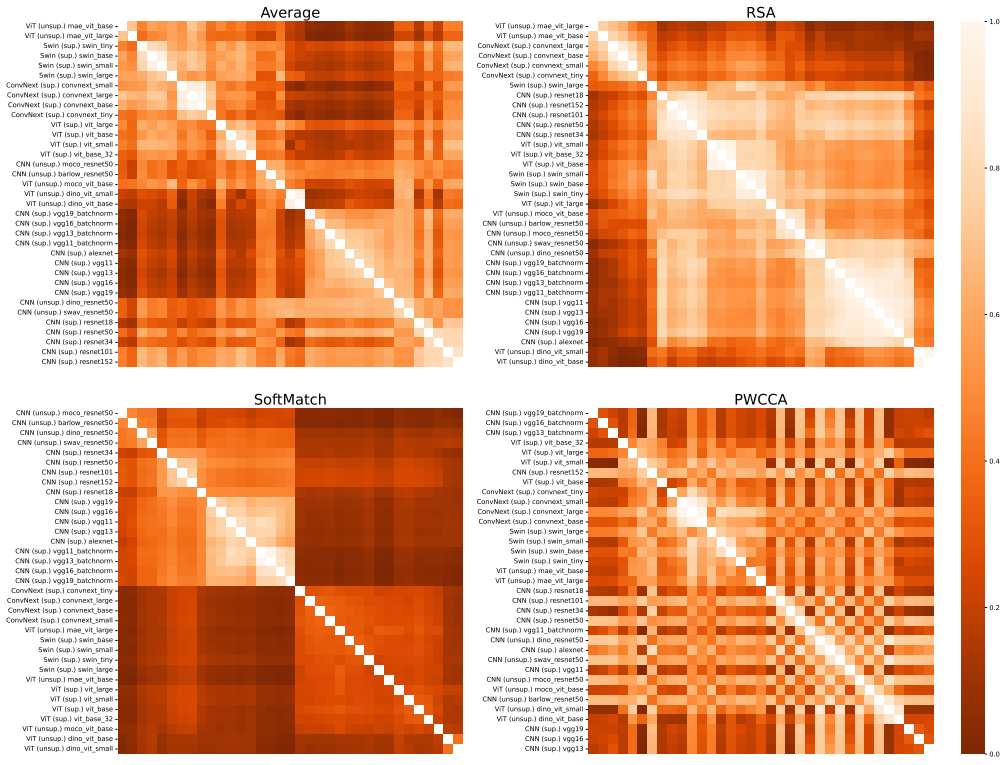

Figure L.6: Same as Figure L.1, but using CIFAR10 instead of ImageNet.

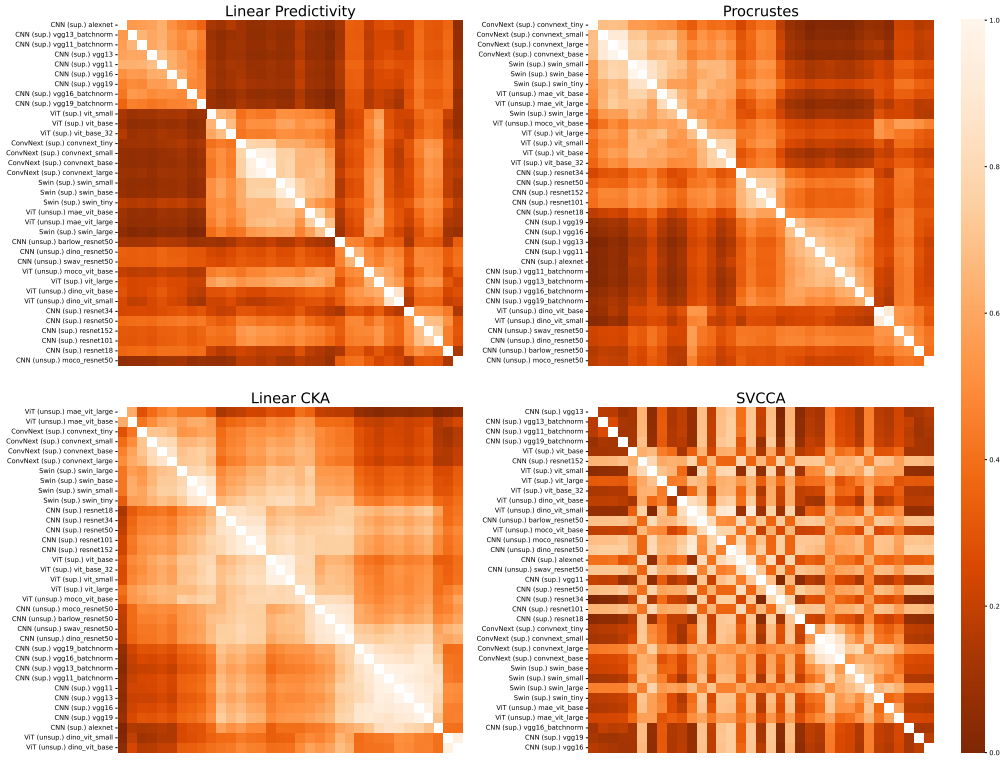

Figure L.7: Same as Figure L.2, but using CIFAR10 instead of ImageNet.

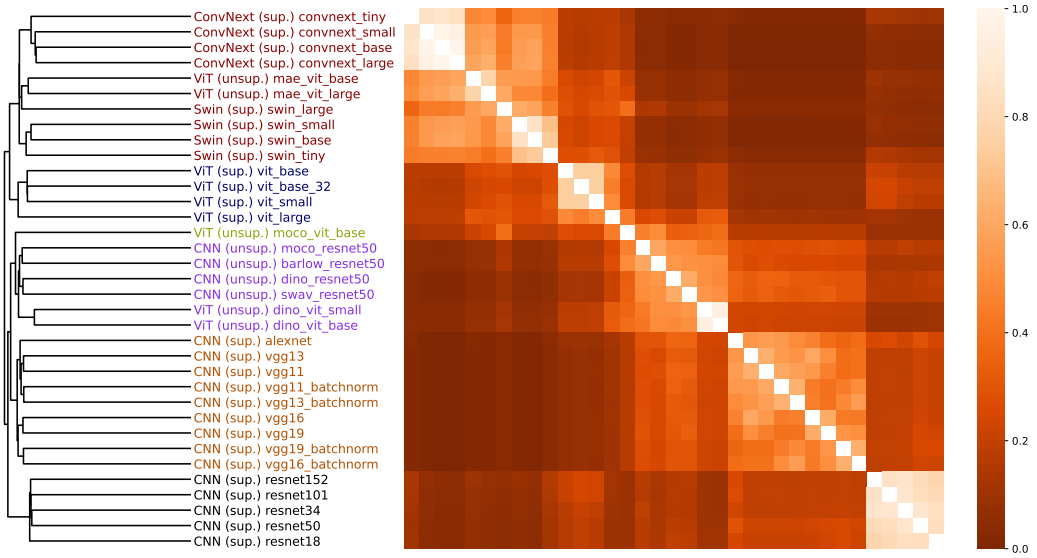

Figure L.8: Similar to Figure 6, but using CIFAR10 instead of ImageNet. Compared to ImageNet in Figure 6, the clustering result is a little different but similar, demonstrating the SNF results and metric's separability are also influenced by datasets, while also preserving a certain degree of stability. The supervised models are clustered in the same way, but the clusters for the self-supervised models changed, demonstrating that the self-supervised way leads to a special but kind of unifying representational geometry.

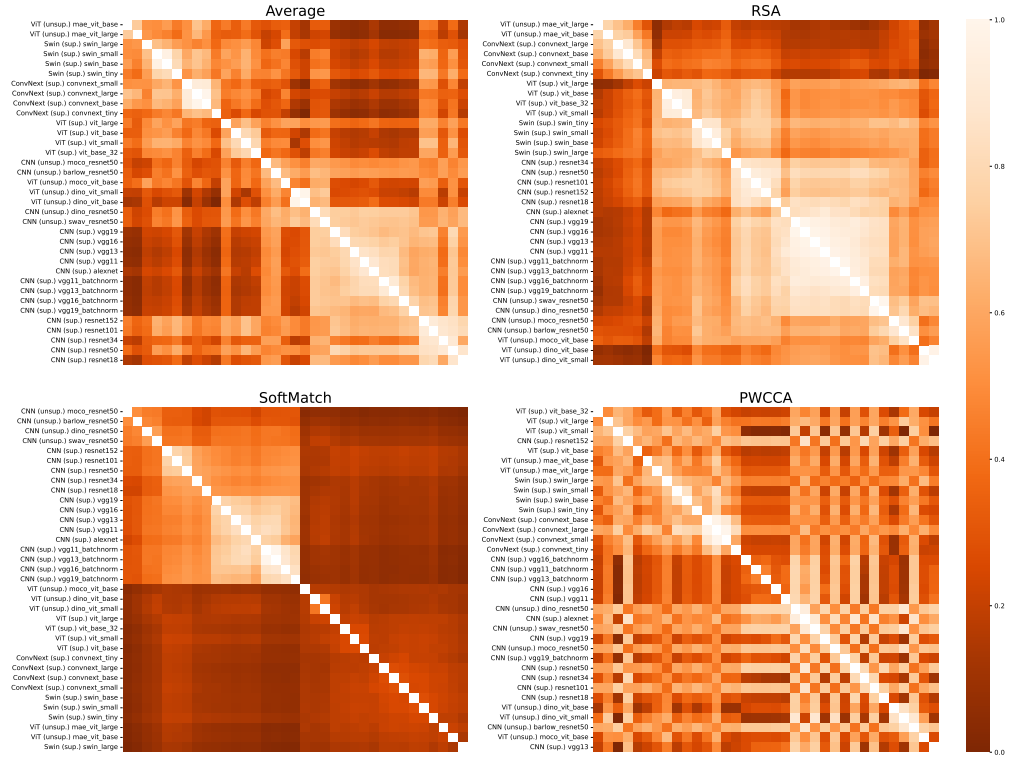

Figure L.9: Same as Figure L.1, but using CIFAR100 instead of ImageNet.

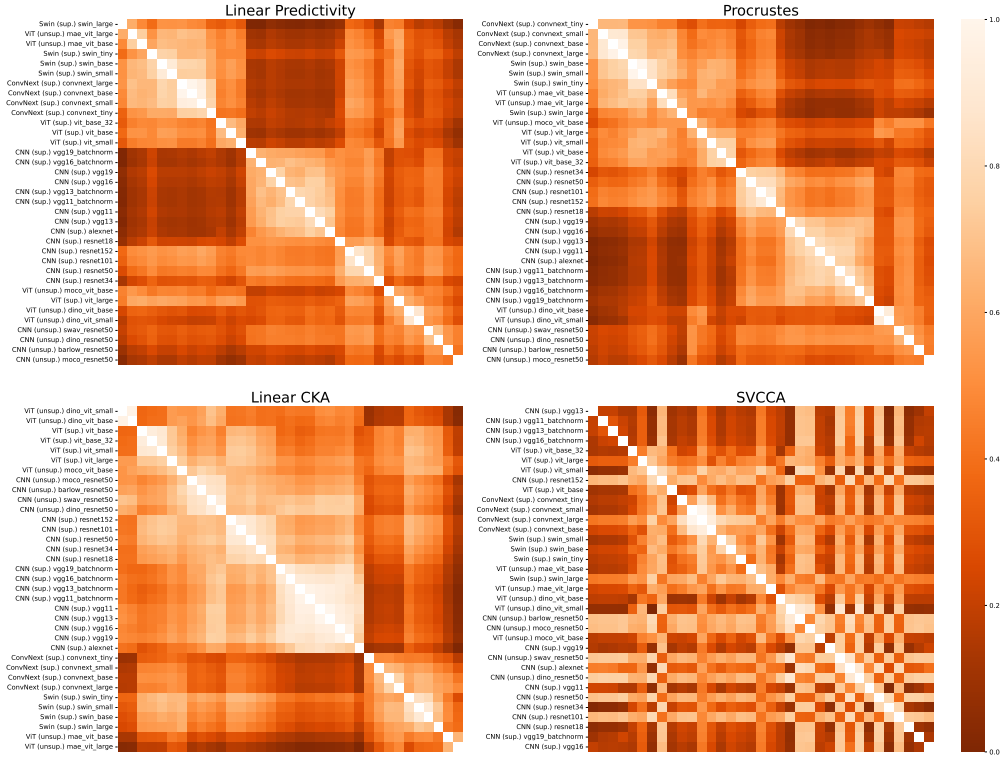

Figure L.10: Same as Figure L.2, but using CIFAR100 instead of ImageNet.

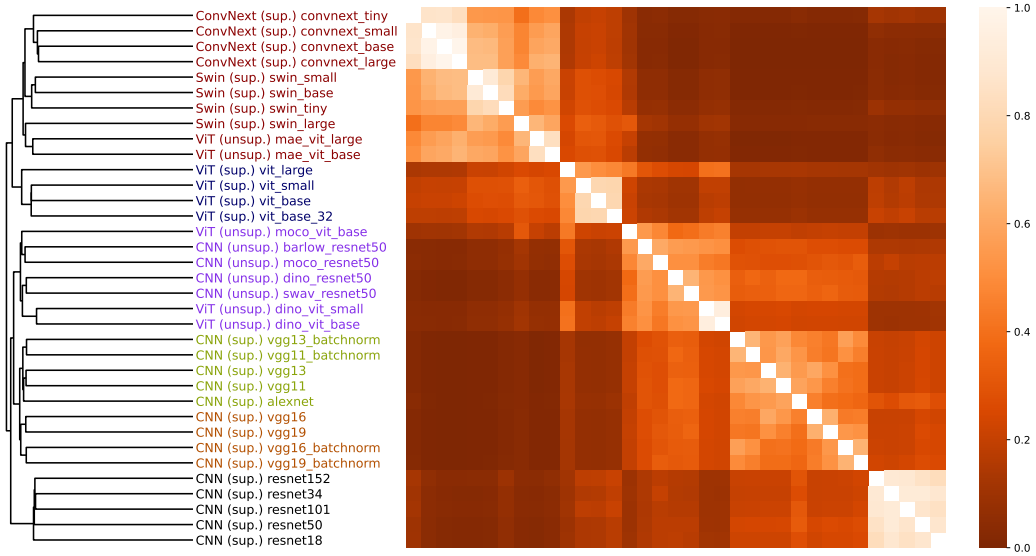

Figure L.11: Same as Figure L.8, but using CIFAR100 instead of CIFAR10.

M Cophenetic Correlation Coefficients for Clustering

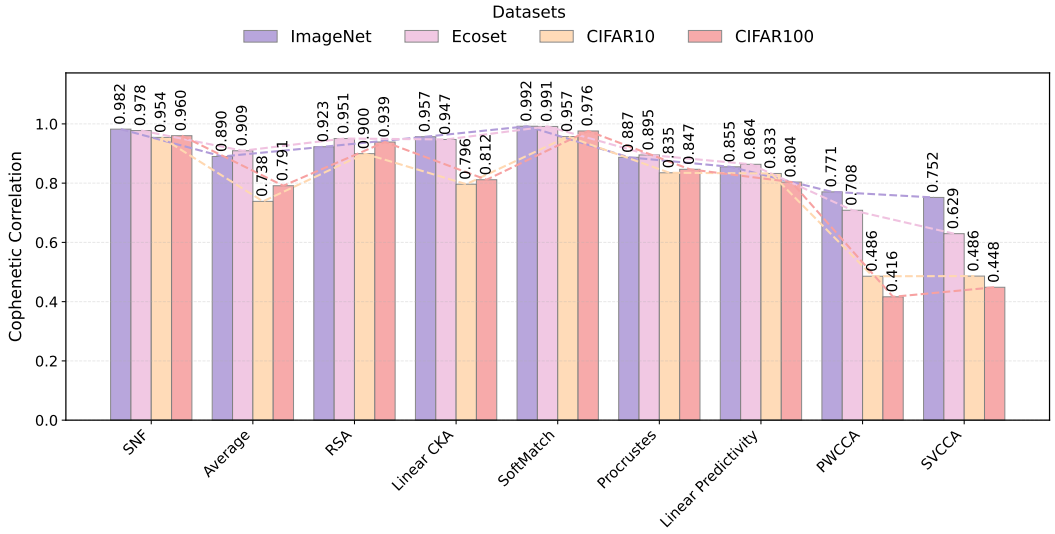

Figure M.1: Cophenetic correlation coefficients (CCC) for hierarchical clusterings induced by each metric on four datasets. Higher CCC (closer to 1) means the clustering more faithfully preserves the original pairwise structure. Bars show CCC on ImageNet, Ecoset, CIFAR10, and CIFAR100 (values annotated; lines trace dataset-wise trends).

N METRICS' CROSS-LAYER CONSISTENCY

Cross-layer consistency asks whether the inter-model similarity pattern persists across depth, a different question from whether each layer separates model families (Appendix I). We compare vectorized off-diagonal similarity entries across layers using Pearson correlations. In the updated results, Linear Predictivity, SoftMatch, and SNF are relatively consistent, whereas Linear CKA and the CCA variants show lower consistency. In particular, Linear Predictivity's cross-layer stability coexists with weaker family separation, so these two criteria should not be conflated.

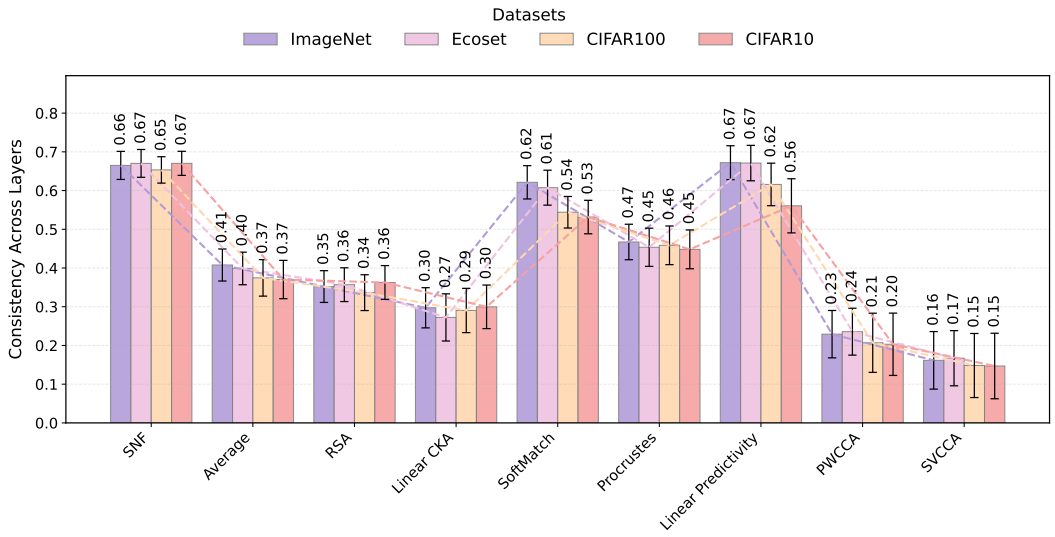

Figure N.1: Cross-layer consistency of inter-model similarity. Bars show the reported mean correlations across depth pairs; higher values indicate greater consistency, not necessarily stronger family separation.

O SIMILARITY NETWORK FUSION DETAILS

We use the existing Similarity Network Fusion algorithm (Wang et al., 2014; Markello, 2020) as a secondary integration baseline. Figure O.1 illustrates the construction. Let n be the number of models, and $\mathcal{V}$ be the set of the representational metrics. For each metric $v \in \mathcal{V}$, we obtain a similarity matrix $\mathbf{S}^v \in \mathbb{R}^{n\times n}$, where each entry $\mathbf{S}^v_{ij}$ measures the similarity between the model i's representation and model j's according to metric v, as described in Section 2.2. Then, for each metric v, we first convert pairwise scores into a dissimilarity matrix $\mathbf{Q}^v$ by $\mathbf{Q}^v_{ij} = 1_{\mathrm{i}\neq\mathrm{j}}\big(1 - (\mathbf{S}^v_{ij} + \mathbf{S}^v_{ji})/2\big)$. We then build an affinity $\mathbf{W}^v \in \mathbb{R}^{n\times n}$ with an exponential kernel:

$$\mathbf{W}^v_{ij} = \frac{1}{\sqrt{2\pi(\sigma^v_{ij})^2}} \exp\left(-\frac{(\mathbf{Q}^v_{ij})^2}{2(\sigma^v_{ij})^2}\right), \quad \sigma^v_{ij} = \mu \cdot \frac{\overline{\mathbf{Q}}^v(i,\widetilde{N_i}^v) + \overline{\mathbf{Q}}^v(j,\widetilde{N_j}^v) + \mathbf{Q}^v_{ij}}{3}$$

Here, $\overline{\mathbf{Q}}^v(i,\widetilde{N_i}^v)$ denotes the average dissimilarity from i to its K nearest neighbors $\widetilde{N_i}^v$ under metric v. We set $\mu = 0.5$ and use $K = 5$ for the model population and NSD and $K = 3$ for ABO. We also carried out a systematic hyperparameter sweep. The results (Figure Q.1,Q.2, Q.3, Q.4, Q.5) indicate that SNF performs well across a fairly wide range of hyperparameters.

We view each $\mathbf{W}^v$ as a weighted graph and aim to fuse them into a single matrix that emphasizes relationships consistently supported across metrics while suppressing spurious ones. Following the implementation, we form a row-normalized full matrix and a KNN-sparse matrix for each metric:

$$\mathbf{C}^v_{ii} = \textstyle\sum_j \mathbf{W}^v_{ij}, \quad \widetilde{\mathbf{W}}^v = (\mathbf{C}^v)^{-1}\mathbf{W}^v, \quad \widehat{\mathbf{W}}^v = \tfrac{1}{2}(\widetilde{\mathbf{W}}^v + (\widetilde{\mathbf{W}}^v)^\top)$$

$$\widetilde{\mathbf{S}}^v_{ij} = \widehat{\mathbf{W}}^v_{ij} / \textstyle\sum_{k\in\widehat{N_i}^v} \widehat{\mathbf{W}}^v_{ik}\text{, if } j \in \widehat{N_i}^v\text{, } K \text{ nearest neighbors based on } \widehat{W}^v; \quad \text{else } 0.$$

We then run the SNF message-passing updates with diagonal regularization, which keeps self-affinity dominant while improving numerical stability. Initialize $\mathbf{P}^v_0 = \widehat{\mathbf{W}}^v$. For $t = 0, \ldots, T-1$:

$$\mathbf{P}^v_{t+1} = \mathcal{B}_\alpha\left(\widetilde{\mathbf{S}}^v \left(\frac{1}{|\mathcal{V}|-1}\textstyle\sum_{u\neq v}\mathbf{P}^u_t\right)\widetilde{\mathbf{S}}^{v\top}\right), \quad \mathcal{B}_\alpha(\mathbf{X}) = \tfrac{1}{2}(\mathbf{X} + \mathbf{X}^\top) + \alpha\mathbf{I}$$

After T iterations, we average the networks and perform a row normalization and symmetrization:

$$\mathbf{P} = \frac{1}{|\mathcal{V}|}\textstyle\sum_{v\in\mathcal{V}}\mathbf{P}^v_T, \quad \mathbf{D}_{ii} = \textstyle\sum_j \mathbf{P}_{ij}, \quad \widetilde{\mathbf{P}} = \mathbf{D}^{-1}\mathbf{P}, \quad \widehat{\mathbf{P}} = \tfrac{1}{2}\big(\widetilde{\mathbf{P}} + \widetilde{\mathbf{P}}^\top + \mathbf{I}\big)$$

Lastly, to form a dendrogram for model typology, we cluster the fused affinity $\widehat{\mathbf{P}}$ with hierarchical clustering using SciPy's linkage function (Virtanen et al., 2020).

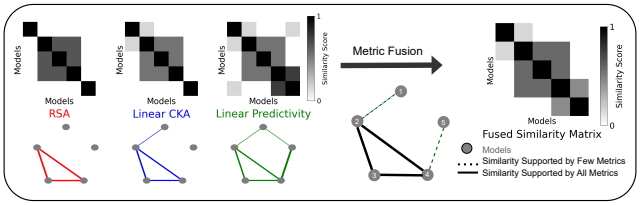

Figure O.1: **Top left:** Each representational metric defines a pairwise similarity matrix over models/brain regions. **Bottom left:** Each matrix is visualized as an affinity graph, with nodes representing models/brain regions and edge widths reflecting pairwise similarity strength; weak similarities below a threshold are omitted for clarity. **Right:** A consensus matrix obtained via Similarity Network Fusion (SNF) highlights relations consistently supported across metrics while leveraging complementary signals. In the fused graph, solid edges denote agreement across all metrics, dotted edges indicate partial support; strong but uncorroborated edges may persist with reduced weight (e.g., edge 4–5); weak and metric-specific connections are typically suppressed (e.g., edge 1–5).

P LEAVE-ONE-OUT FUSION

To test whether SNF depends on any single input, we remove one metric at a time and fuse the remaining ones (Table P.1). Both analyses use $K = 5$, as in the main text. On ImageNet, removing any one metric keeps d' between 11.70 and 13.20, far above the best single metric (RSA, 3.95), and changes silhouette score and contrastive ratio by about 0.02 or less. Removing Linear CKA or Procrustes lowers d' the most, whereas removing RSA or Linear Predictivity raises it. On NSD, removing SoftMatch lowers d' the most (22.52 to 18.91) and removing Procrustes raises it (27.85), while silhouette score and contrastive ratio change by less than 0.01. The fused separation is therefore not driven by one metric, but its magnitude depends on which metrics are included.

Table P.1: Leave-one-out SNF separation (mean $\pm$ standard error across group pairs). "None" fuses all metrics; other rows omit the named metric. NSD uses five metrics, as PWCCA and SVCCA are excluded there.

Dataset	Removed metric	d'	Silhouette	Contrastive ratio
ImageNet	None	12.42 ± 2.93	0.695 ± 0.067	0.584 ± 0.083
	RSA	13.20 ± 3.36	0.678 ± 0.070	0.567 ± 0.086
	Linear CKA	11.70 ± 2.76	0.680 ± 0.069	0.567 ± 0.085
	SoftMatch	12.75 ± 2.95	0.707 ± 0.066	0.597 ± 0.084
	Procrustes	11.73 ± 2.80	0.690 ± 0.068	0.578 ± 0.084
	Linear Predictivity	13.02 ± 3.12	0.693 ± 0.067	0.582 ± 0.084
	PWCCA	12.34 ± 2.89	0.716 ± 0.063	0.604 ± 0.080
	SVCCA	12.16 ± 2.80	0.716 ± 0.063	0.607 ± 0.081
NSD	None	22.52 ± 2.46	0.788 ± 0.047	0.741 ± 0.052
	RSA	22.60 ± 2.28	0.787 ± 0.048	0.741 ± 0.053
	Linear CKA	21.03 ± 2.29	0.784 ± 0.048	0.737 ± 0.053
	SoftMatch	18.91 ± 1.97	0.782 ± 0.048	0.734 ± 0.053
	Procrustes	27.85 ± 3.43	0.793 ± 0.047	0.746 ± 0.052
	Linear Predictivity	22.50 ± 2.46	0.796 ± 0.046	0.748 ± 0.052

Q HYPERPARAMETER SWEEP

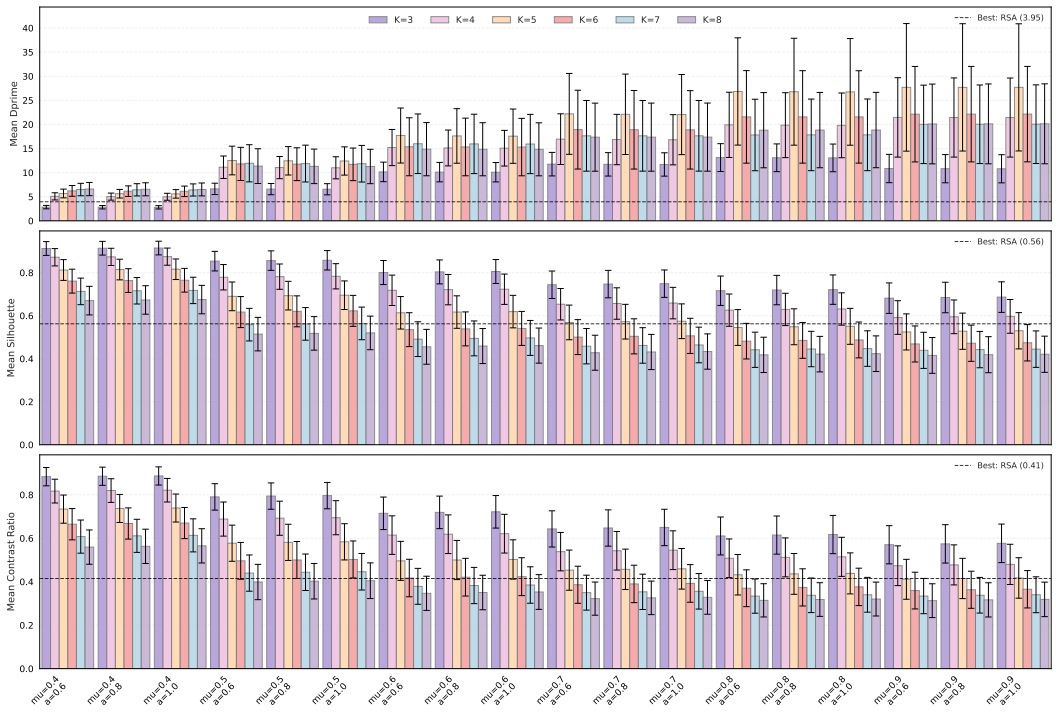

Figure Q.1: Grouped barplots of SNF hyperparameter sweep results on ImageNet under d', silhouette score and contrast ratio. The three rows correspond to the three separability metrics. Within each row, bar colors denote different neighborhood sizes K, and the x-axis enumerates (μ, α) combinations, with $T = 20$ iterations. Each bar shows the mean model-family separability with standard error. The dashed horizontal line marks the strongest single representational metric among RSA, SoftMatch, Procrustes, Linear CKA, Linear Predictivity, PWCCA, and SVCCA for the corresponding measure. Scores are shown in their native scales and are not directly comparable across measures.

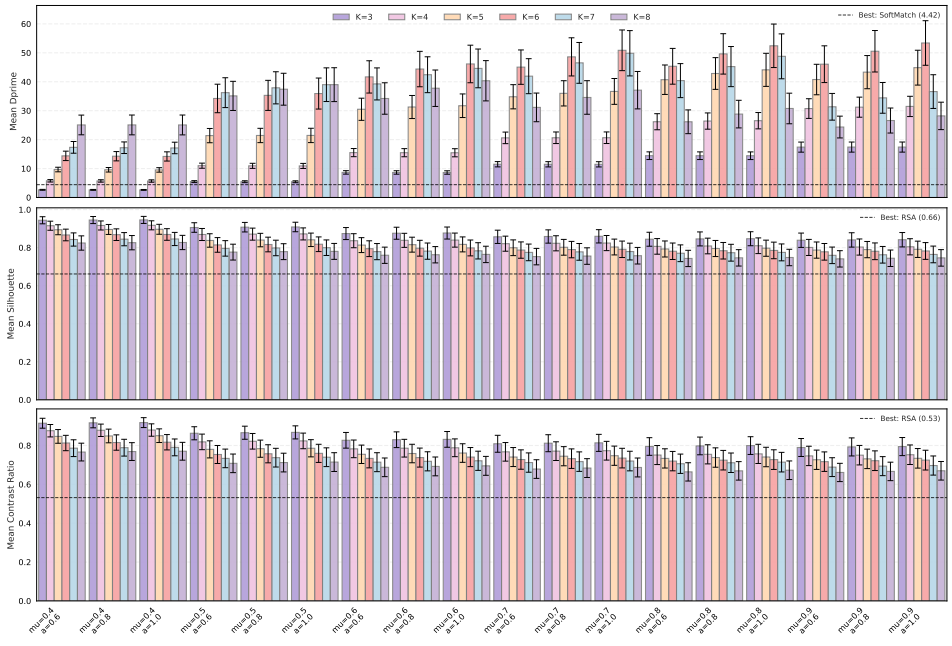

Figure Q.2: Similar to Figure Q.1, but on human brain responses (NSD dataset) and excluding SVCCA and PWCCA.

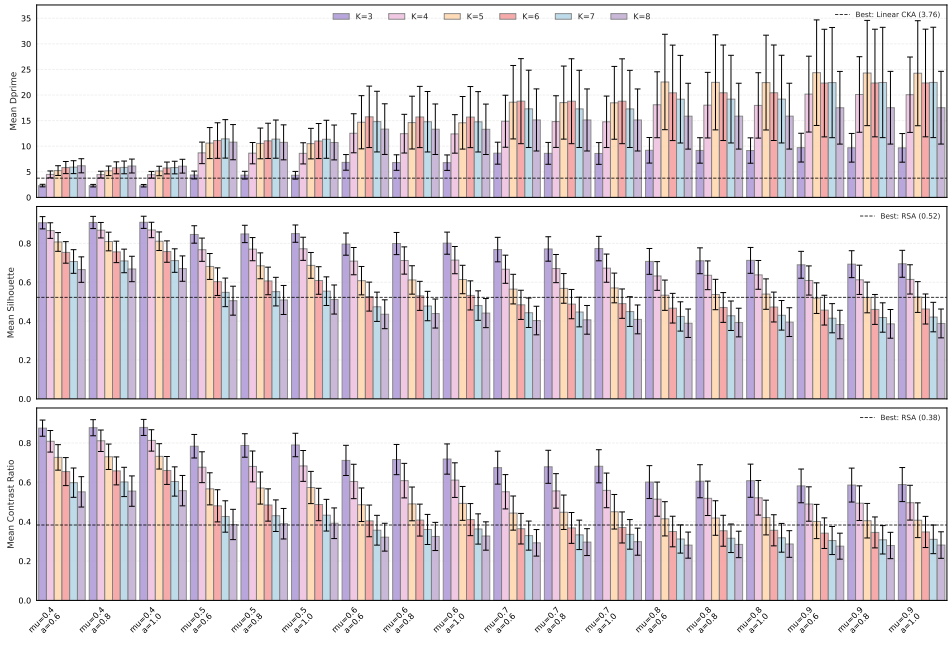

Figure Q.3: Same as Figure Q.1, but using Ecoset instead of ImageNet.

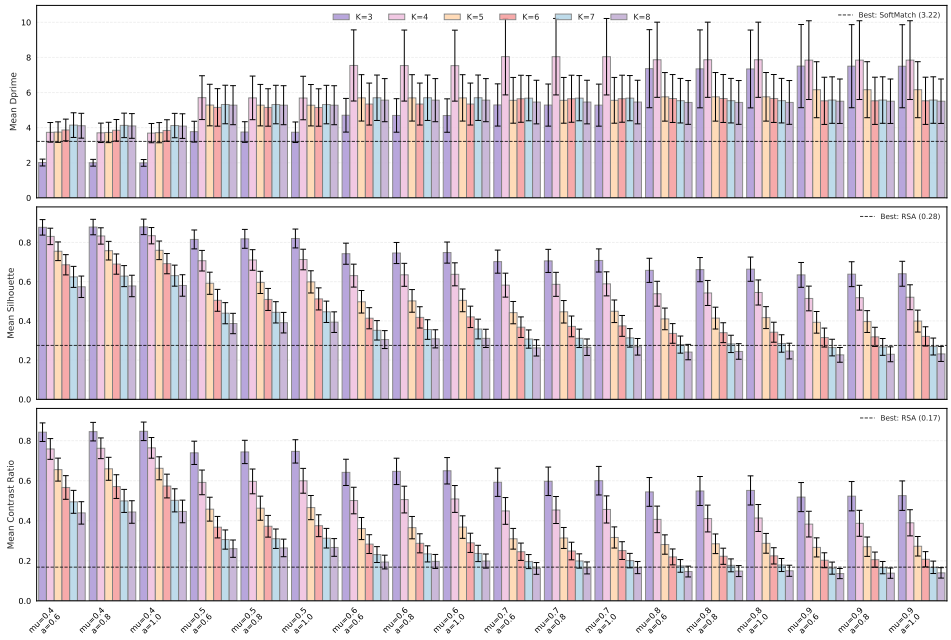

Figure Q.4: Same as Figure Q.1, but using CIFAR10 instead of ImageNet.

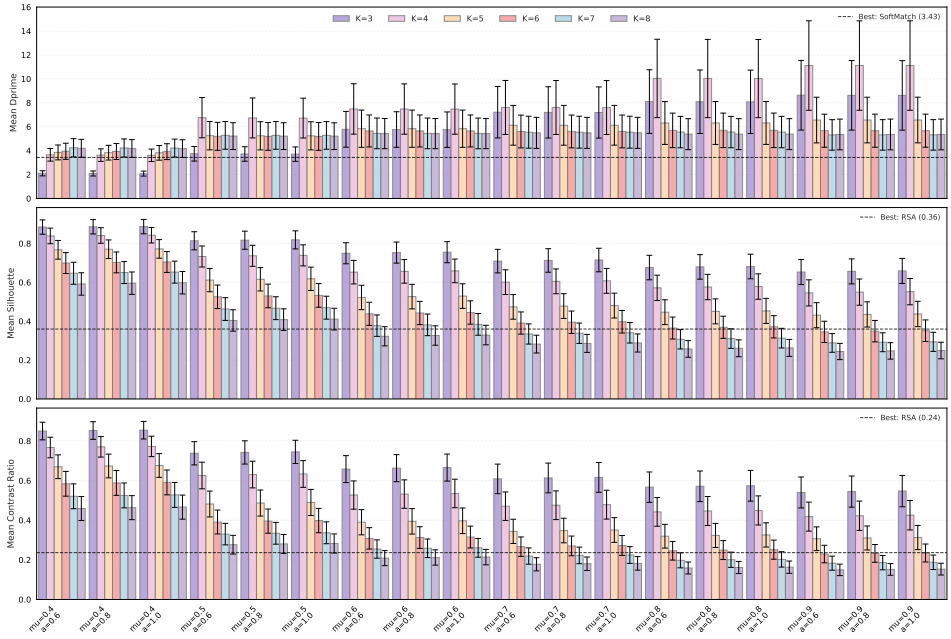

Figure Q.5: Same as Figure Q.1, but using CIFAR100 instead of ImageNet.

R SIMILARITY MATRIX PERMUTATION

To show that the fused similarity depends on genuine signal integration rather than graph-sharpening artifacts, we performed a control analysis using two complementary metrics, RSA and SoftMatch. We permuted one input similarity matrix while keeping the other fixed. As shown in Figure R.1, when one input similarity matrix is progressively permuted, SNF performance degrades accordingly, indicating that SNF depends on real and shared representational structure.

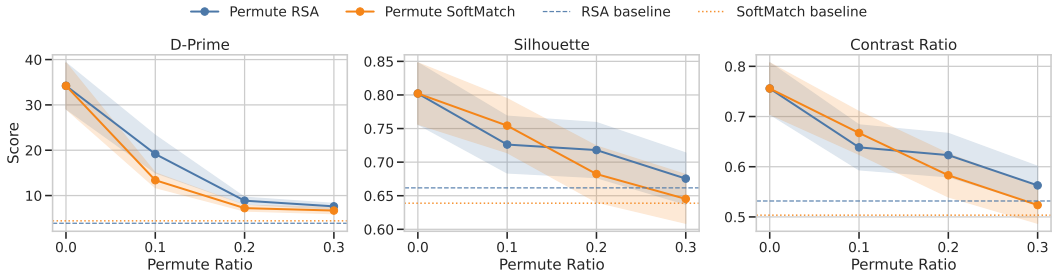

Figure R.1: Control experiment with partial entry-wise permutation of raw similarity matrices on NSD. Starting from the raw RSA and SoftMatch similarity matrices, one matrix is kept fixed while the other is perturbed by shuffling a randomly selected subset of off-diagonal upper-triangular entries at ratios of 0, 0.1, 0.2, and 0.3. The shuffled entries are mirrored to the lower triangle to preserve symmetry. The resulting pair of matrices is fused with SNF. The three panels report the resulting brain region separability under d', silhouette score and contrast ratio. The blue and orange curves correspond to perturbing RSA and SoftMatch, respectively; the horizontal dashed and dotted lines indicate the corresponding RSA and SoftMatch baselines. For each panel, the plotted score is the mean over the 45 unique off-diagonal entries of the 10×10 region-level matrix, with standard error shown as the shaded band. Scores are shown in their native scales and are not directly comparable across measures. The permutation uses a random seed $= 42$.