

Comparing and Integrating Different Notions of Representational Correspondence in Neural Systems

Jialin Wu¹ Shreya Saha² Yiqing Bo¹ Meenakshi Khosla^{1,3}

Abstract

The extent to which different biological and artificial neural systems rely on equivalent internal representations to support similar tasks remains a central question in neuroscience and machine learning. Prior work typically compares systems using a single representational similarity metric, even though different metrics emphasize distinct facets of representational correspondence. Here we evaluate a suite of representational similarity measures by asking how well each metric recovers known structure across two domains: for artificial models, whether procedurally dissimilar models (differing in architecture or training paradigm) are assigned lower similarity than procedurally matched models; and for neural data, whether responses from distinct cortical regions are separated while responses from the same region align across subjects. Across both vision models and neural recordings, metrics that preserve representational geometry or tuning structure more reliably separate this structure than more flexible mappings such as linear predictivity. To integrate these complementary facets, we adapt Similarity Network Fusion, originally developed for multi-omics integration, to combine similarity graphs across metrics. The resulting fused similarity yields sharper separation of procedurally defined model families and, when applied to neural data, recovers a clearer hierarchical organization of the ventral visual stream that aligns more closely with established anatomical and functional hierarchies than single metrics. Overall, this approach reveals which dimensions of representational correspondence recover meaningful structure in models and brains, and how complementary notions of similarity can be integrated.

¹Department of Computer Science and Engineering, UC San Diego, San Diego, CA 92037, USA ²Department of Electrical and Computer Engineering, UC San Diego, San Diego, CA 92037, USA ³Department of Cognitive Science, UC San Diego, San Diego, CA 92037, USA. Correspondence to: Meenakshi Khosla <mkhosla@ucsd.edu>.

Preprint. February 24, 2026.

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.), 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 inherently multidimensional. Some aspects of representation may be widely shared across systems, reflecting convergent computational solutions, while others may encode signatures that are specific to particular architectures, training objectives, or biological substrates. As a result, representational comparison remains methodologically fragmented, with different similarity measures often yielding conflicting views of correspondence. This raises a fundamental question: which dimensions of representational correspondence can recover meaningful structure?

Answering this question requires domain-appropriate criteria. In artificial neural networks, a natural benchmark is *procedural correspondence*: whether a similarity measure distinguishes models trained under different architectures or learning paradigms while grouping procedurally matched models. In neural data, an analogous benchmark is *anatomical-functional organization*: whether responses from distinct cortical regions are separable while responses from the same region align across subjects. Together, these criteria provide principled tests for evaluating which representational similarity measures capture meaningful structure.

Using vision models and visual cortical recordings as a controlled testbed, this work makes three central contributions.

First, we provide a systematic evaluation of representational similarity measures capturing distinct dimensions of correspondence (e.g. geometry, unit-level tuning, and linearly accessible information) using principled criteria in both domains. We find that metrics preserving representational geometry or unit-level tuning (e.g., RSA and soft-matching variants (Khosla & Williams, 2024)) more reliably recover known structure, separating procedurally distinct models and differentiating cortical regions, yet more flexible mappings such as linear predictivity yield weaker discrimination.

Second, recognizing that these representational dimensions are complementary and that no single metric provides a complete characterization, we introduce an integrative framework based on Similarity Network Fusion (SNF) (Wang et al., 2014). By combining similarity graphs derived from multiple metrics, the fused representation sharpens recovery of procedural structure among models and reveals clearer anatomical-functional organization across cortical regions than any individual metric alone.

Finally, we show that integrating complementary similarity dimensions enables the construction of meaningful, data-driven typologies of both artificial models and neural systems. This approach follows empirical traditions in psychology, neuroscience, and genetics (Wang et al., 2014; Letwin et al., 2006; Echtermeyer et al., 2011; Mukamel & Ngai, 2019) where researchers identify natural groupings by combining multiple behavioral or molecular indices. For AI models, clustering the SNF-fused similarity matrix reveals groupings that emerge from how systems organize information rather than from prior assumptions about architecture or training tasks. For neural data, the fused similarities expose clear cross-subject structure: responses from the same cortical area form compact clusters that are invariant across subjects, with robust separation between areas. These results demonstrate that integrating across representational dimensions not only improves discriminability, but also supports principled discovery of organization consistent with known anatomical and functional hierarchies.

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, super-

vised 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. 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 the Natural Scenes Dataset (NSD) (Allen et al.), spanning 10 visual regions across 4 subjects. This dataset enables analogous population-level comparisons in the brain, allowing us to evaluate whether representational similarity measures separate responses from distinct cortical regions while preserving cross-subject alignment within the same region.

2.2. Representational Metrics

We evaluate widely used similarity metrics that differ in the flexibility of the mappings they permit—from permutation-based alignments (soft-matching) to rigid geometric transformations (Procrustes) to looser linear mappings (linear predictivity) as well as non-fitting approaches that compare representational geometry directly (RSA). Consider two representations $\mathbf{X}_i \in \mathbb{R}^{M \times N_i}$ and $\mathbf{X}_j \in \mathbb{R}^{M \times N_j}$ from different models, 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 top N'_i and N'_j singular vectors that explain 99% of the variance yields the reduced representations: $\mathbf{X}'_i = \mathbf{U}_i^{(N'_i)\top} \mathbf{X}_i$ and $\mathbf{X}'_j = \mathbf{U}_j^{(N'_j)\top} \mathbf{X}_j$ whose dominant singular directions capture a disproportionate share of the total information. 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: $\mathbf{Q} = \max_{\mathbf{A}, \mathbf{B}} \text{corr}(\mathbf{A}\mathbf{X}'_i, \mathbf{B}\mathbf{X}'_j)$ subject to unit-variance constraints, providing a final similarity score that reflects how closely the subspaces of the two representations align.

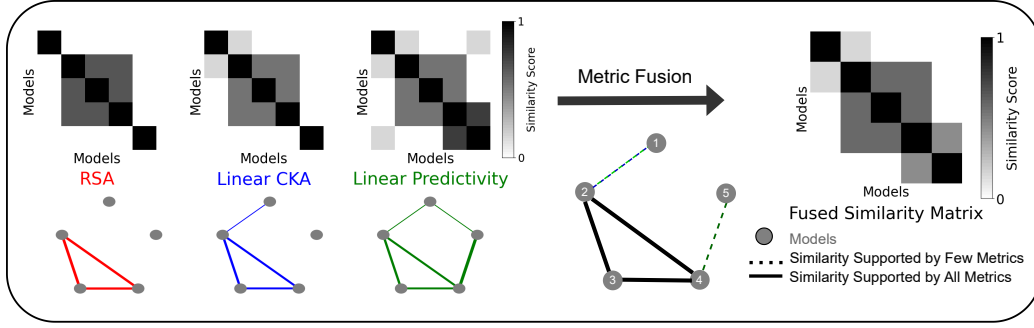


Figure 1. Top left: Each representational metric defines a pairwise similarity matrix over models/brain regions. **Bottom left:** Each matrix is visualized as an affinity graph, with nodes representing models/brain regions and edge widths reflecting pairwise similarity strength; weak similarities below 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).

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 directions according to their contribution to the original representation. After CCA, we obtain canonical vectors $\mathbf{A}$ and $\mathbf{B}$ and their corresponding correlations q_i for $i = 1, 2, \dots, k$, where $k = \min\{N_i, N_j\}$. Instead of giving each direction equal weight, PWCCA projects the unreduced representation $\mathbf{X}_i$ onto its own canonical vectors to measure how strongly each one reconstructs the data. The projection weight for the i^{th} canonical direction is $\alpha_i = \frac{\|\mathbf{X}_i \mathbf{a}_i\|_1}{\sum_{j=1}^k \|\mathbf{X}_i \mathbf{a}_j\|_1}$, where $\mathbf{a}_i$ is the i^{th} column of $\mathbf{A}$. The final PWCCA similarity is then the weighted sum of canonical correlations: $\sum_{i=1}^k \alpha_i q_i$ which emphasizes directions that explain the greatest fraction of variance in $\mathbf{X}_i$ and down-weights noisy, low-variance components, yielding a more faithful measure of representational similarity.

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 $\|\cdot\|_F$ is the Frobenius norm, and $\mathbf{X}_i$ and $\mathbf{X}_j$ can be assumed to be normalised. 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.). 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 $1 - \text{Pearson correlation}$, 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 $\mathbf{P}^*$ 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 $\mathbf{X}_i \mathbf{P}^*$.

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 $\mathcal{O}(N) = \{\mathbf{R} \in \mathbb{R}^{N \times N} : \mathbf{R}^T \mathbf{R} = \mathbf{I}\}$. The optimal transformation $\mathbf{R}^*$ is obtained via singular value decomposition. The similarity score is the mean unit-wise correlation between $\mathbf{X}_j$ and $\mathbf{X}_i \mathbf{R}^*$.

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 $\mathbf{L}^*$ is estimated via ordinary least squares. The final similarity score is the mean unit-wise correlation between $\mathbf{X}_j$ and $\mathbf{X}_i \mathbf{L}^*$.

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 Metrics

We next describe the measures used to assess how well representational similarity metrics recover procedural distinctions among models (i.e. how well they separate different model families).

Contrastive Ratio. The contrastive ratio quantifies the relative separation between intra-family and inter-family similarities. We consider the similarity values between different models within the same family and take the average of them to obtain μ_{within} , and the similarities between models from two model families and obtain the average inter-family μ_{between} . The ratio is then defined as $CR = (\mu_{\text{within}} - \mu_{\text{between}}) / (\mu_{\text{within}} + \mu_{\text{between}})$. A value approaching 1 suggests strong within-family coherence relative to cross-family similarity; a value approaching 0 suggests no difference, and a negative value implies that inter-family similarity exceeds intra-family similarity.

D-Prime (Bo et al., 2024). Similarly to contrastive ratio but considering variances, the D-Prime (d') also quantifies the separation between intra-family and inter-family similarities. It is defined as $(\mu_{\text{within}} - \mu_{\text{between}}) / \sqrt{0.5(\sigma_{\text{within}}^2 + \sigma_{\text{between}}^2)}$, where μ and σ^2 denote the mean and variance of the respective distributions. Higher values indicate tighter clustering within a family and greater spread across families.

Silhouette Score (Rousseeuw, 1987). For each model i , we compute the average distance $a(i)$ to all other models in the same family and the average distance $b(i)$ to models in the other family. The silhouette value is then $s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$, $s(i) \in [-1, 1]$. Values near 1 indicate that the model is well grouped with its own family, values near 0 suggest boundary placement, and negative values imply greater similarity to another family. The overall silhouette score is obtained by averaging $s(i)$ across models.

2.4. Similarity Network Fusion

Next, we sought to reconcile the results across the different evaluation metrics. As elaborated in Section 2.3, each metric captures distinct aspects of model representations and varies in its ability to differentiate between model families. To integrate these metrics, we adopt a unified approach inspired by Similarity Network Fusion (Wang et al., 2014; Markello, 2020). 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 can get a similarity matrix $\mathbf{S}^v \in \mathbb{R}^{n \times n}$, where each entry S_{ij}^v 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 $\mathbf{Q}^v$ as this equation, $Q_{ij}^v = 1_{i \neq j} (1 - (S_{ij}^v + S_{ji}^v)/2)$. We then build an affinity $\mathbf{W}^v \in \mathbb{R}^{M \times M}$ with a exponential kernel:

$$\mathbf{W}^v(i, j) = \frac{1}{\sqrt{2\pi(\sigma_{ij}^v)^2}} \exp\left(-\frac{(\mathbf{Q}_{ij}^v)^2}{2(\sigma_{ij}^v)^2}\right),$$

$$\sigma_{ij}^v = \mu \cdot \frac{\overline{\mathbf{Q}}^v(i, N_i) + \overline{\mathbf{Q}}^v(j, N_j) + \mathbf{Q}_{ij}^v}{3}.$$

Here, $\overline{\mathbf{Q}}^v(i, N_i)$ denotes the average dissimilarity from i to its K nearest neighbors N_i under metric v . We set the hyperparameter $\mu \in (0, 1)$ to 0.5 and K to 5 following the original paper.

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:

$$\mathbf{C}_{ii}^v = \sum_j \mathbf{W}_{ij}^v, \quad \widetilde{\mathbf{W}}^v = (\mathbf{C}^v)^{-1} \mathbf{W}^v,$$

$$\widehat{\mathbf{W}}^v = \frac{1}{2} (\widetilde{\mathbf{W}}^v + (\widetilde{\mathbf{W}}^v)^\top),$$

$$\mathbf{S}_{ij}^v = \widehat{\mathbf{W}}_{ij}^v / \sum_{k \in N_i} \widehat{\mathbf{W}}_{ik}^v, \text{ if } j \in N_i; \quad \text{else } 0.$$

We then run the SNF message-passing updates with diagonal regularization, which keeps self-affinity dominant while improving numerical stability. Initialize $\mathbf{P}_0^{(v)} = \widehat{\mathbf{W}}^v$. For $t = 0, \dots, T-1$:

$$\mathbf{P}_{t+1}^{(v)} = \mathcal{B}_\alpha \left(\mathbf{S}^{(v)} \left(\frac{1}{|\mathcal{V}|-1} \sum_{u \neq v} \mathbf{P}_t^{(u)} \right) \mathbf{S}^{(v)\top} \right)$$

$$\mathcal{B}_\alpha(\mathbf{X}) = \frac{1}{2} (\mathbf{X} + \mathbf{X}^\top) + \alpha \mathbf{I}$$

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

$$\mathbf{P} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbf{P}_T^{(v)}, \quad \mathbf{D}_{ii} = \sum_j \mathbf{P}_{ij}$$

$$\tilde{\mathbf{P}} = \mathbf{D}^{-1} \mathbf{P}, \quad \hat{\mathbf{P}} = \frac{1}{2} (\tilde{\mathbf{P}} + \tilde{\mathbf{P}}^\top + \mathbf{I}).$$

Lastly, to form a dendrogram for model typology, we cluster the fused affinity $\hat{\mathbf{P}}$ with hierarchical clustering using 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 Methods, which contrast within-family and across-family similarities. Results reported in this section are based on ImageNet representations;

qualitatively similar patterns are observed across additional datasets (Appendix C). Across metrics, we observe substantial variation in the degree to which procedural structure is recovered (Figures 2, C.1, C.2, C.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. 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 that are lost by arbitrary linear transformations.

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 transformation 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.

3.2. Integration achieves superior discrimination

Critically, our SNF approach, which integrates information across all representational dimensions, achieves dramatically superior family separation compared to any single metric. SNF attains a d' of 12.42—nearly three times higher than the best-performing single measure—and consistently outperforms all baselines across separation criteria. Importantly, as shown in Figure 2, SNF maintains high and balanced discrimination across nearly all family pairs. By contrast, individual metrics often exhibit uneven performance, separating some families while failing for others.

Averaging similarities across metrics does not resolve this limitation: simple means dilute complementary signals and retain conflicting noise. In contrast, SNF’s diffusion-based fusion reinforces consistent neighborhood structure across metrics while attenuating discordant components, yielding both stronger global separation and greater local stability.

This superior performance demonstrates that different representational dimensions provide complementary information about model families. While geometry and tuning capture family-specific computational strategies, and linearly accessible features potentially reflect more universal solutions, the integration of these diverse facets yields comprehensive signatures that most reliably distinguish model families.

To test whether SNF recovers the shared structure across metrics rather than replicating any single one, we try to quantify the inter-metric agreement. Specifically, to ensure

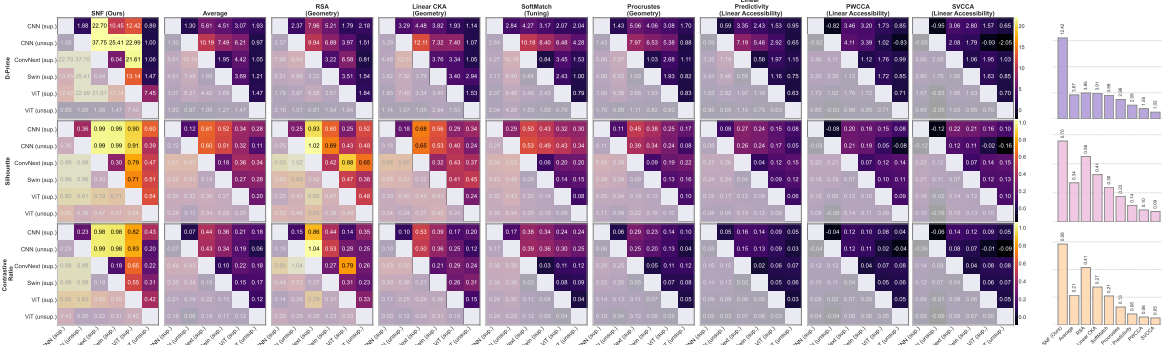


Figure 2. Model-family separability on ImageNet under d' , silhouette score and contrastive ratio. Columns correspond to nine similarity metrics, including two fusion-based methods (SNF, average) and seven commonly used representational metrics (Distinct aspects of representation emphasized by each metric are shown in the bracketed text). The barplot on the right side of each row shows the mean model-family separability. Scores are shown in their native scales and are not directly comparable across measures.

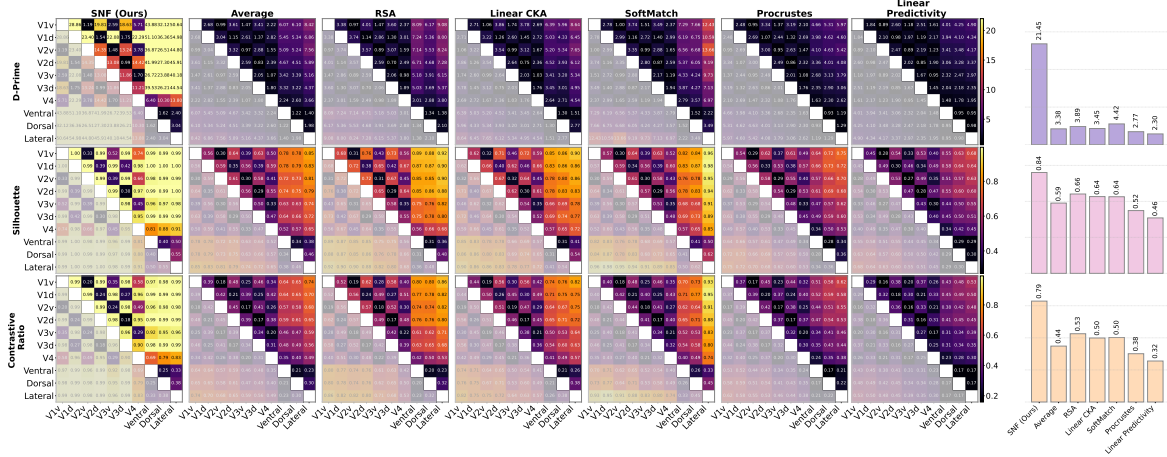


Figure 3. Similar to Figure 2 but on the NSD dataset with human responses and excluding SVCCA and PWCCA.

comparability across metrics, we symmetrize every similarity matrix by averaging it with its transpose, remove the diagonal (self-similarity), vectorize the remaining upper-triangle entries, and then compute correlation between vectors. As shown in Figure D.1, geometry-preserving metrics (RSA, SoftMatch, CKA) show strong mutual agreement, mapping-based metrics (Procrustes, Linear Predictivity) are moderately aligned, and CCA-variants highly correlate with each other but less agree with other metrics. SNF aligns only moderately with any single metric and is clearly distinct from simple averaging, indicating it fuses complementary facets instead of collapsing to one metric.

3.3. Which similarity metrics recover known anatomical–functional structure in visual cortex?

We next evaluate representational similarity metrics using neural data from the Natural Scenes Dataset (NSD), asking whether the patterns observed in artificial models generalize to biological visual systems. We compare representational

correspondences across cortical regions using responses to a shared set of natural images, focusing on cross-subject comparisons. 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 brain analyses, as their alignment values were systematically inflated by the large voxel dimensionality relative to the number of stimuli. To assess recovery of anatomical–functional structure, we quantify region discriminability using the separability measure d' , contrasting within-region similarity across subjects with across-region similarity. Across single metrics, we observe patterns consistent with the model results: metrics that preserve representational geometry or unit-level tuning better differentiate cortical regions, whereas metrics with more permissive mappings show weaker discrimination. Critically, integrating information across representational dimensions using Similarity Network Fusion (SNF) yields substantially stronger recovery of anatomical–functional structure than any in-

dividual metric. As shown in Figure 3, SNF achieves a mean d' of 21.45—nearly five times higher than the best-performing single measure—and consistently outperforms all baselines across evaluation criteria.

3.4. Data-driven typologies in models and brains

Having shown that integrating complementary representational similarity dimensions yields the most informative correspondence structure, we next examine the population-level organization revealed by the SNF-fused similarity matrices. Rather than evaluating individual pairwise comparisons, this analysis asks whether integrated similarity supports the discovery of coherent structure across populations of systems, artificial models and biological brain regions alike, without relying on prior assumptions about architecture, training paradigm, or anatomy.

3.4.1. POPULATION STRUCTURE IN VISION MODELS

As a baseline, clustering based on individual similarity metrics produces fragmented and inconsistent organization (Figure E.1, E.2, E.3, E.4, E.6, E.7 E.9 E.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. Together, these results underscore the limitations of single similarity metrics: each emphasizes a particular representational facet, leading to population structures that are incomplete or unstable.

In contrast, hierarchical clustering of the fused similarity matrix reveals well-defined groupings that both recover expected relationships and reveal additional organizational structure (Figure 4, E.5, E.8, E.11)). The fidelity of these clusterings to the underlying similarity structure is quantified using the cophenetic correlation coefficient (Figure F.1), which shows that the SNF-based organization more faithfully reflects pairwise similarities than individual metrics.

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 4), the most striking result is that training paradigm can override architectural boundaries. All self-supervised models, regardless of architecture, form a unified cluster that transcends the CNN–Transformer divide. Self-supervised ResNets group more closely with self-supervised ViTs than with their supervised architectural

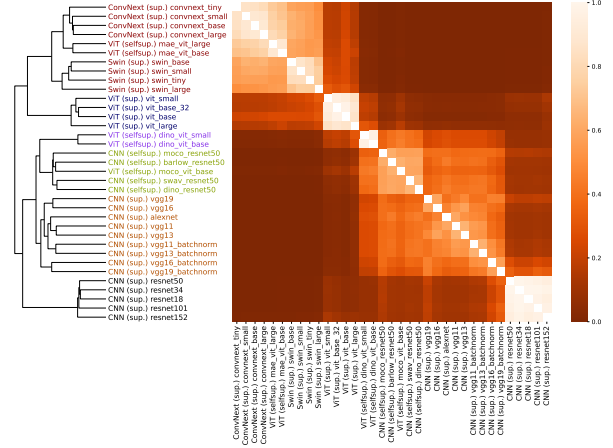


Figure 4. SNF-based clustering reveals that models naturally group by architecture and supervision regime. The heatmap shows the SNF-fused similarity matrix reordered by leaf ordering. Leaf labels are colored by the cluster (formed by SNF) they belong to.

counterparts, suggesting that self-supervised objectives may induce shared computational strategies that dominate over architectural differences. Perhaps most surprisingly, 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 all converge on similar representational structures. This convergence suggests that architectural modernization and masked reconstruction, though distinct in implementation, arrive at related computational solutions.

The typology reveals the “species” of vision models, defined not just by their surface characteristics but by how they fundamentally process and organize information.

3.4.2. POPULATION STRUCTURE IN VISUAL CORTEX

We observe a closely parallel pattern in neural data. Applying the same SNF-based clustering approach to representational similarities across visual cortical regions reveals coherent, cross-subject organization consistent with known anatomical and functional structure. Visualization using PCA shows that while individual metrics recover broad trends, the SNF-fused similarity matrix produces the clearest and most interpretable 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 5).

To quantify how well a representational similarity measure preserves the expected ventral-stream hierarchy, following Thobani et al. (2025), we compute a *hierarchy alignment score*. This score measures the rank corre-

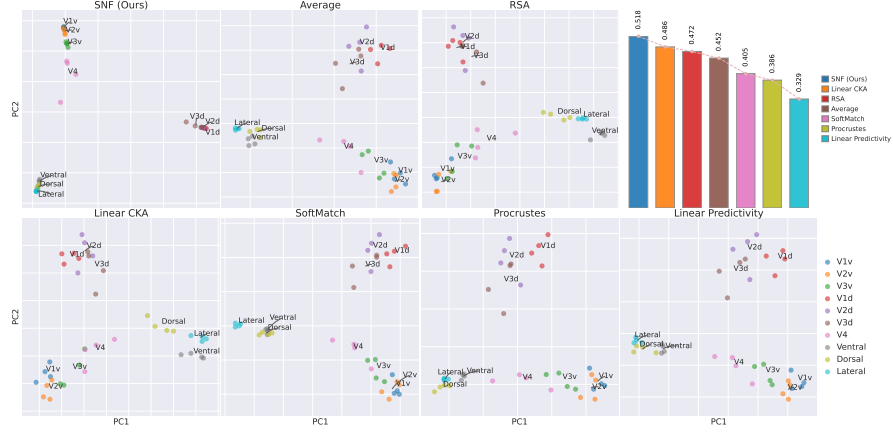


Figure 5. PCA visualization and hierarchical score comparison of cross-brain-region representational relationships in the ventral visual pathway of the NSD Dataset. Each panel corresponds to one metric, points represent brain region instances, and text labels indicate centroid positions. The top-right bar chart displays the raw hierarchical scores for each metric, with colors corresponding to metrics.

lation (Kendall’s τ) between representational similarity and approximated hierarchical distance between cortical regions. Each region r is assigned a discrete hierarchical level $L(r) \in \{1, \dots, 5\}$, as $L(V1v) = L(V1d) = 1$, $L(V2v) = L(V2d) = 2$, $L(V3v) = L(V3d) = 3$, $L(V4) = 4$, $L(\text{ventral_visual}) = 5$. Given a brain-region-by-brain-region similarity matrix $S \in \mathbb{R}^{R \times R}$ and the corresponding region labels $\{r_1, \dots, r_R\}$, we first construct a hierarchy-distance matrix, $H_{ij} = |L(r_i) - L(r_j)|$, $1 \leq i, j \leq R$. We then extract all off-diagonal entries of H and S , $\mathbf{h} = \{H_{ij} : i \neq j\}$, $\mathbf{s} = \{S_{ij} : i \neq j\}$, and define $\mathbf{y} = -\mathbf{s}$ so that smaller hierarchy distance corresponds to larger similarity. The hierarchy score is the Kendall’s τ -b rank correlation between $\mathbf{h}$ and $\mathbf{y}$: $\tau_{\text{hier}} = \tau_{\text{Kendall}}(\mathbf{h}, \mathbf{y})$. The results further support the superior performance of SNF in revealing known biological correspondence structure (Figure 5). The ability of individual similarity metrics to recover this hierarchical alignment mirrors the trends observed in the separability analyses: metrics that preserve representational geometry (e.g., RSA, linear CKA) achieve the strongest alignment with the ventral-stream hierarchy, while mapping-based metrics exhibit progressively weaker alignment as the allowable transformations become more flexible, with SoftMatch (which imposes the strictest correspondence) yielding the highest hierarchy alignment among mapping-based approaches.

4. Discussion

This work reframes representational comparison as a question of *what kinds of structure different similarity metrics are capable of recovering*. Rather than treating representational similarity as a single quantity, we systematically evaluated multiple correspondence dimensions (e.g. geometric organization, unit-level tuning, and linearly accessible information) and assessed how well each recovers known

structure in two domains: procedural differences among artificial vision models and anatomical–functional organization in human visual cortex. Across both domains, we find a consistent ordering of representational similarity measures. Metrics that preserve representational geometry or unit-level tuning (e.g., RSA, linear CKA, SoftMatch) most reliably recover known structure, distinguishing procedurally distinct model families and differentiating cortical regions across subjects. In contrast, more flexible mappings that emphasize linearly accessible information (e.g., Linear Predictivity and CCA-based methods) show substantially weaker discrimination. These results suggest that while linearly decodable signals may reflect more broadly shared constraints, finer-grained geometric and tuning structure carries stronger signatures of procedural and anatomical organization.

Another contribution of this work is to show that these representational dimensions are complementary rather than redundant. No single metric provides a complete characterization of representational correspondence. To address this, we adapt Similarity Network Fusion (SNF), originally developed for multi-omics integration, to combine similarity graphs derived from multiple metrics. The fused similarity consistently outperforms individual measures in recovering known structure in both models and brains. We note that SNF provides a stable global summary when different metrics capture complementary aspects of representation, but does not replace individual similarity measures. While the fused similarity is well-suited for identifying broad organizational structure, such as model families or cortical hierarchies, individual metrics are necessary to interpret which representational properties drive the observed similarity. More broadly, our results suggest that representational similarity measures should be interpreted in terms of the structure they are capable of recovering, rather than interchangeable proxies for representational correspondence.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

- Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., and Kay, K. A Massive 7T fMRI Dataset to Bridge Cognitive Neuroscience and Artificial Intelligence. 25(1):116–126. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL <https://doi.org/10.1038/s41593-021-00962-x>.
- Bo, Y., Soni, A., Srivastava, S., and Khosla, M. 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>.
- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. 2020.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
- Chen*, X., Xie*, S., and He, K. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.
- Cloos, N., Li, M., Siegel, M., Brincat, S. L., Miller, E. K., Yang, G. R., and Cueva, C. J. Differentiable optimization of similarity scores between models and brains. *arXiv preprint arXiv:2407.07059*, 2024.
- De Loera, J. A. and Kim, E. D. Combinatorics and geometry of transportation polytopes: An update. *Discrete geometry and algebraic combinatorics*, 625:37–76, 2013.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- Ding, F., Denain, J.-S., and Steinhart, J. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34: 1556–1568, 2021.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. 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>.
- Echtermeyer, C., Han, C. E., Rotarska-Jagiela, A., Mohr, H., Uhlhaas, P. J., and Kaiser, M. Integrating temporal and spatial scales: human structural network motifs across age and region of interest size. *Frontiers in neuroinformatics*, 5:10, 2011.
- Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. pp. 770–778. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.
- Khosla, M. and Williams, A. H. 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.
- Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMIR, 2019.
- Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience. 2. 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/full>.
- Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. URL https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

- Letwin, N. E., Kafkafi, N., Benjamini, Y., Mayo, C., Frank, B. C., Luu, T., Lee, N. H., and Elmer, G. I. Combined application of behavior genetics and microarray analysis to identify regional expression themes and gene–behavior associations. *Journal of Neuroscience*, 26(20):5277–5287, 2006.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A ConvNet for the 2020s, 2022. URL <https://arxiv.org/abs/2201.03545>.
- maintainers, T. and contributors. Torchvision: Pytorch’s computer vision library. <https://github.com/pytorch/vision>, 2016.
- Markello, R. snfpy: Similarity Network Fusion in Python, 2020. URL <https://github.com/rmarkello/snfpy>.
- Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. 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.
- Morcos, A., Raghu, M., and Bengio, S. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.
- Mukamel, E. A. and Ngai, J. Perspectives on defining cell types in the brain. *Current opinion in neurobiology*, 56: 61–68, 2019.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.
- Rousseeuw, P. J. 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](https://doi.org/10.1016/0377-0427(87)90125-7). URL <https://www.sciencedirect.com/science/article/pii/0377042787901257>.
- Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. URL <http://arxiv.org/abs/1409.1556>.
- Soni, A., Srivastava, S., Kording, K., and Khosla, M. Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, pp. 2024–08, 2024.
- Thobani, I., Sagastuy-Brena, J., Nayebi, A., Prince, J., Cao, R., and Yamins, D. Model-brain comparison using inter-animal transforms, 2025. URL <https://arxiv.org/abs/2510.02523>.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3):261–272, 2020.
- Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.
- Wightman, R. PyTorch Image Models, year = 2019, publisher = GitHub, journal = GitHub repository, doi = 10.5281/zenodo.4414861, howpublished = <https://github.com/rwightman/pytorch-image-models>.
- Williams, A. H. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. *bioRxiv*, pp. 2024–10, 2024.
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. 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 Tabel 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
ECOSSET (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](#)), 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.

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.](#)), VGG-11/13/16/19 (with/without batch normalization) ([Simonyan & Zisserman](#)), and ResNet-18/34/50/101/152 ([He et al.](#)).

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. Separation Performance on Other Datasets

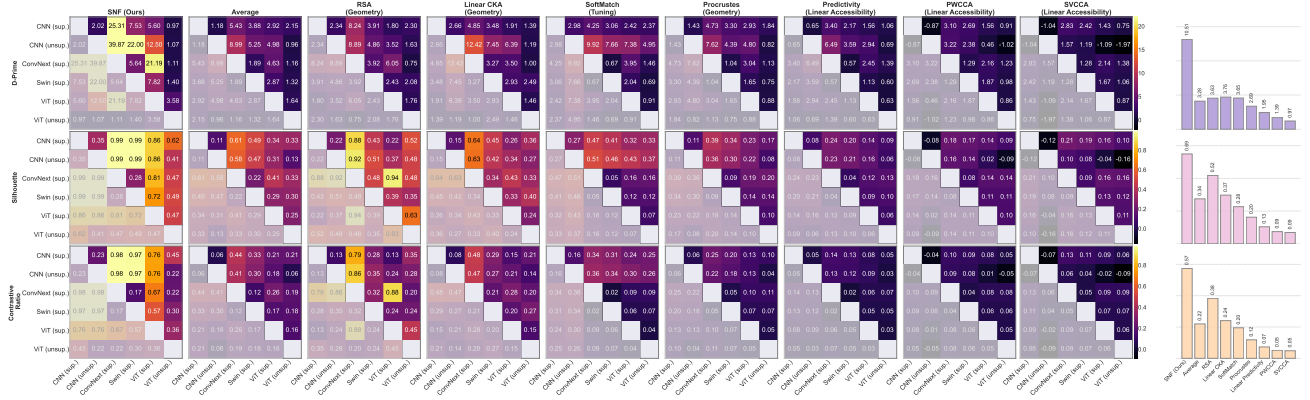


Figure C.1. Same as Figure 2, but using EcoSet instead of ImageNet.

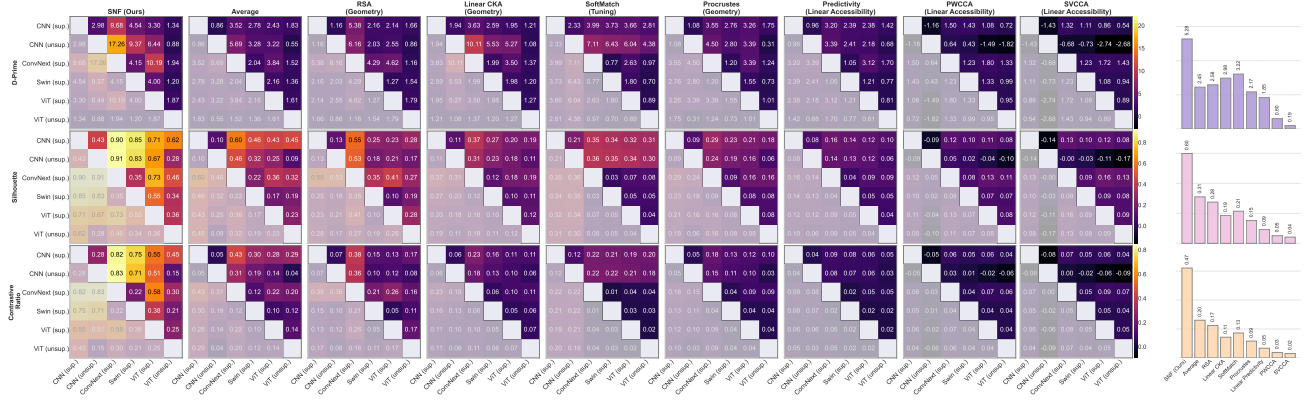


Figure C.2. Same as Figure 2, but using CIFAR10 instead of ImageNet.

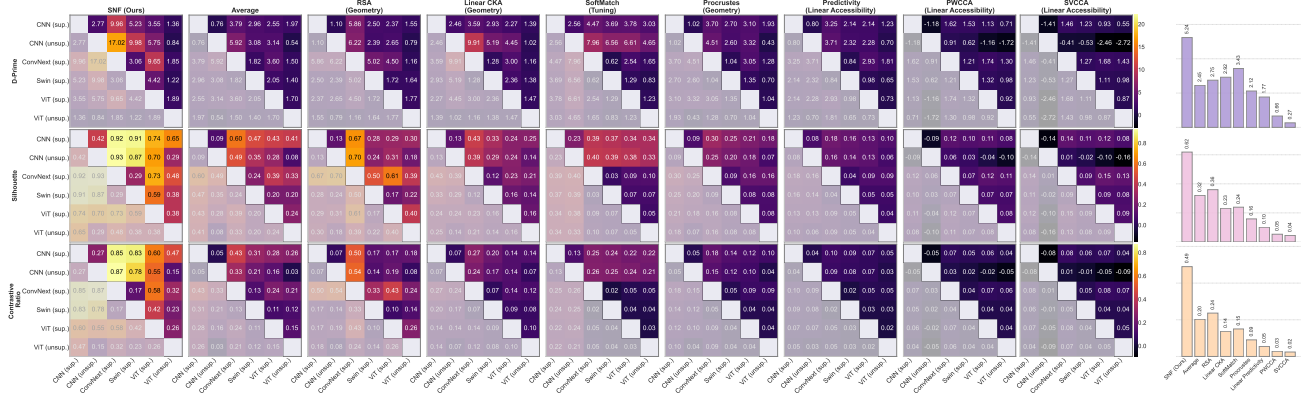


Figure C.3. Same as Figure 2, but using CIFAR100 instead of ImageNet.

D. Metrics' Similarity Scores Consistency

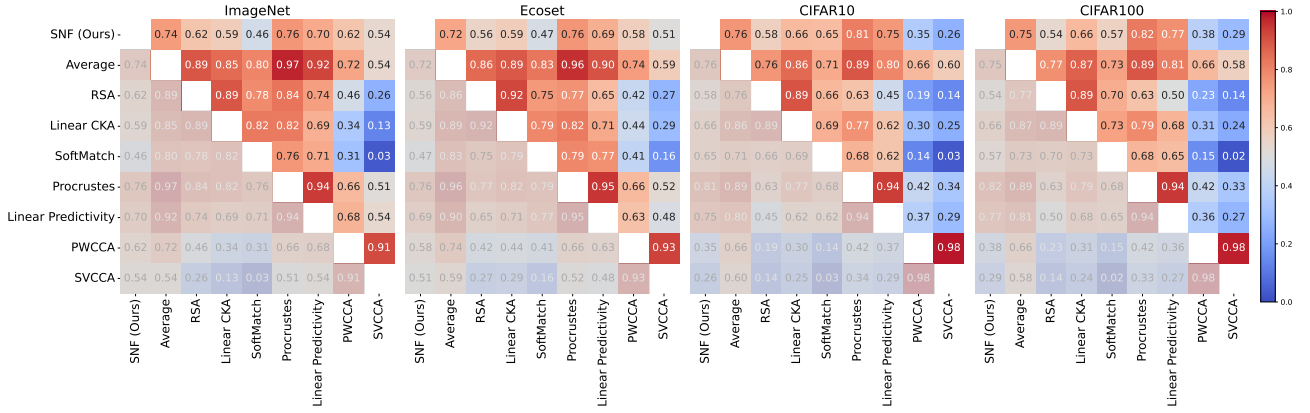
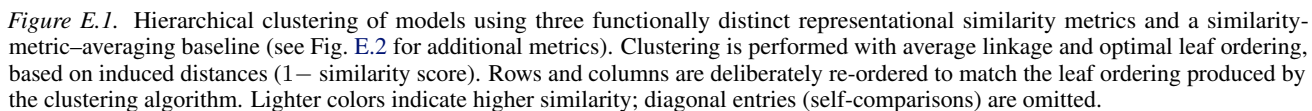


Figure D.1. Each subplot (ImageNet, Ecoset, CIFAR10, CIFAR100) shows pairwise Pearson correlations between the vectorized upper-triangle entries of the symmetrized model-model similarity matrices produced by nine metrics. Higher values indicate that two metrics have more similar relational geometry among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



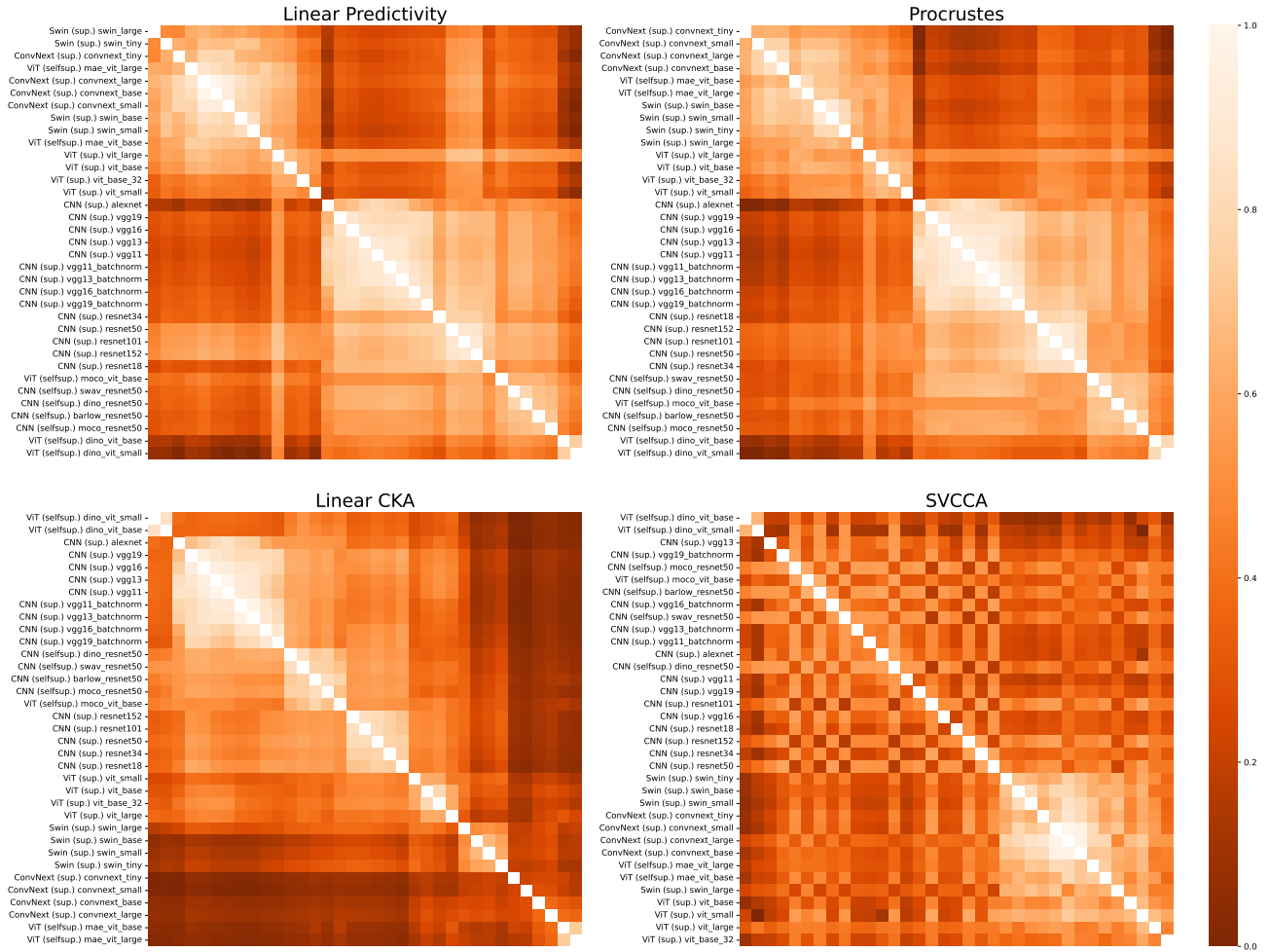


Figure E.2. Same as Figure E.1, but for the other 4 metrics.

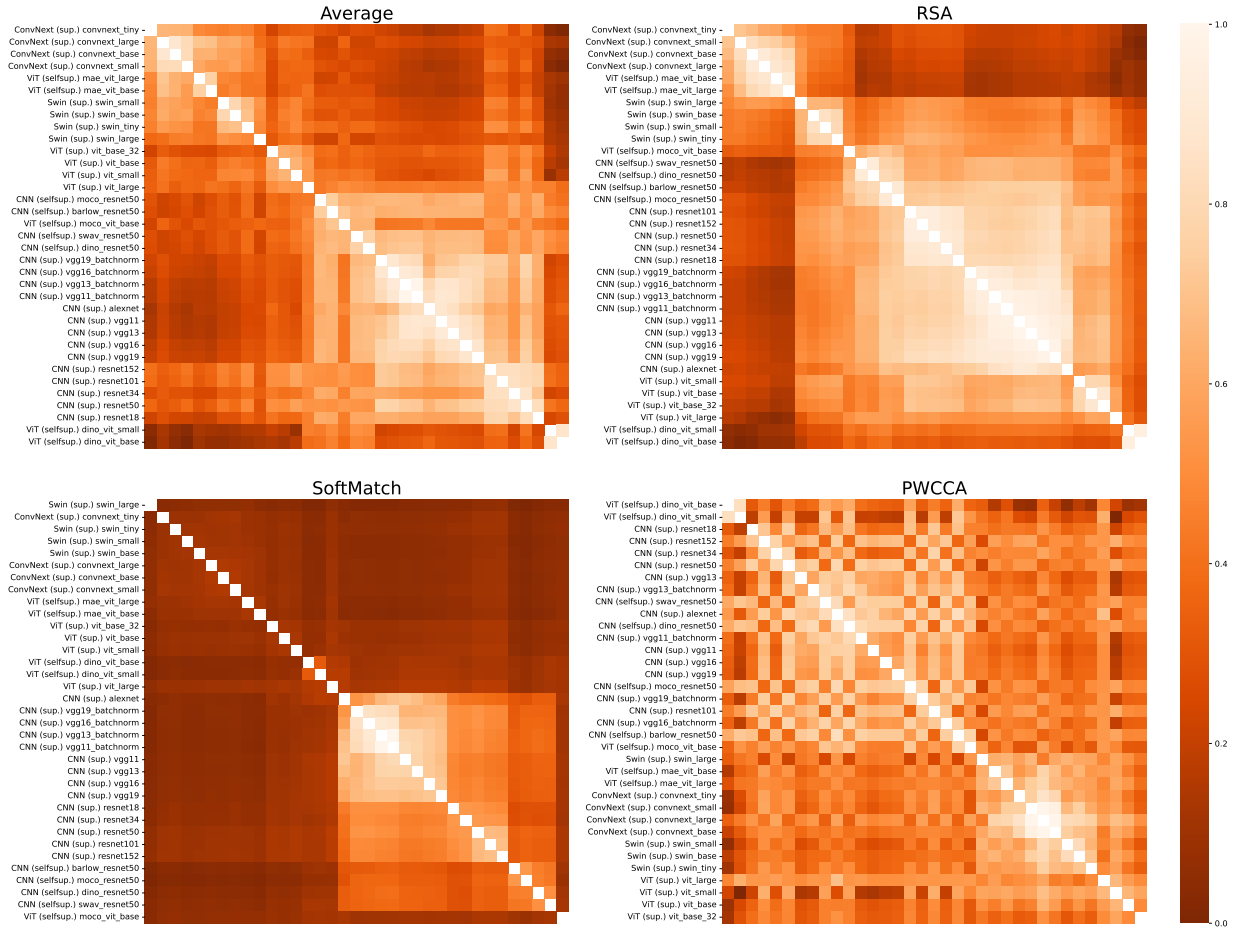


Figure E.3. Same as Figure E.1, but using Ecoset instead of ImageNet.

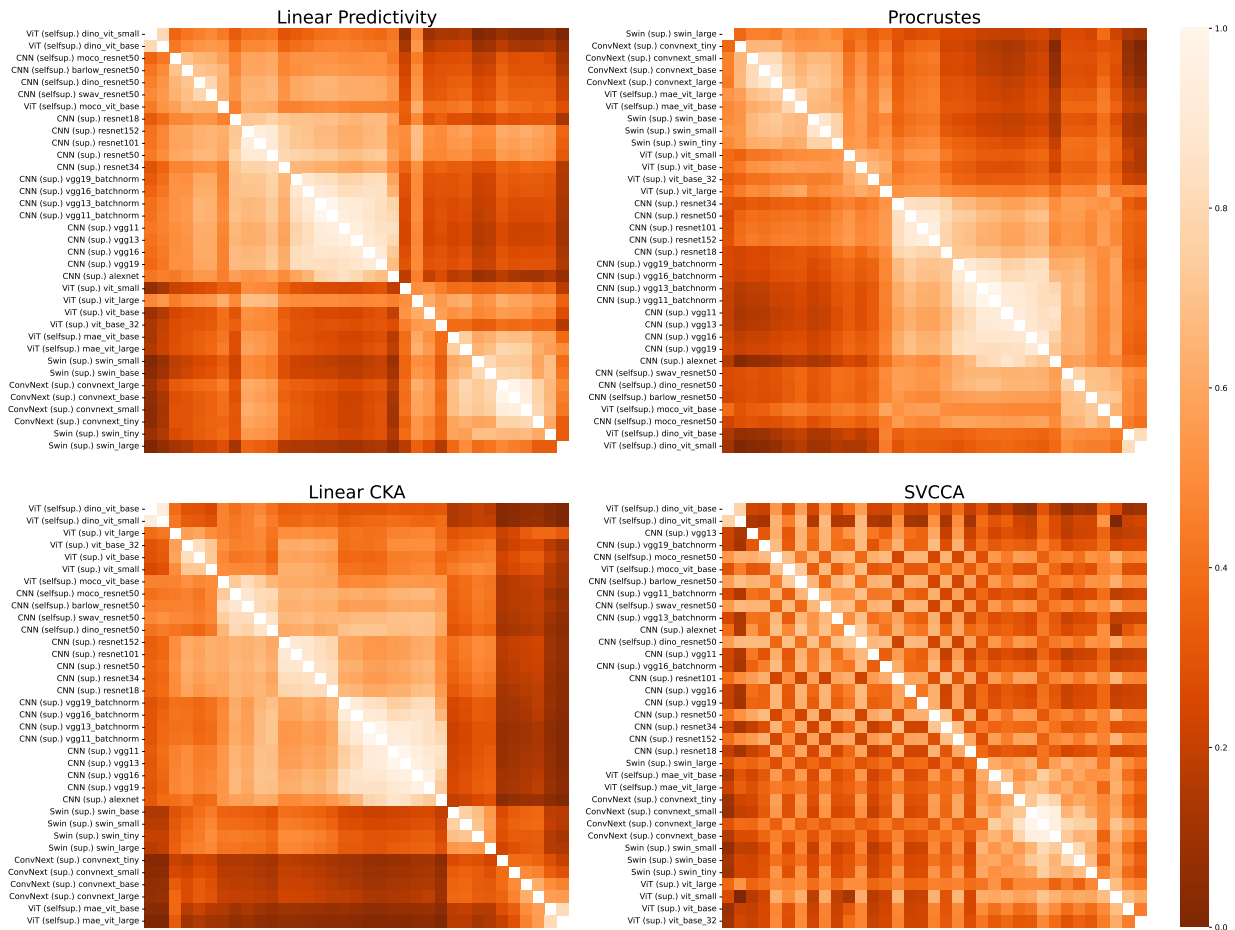


Figure E.4. Same as Figure E.2, but using Ecocost instead of ImageNet.

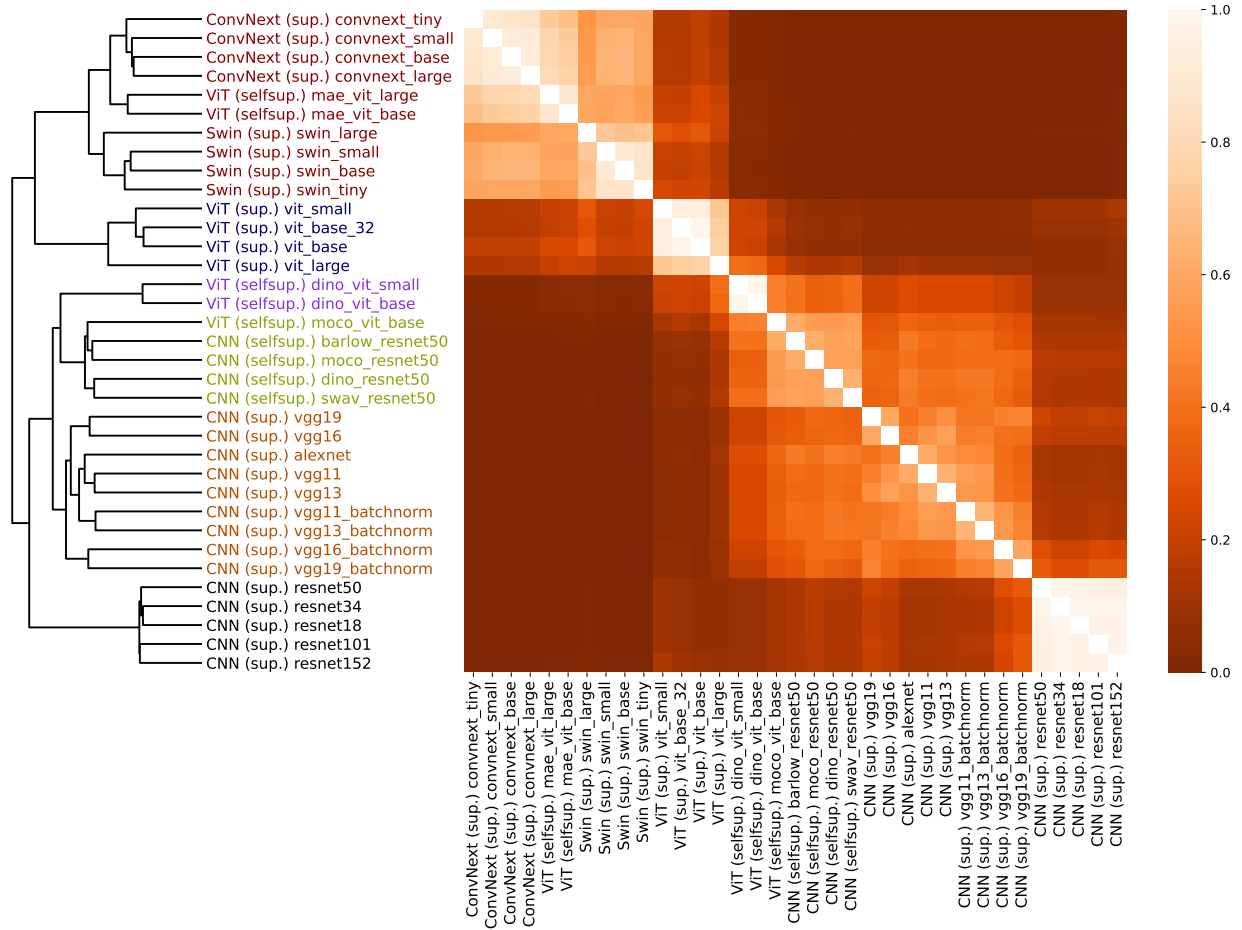


Figure E.5. Same as Figure 4, but using Ecoset instead of ImageNet.

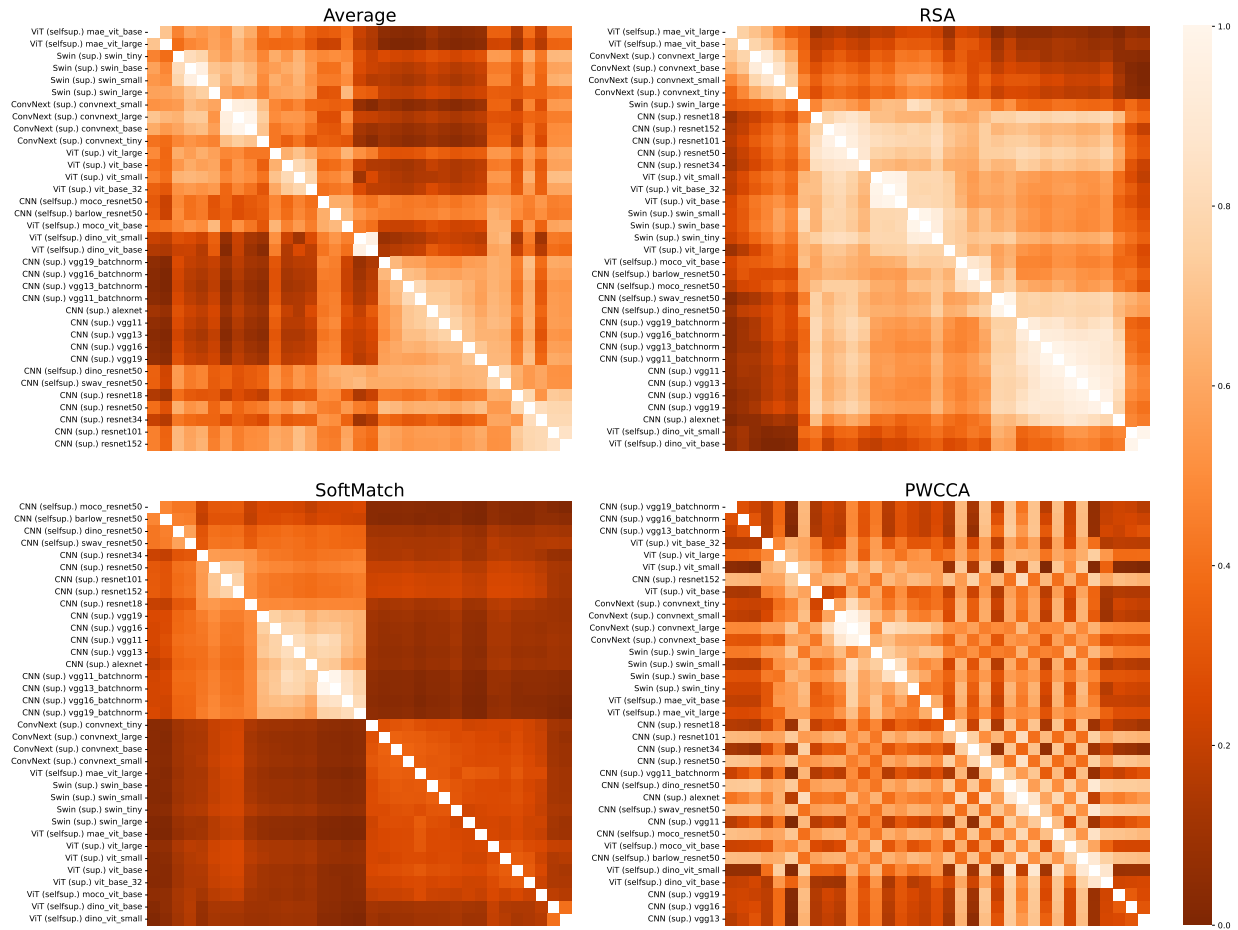


Figure E.6. Same as Figure E.1, but using CIFAR10 instead of ImageNet.

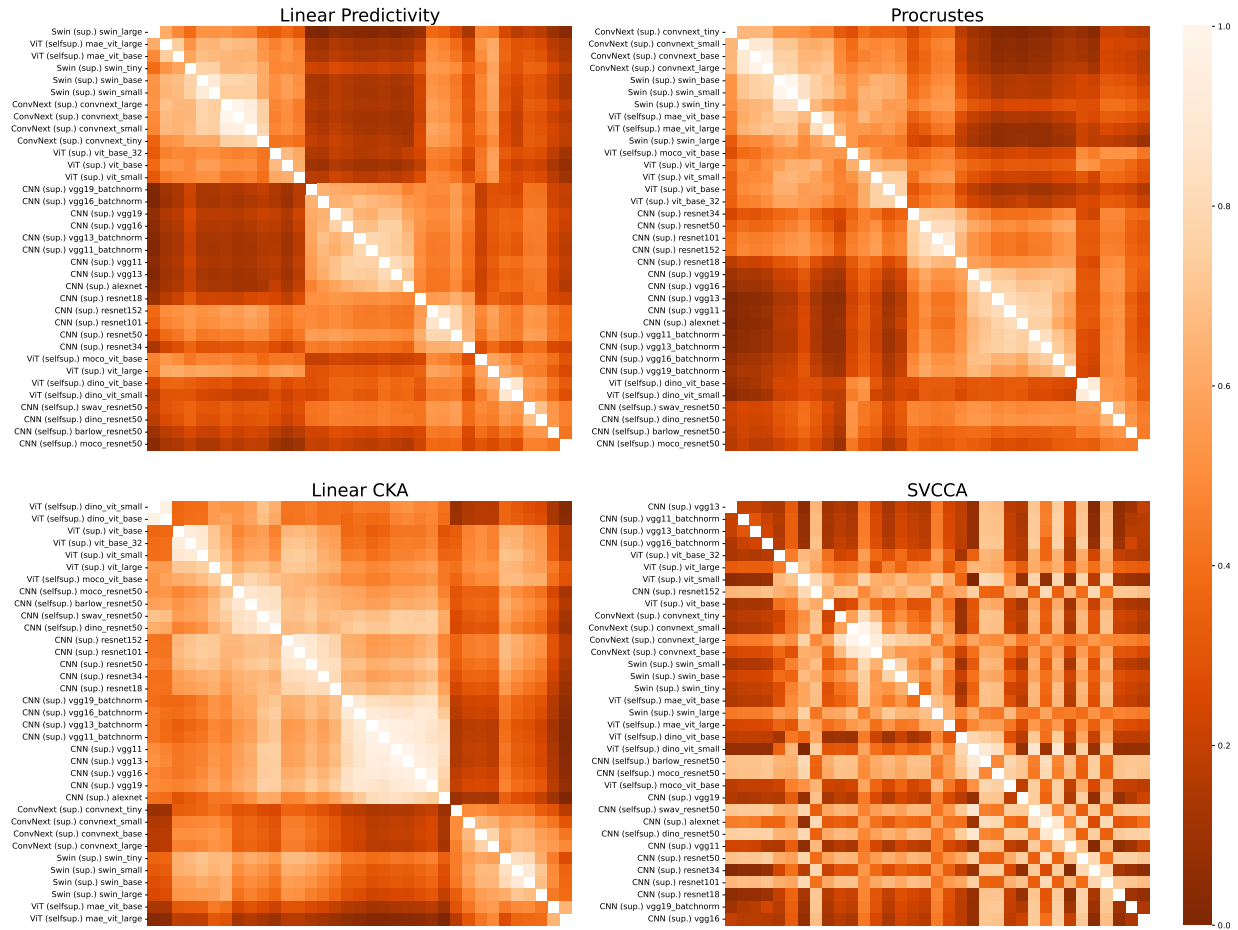


Figure E.7. Same as Figure E.2, but using CIFAR10 instead of ImageNet.

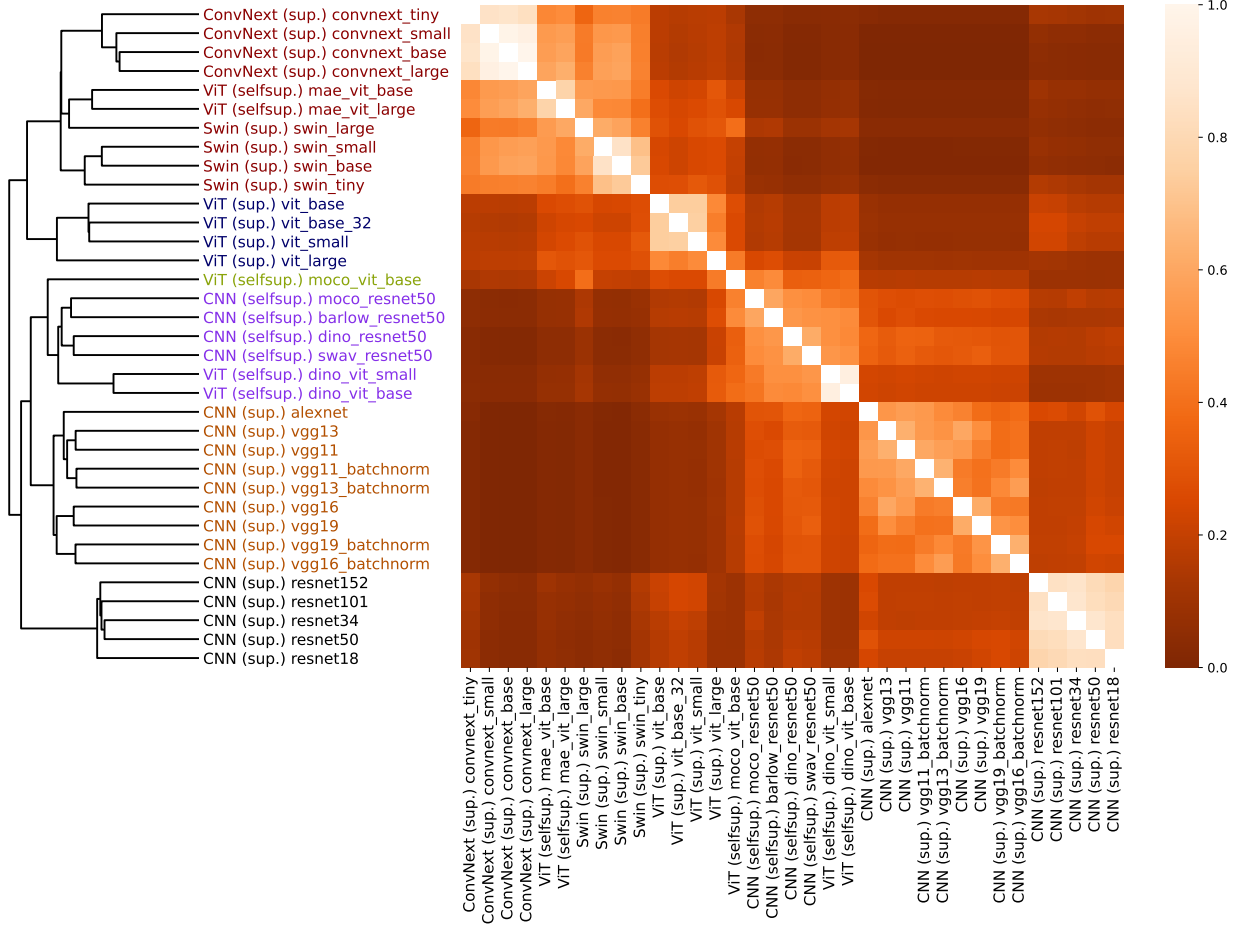


Figure E.8. Compared to ImageNet in Figure 4, the clustering result is a little different but similar, demonstrating the SNF results and metric’s separability are also influenced by datasets, while also preserving a certain degree of stability. The supervised models are clustered in the same way, but the clusters for the self-supervised models changed, demonstrating that the self-supervised way leads to a special but kind of unifying representational geometry.

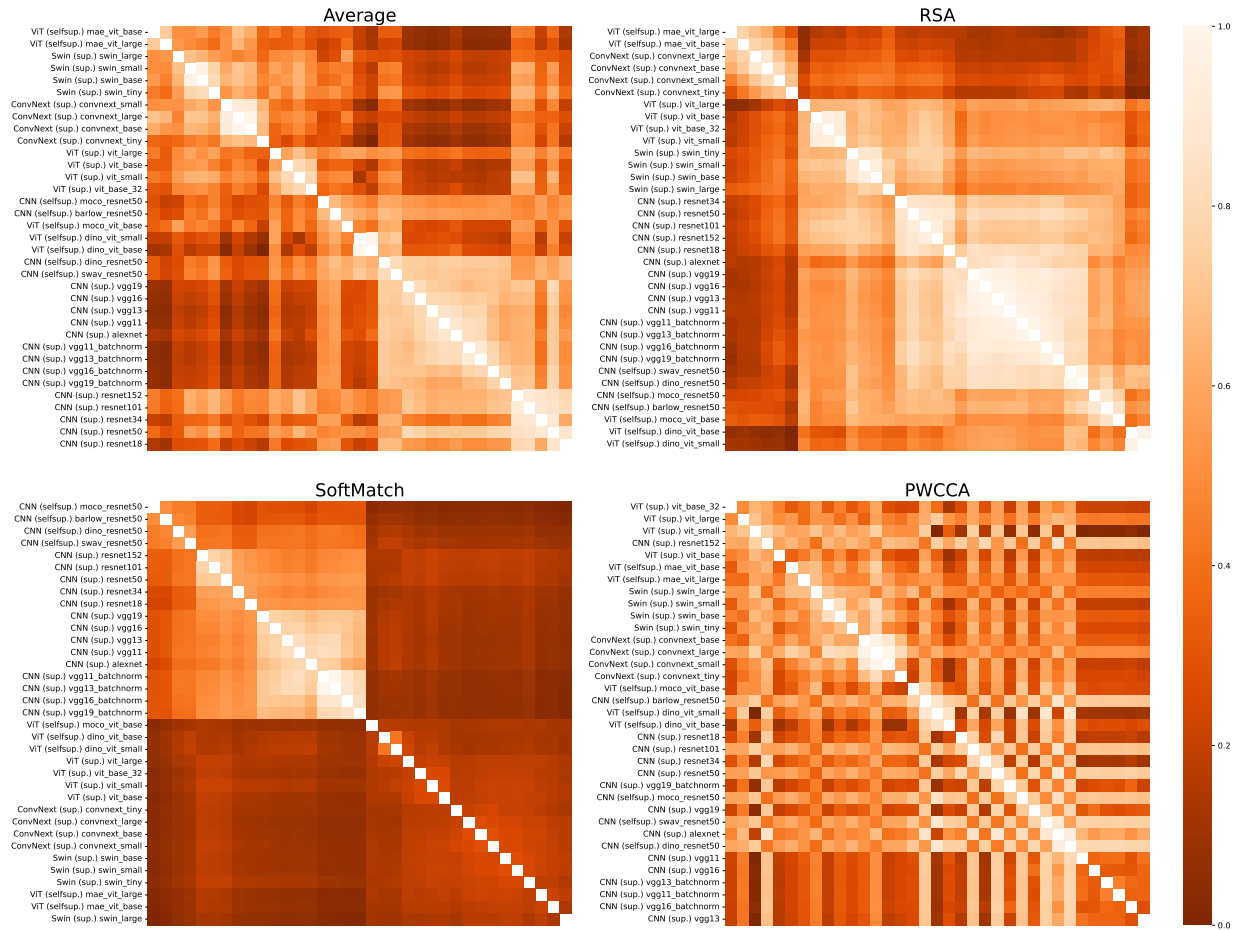


Figure E.9. Same as Figure E.1, but using CIFAR100 instead of ImageNet.

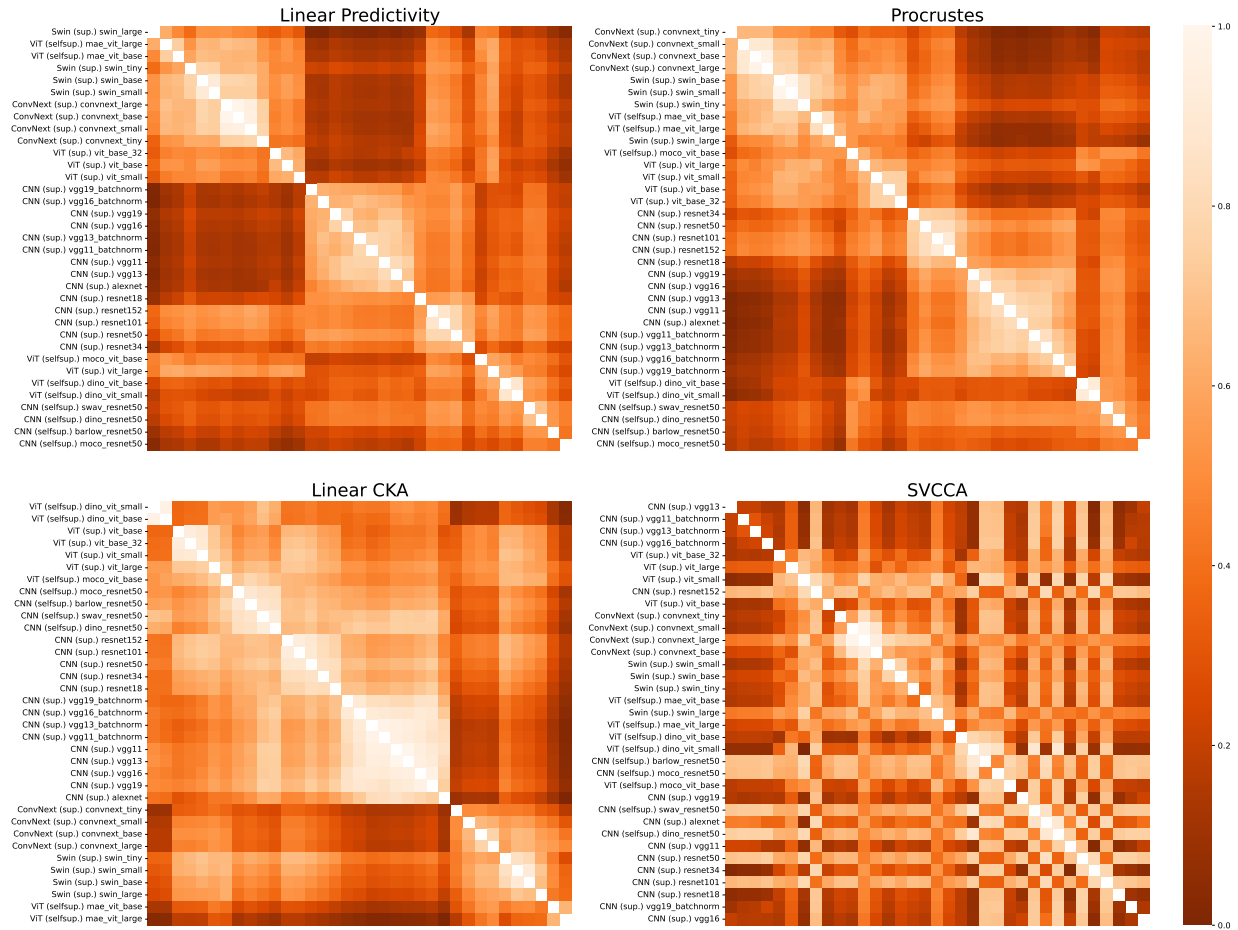


Figure E.10. Same as Figure E.2, but using CIFAR100 instead of ImageNet.

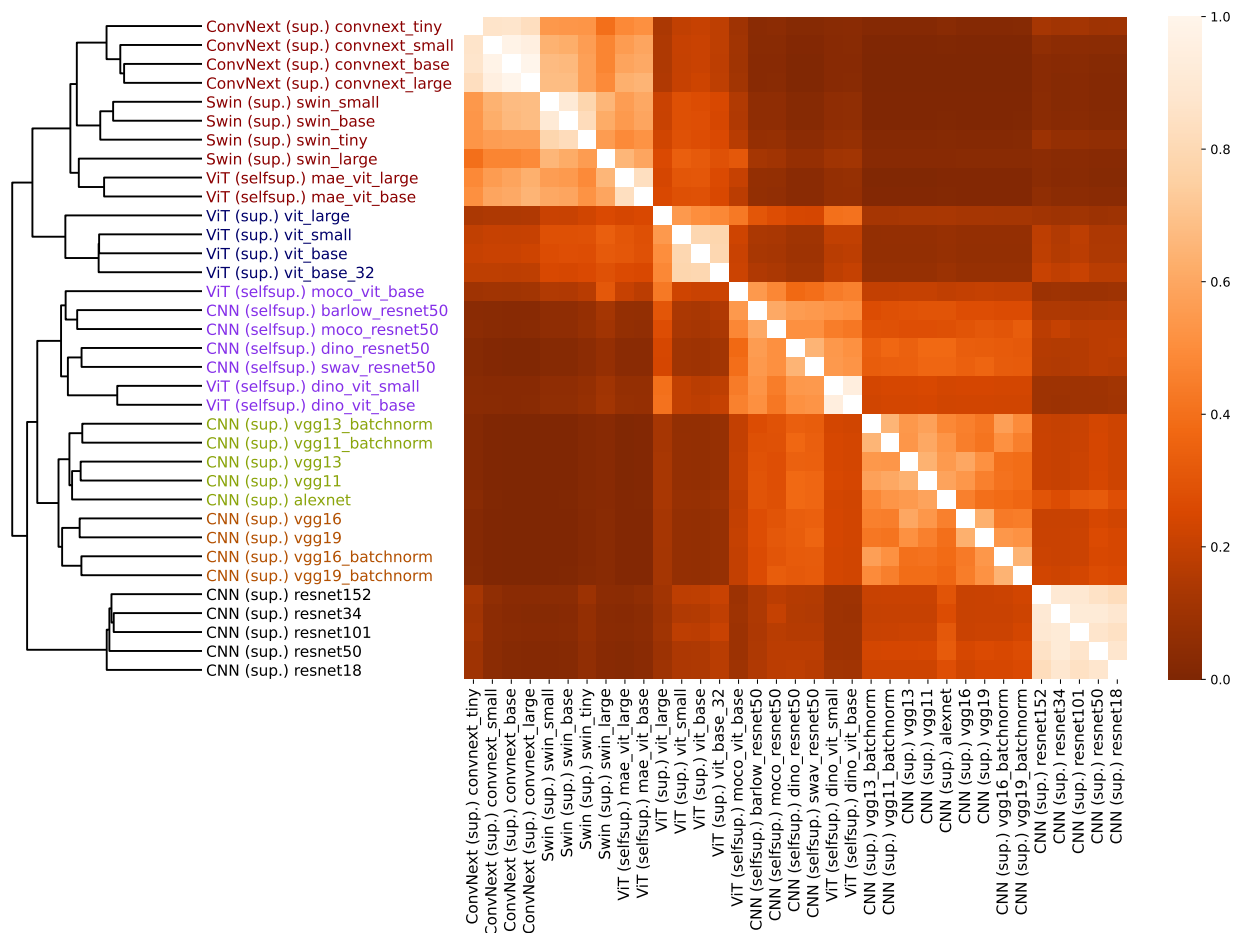


Figure E.11. Same as Figure E.8, but using CIFAR100 instead of CIFAR10.

F. Cophenetic Correlation Coefficients for Clustering

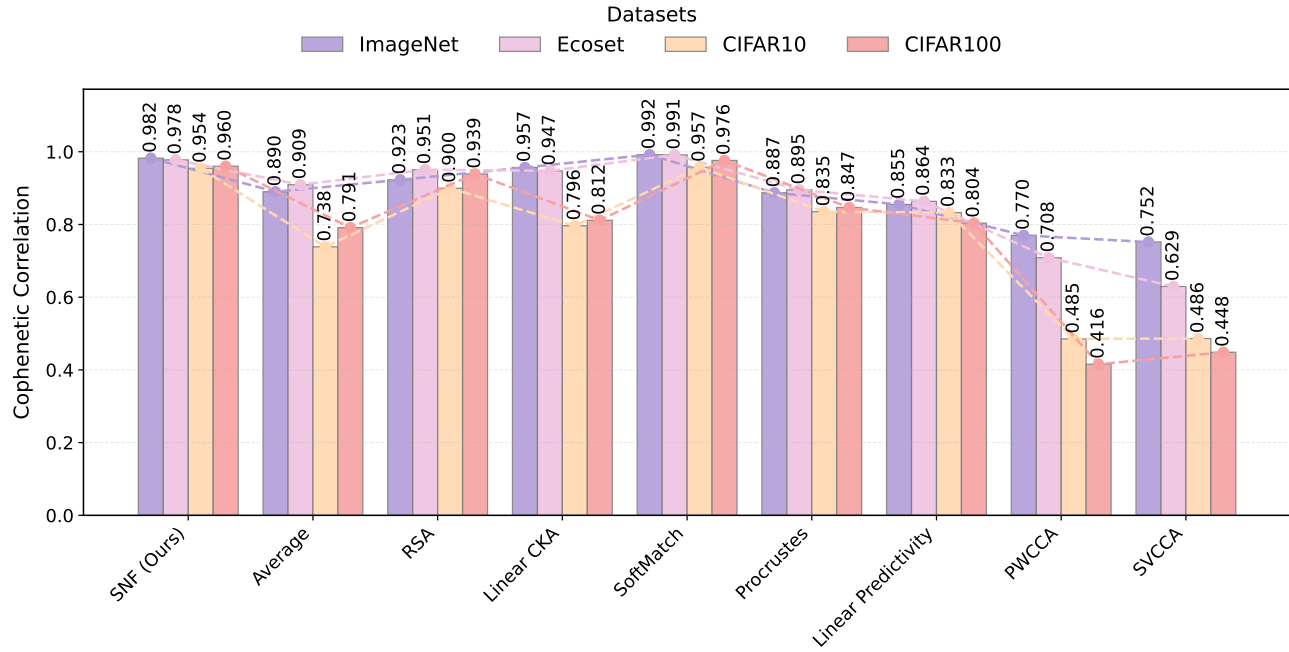


Figure F.1. Cophenetic correlation coefficients (CCC) for hierarchical clusterings induced by each metric on four datasets. Higher CCC (closer to 1) means the clustering more faithfully preserves the original pairwise structure. Bars show CCC on ImageNet, Ecoset, CIFAR10, and CIFAR100 (values annotated;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

G. Metrics' Cross Layer Consistency

For each metric and dataset, we also test whether the similarity structure is stable across depth. For CNNs, if batch normalization layers exist (e.g. ResNet), we count layers by them; if not (e.g., VGG, AlexNet), we count layers by ReLU units. For ViTs, one feature extraction unit (first layer normalization + attention block + second layer normalization) is counted as one layer, and we take the output after the second layer normalization. We select the layer for a normalized depth $d \in (0, 1]$ via $\ell = \lfloor dL \rfloor$, where L is the total number of layers. For each metric, one depth corresponds to one matrix. We vectorized the off-diagonal entries, and computed Pearson correlations between the 3 depth pairs. We found that SNF and SoftMatch show the highest depth consistency, whereas the CKA and CCA variants are less stable. Although the representations in different layers could be greatly different, the SNF could still identify the layer-model belonging relationship and the difference between model families.

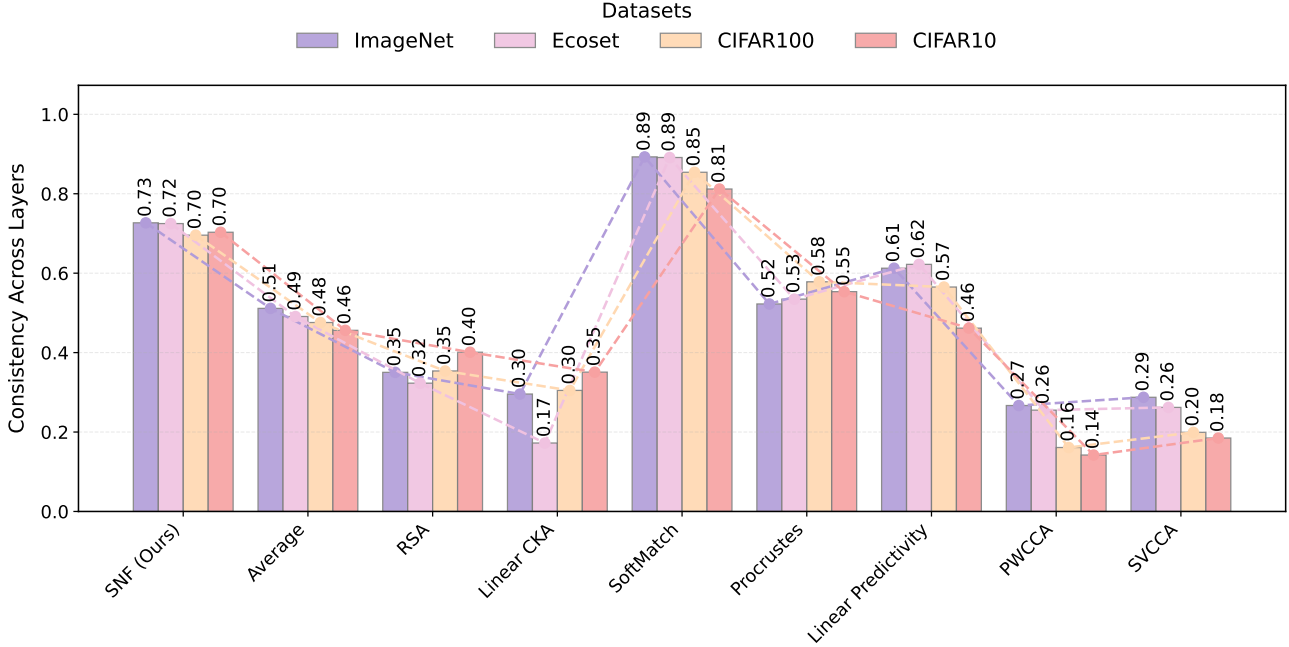


Figure G.1. Cross-layer consistency of inter-model similarity. The bar height is the mean of the three depth pairs correlations; higher values indicate greater consistency.

H. The Usage of Large Language Models (LLMs)

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.