Generalized Shape Metrics on Neural Representations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Understanding the operation of biological and artificial networks remains a difficult and important challenge. To identify general principles of neural computation, researchers are increasingly interested in surveying large collections of networks that are trained on, or biologically adapted to, similar tasks. A standardized set of analysis tools is now needed to identify how network-level covariates—such as architecture, learning algorithm, anatomical brain region, model organism, etc.—systematically impact neural representations (hidden layer activations). Here, we provide a rigorous foundation for these analyses by defining a broad family of metric spaces that quantify representational dissimilarity between networks. Using this framework, we modify existing measures based on canonical correlation analysis and centered kernel alignment to satisfy the triangle inequality, formulate a novel metric that respects the inductive biases in convolutional layers, and identify approximate Euclidean embeddings that enable network representations to be incorporated into essentially any off-the-shelf machine learning method. We demonstrate these methods on large-scale datasets from biology (Allen Institute Brain Observatory) and deep learning (NAS-Bench-101), and identify relationships between neural representations that are interpretable in terms of anatomical hierarchies and model performance.

1 Introduction

The extent to which different deep networks or neurobiological systems use equivalent representations in support of similar task demands is a topic of persistent interest in machine learning and neuroscience [1]. Several methods including linear regression [2, 3], canonical correlation analysis (CCA; [4, 5]), representational similarity analysis (RSA; [6]), and centered kernel alignment (CKA; [7, 8]) have been used to quantify the similarity of hidden layer activation patterns. These similarity measures are often interpreted on an ordinal scale and are employed to compare a limited number of networks—e.g., they can indicate whether networks A and B are more or less similar than networks A and C. While these comparisons have yielded important insights [2–7, 9–11], the underlying methodologies have not been extended to quantitative analyses spanning thousands of networks.

To unify existing approaches and enable more sophisticated analyses, we draw on ideas from statistical shape analysis [12–14] to develop dissimilarity measures that are proper metrics—i.e., are symmetric and respect the triangle inequality—on neural representations. This enables several off-the-shelf methods with theoretical guarantees for classification (e.g. k-nearest neighbors, [15]) and clustering (e.g. hierarchical clustering [16]) on network representations. Existing similarity measures can violate the triangle inequality, which complicates these downstream analyses [17–19]. However, by introducing simple modifications to these methods, we can satisfy the triangle inequality and view existing approaches as special cases of the framework we outline. This framework also facilitates the discovery of new metrics, including metrics specialized to convolutional layer representations.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

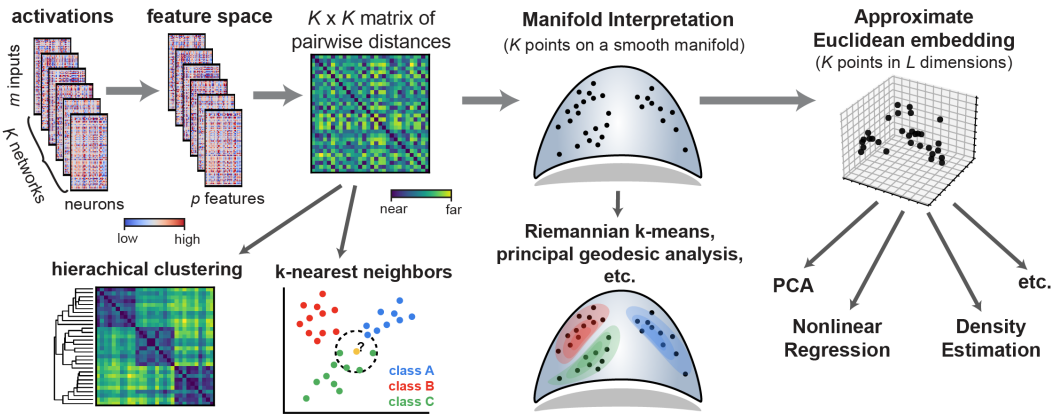

Figure 1: Machine learning workflows enabled by generalized shape metrics.

Moreover, we show empirically that neural representations can be embedded with low distortion into Euclidean spaces with a moderate number of dimensions. This enables an even broader variety of modeling approaches for regression, classification, and unsupervised analysis. Applying this approach to representations of visual inputs in mice (Allen Brain Observatory; [20]), we predict an anatomically derived feature of brain regions ("hierarchy score") based on features derived from representational similarity. We also apply this approach to analyze hidden layer representations in a database of 432K deep networks (NAS-Bench-101; [21]), and find a surprising degree of correlation between early layer and deep layer representations. One consequence of this is that we can predict a network's test set accuracy using only embedded representations of early layers.

Overall, we provide a theoretical grounding which explains why existing similarity measures are useful (they are often close to metric spaces, and can be modified to fulfill metric space axioms precisely), draw new connections between these approaches and other established research areas [14, 22], utilize this framework to propose novel metrics, and demonstrate a practical and general-purpose machine learning workflow that scales to datasets containing thousands of networks.

2 Methods

This section outlines a workflow (Fig. 1) to analyze representations across large collections of networks. After briefly summarizing prior approaches (sec. 2.1), we cover background material on metric spaces and discuss their theoretical advantages over existing dissimilarity measures (sec. 2.2). We then present a class of metrics that capture these advantages (sec. 2.3) and cover a special case that is suited to convolutional layers (sec. 2.4). We conclude by developing even stronger geometric insights (sec. 2.6). First, we note that certain metrics correspond to geodesic distances on shape manifolds [23], enabling manifold optimization and learning algorithms to be applied to neural representations. Moreover, we can use multidimensional scaling methods [24] to find Euclidean embeddings that approximate this metric space, which we demonstrate empirically in Section 3.

2.1 Prior work and problem setup

Neural network representations are often summarized over a set of m reference inputs (e.g. test set images). Let $\boldsymbol{X}_i \in \mathbb{R}^{m \times n_i}$ and $\boldsymbol{X}_j \in \mathbb{R}^{m \times n_j}$ denote the responses of two networks (with n_i and n_j neurons, respectively) to a collection of these inputs. Quantifying the similarity between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is complicated by the fact that, while the m inputs are the same, there is no direct correspondence between the neurons. Even if $n_i = n_j$, the typical Frobenius inner product, $\langle \boldsymbol{X}_i, \boldsymbol{X}_j \rangle = \mathrm{Tr}[\boldsymbol{X}_i^\top \boldsymbol{X}_j]$, and metric, $\|\boldsymbol{X}_i - \boldsymbol{X}_j\| = \langle \boldsymbol{X}_i - \boldsymbol{X}_j, \boldsymbol{X}_i - \boldsymbol{X}_j \rangle^{1/2}$, fail to capture the desired notion of dissimilarity. For instance, let $\boldsymbol{\Pi}$ denote some $n \times n$ permutation matrix and let $\boldsymbol{X}_i = \boldsymbol{X}_j \boldsymbol{\Pi}$. Intuitively, we should consider $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ to be identical in this case since the ordering of neurons is arbitrary. Yet, clearly $\|\boldsymbol{X}_i - \boldsymbol{X}_j\| \neq 0$, except in very special cases.

Representational similarity analysis (RSA; [6]) addresses this problem by computing the dissimilarity between network responses for each pair of inputs, resulting in an $m \times m$ representational dissimilarity

matrix (RDM) for each network. Then, RSA uses the Spearman rank correlation between the RDMs from two networks to assess their similarity. A simpler approach is to use linear regression to predict the responses of one network given the responses of another, using the coefficient of determination (R^2) as a measure of similarity [2, 3]. In machine learning, canonical correlation analysis (CCA; [4, 5]) and centered kernel alignment (CKA; [7, 8]) have been used to quantify the similarity of hidden layer activation patterns. These procedures allow for various linear and nonlinear transformations of representations, but without a theoretical framework it is unclear how to choose among them, use their outputs for downstream tasks, or generalize them to new domains.

2.2 Feature space mapping, metrics, and equivalence relations

Our first contribution will be to establish formal notions of distance (metrics) between neural representations. To accommodate the common scenario when the number of neurons varies across networks (i.e. when $n_i \neq n_j$), we must first map the representations into a common feature space so that a single distance function may be defined between them. For each set of representations, $\boldsymbol{X}_i$, we suppose there is a mapping into a p-dimensional feature space, $\boldsymbol{X}_i \mapsto \boldsymbol{X}_i^\phi$, where $\boldsymbol{X}_i^\phi \in \mathbb{R}^{m \times p}$. In the special case where all networks have equal size, $n_1 = n_2 = \ldots = n$, we can express the feature mapping as a single function $\phi : \mathbb{R}^{m \times n} \mapsto \mathbb{R}^{m \times p}$, so that $\boldsymbol{X}_i^\phi = \phi(\boldsymbol{X}_i)$. When networks have dissimilar sizes, we can reduce the representations down to some common dimension $p \leq \min(n_1, \ldots, n_K)$ using, for example, PCA [4]. Alternatively, we can zero pad representations up to $p = \max(n_1, \ldots, n_K)$ dimensions, which preserves the geometry of all network representations.

Next, we seek to establish *metrics* within the feature space, which are distance functions that satisfy:

$$\text{Equivalence:} \quad d(\boldsymbol{X}_i^\phi, \boldsymbol{X}_j^\phi) = 0 \iff \boldsymbol{X}_i^\phi \sim \boldsymbol{X}_j^\phi \tag{1}$$

$$\text{Symmetry:} \quad d(\boldsymbol{X}_i^\phi, \boldsymbol{X}_j^\phi) = d(\boldsymbol{X}_j^\phi, \boldsymbol{X}_i^\phi) \tag{2}$$

$$\text{Triangle Inequality:} \quad d(\boldsymbol{X}_i^\phi, \boldsymbol{X}_j^\phi) \leq d(\boldsymbol{X}_i^\phi, \boldsymbol{X}_k^\phi) + d(\boldsymbol{X}_k^\phi, \boldsymbol{X}_j^\phi) \tag{3}$$

for all $\boldsymbol{X}_i^\phi$, $\boldsymbol{X}_j^\phi$, and $\boldsymbol{X}_k^\phi$ in the feature space. The symbol '$\sim$' denotes an *equivalence relation* between two elements. That is, the expression $\boldsymbol{X}_i^\phi \sim \boldsymbol{X}_j^\phi$ means that "$\boldsymbol{X}_i^\phi$ is equivalent to $\boldsymbol{X}_j^\phi$." Formally, distance functions satisfying Eqs. (1) to (3) define a metric over a quotient space defined by the equivalence relation (see Supplement A). Intuitively, by specifying different equivalence relations we can account for symmetries in network representations, such as permutations over arbitrarily labeled neurons (other options are discussed below in sec. 2.3).

Metrics quantify dissimilarity in a way that agrees with our intuitive notion of distance. For example, Eq. (2) ensures that the distance from $\boldsymbol{X}_i^\phi$ to $\boldsymbol{X}_j^\phi$ is the same as the distance from $\boldsymbol{X}_j^\phi$ to $\boldsymbol{X}_i^\phi$. Linear regression is an approach that violates this condition—the similarity measured by R^2 depends on which network is treated as the dependent variable. Further, Eq. (3) ensures that distances are self-consistent in the sense that if two elements ($\boldsymbol{X}_i^\phi$ and $\boldsymbol{X}_j^\phi$) are close to a third ($\boldsymbol{X}_k^\phi$), then they are necessarily close to each other. Existing measures based on CCA, RSA, and CKA, are symmetric, but do not satisfy the triangle inequality. By modifying these approaches, we avoid potential pitfalls and can leverage theoretical guarantees on learning in proper metric spaces [15–19].

2.3 Generalized shape metrics and group invariance

In this section, we outline a new framework to quantify representational dissimilarity, which leverages a well-developed mathematical literature on *shape spaces* [12–14]. The key idea is to treat $\boldsymbol{X}_i^\phi \sim \boldsymbol{X}_j^\phi$ if and only if there exists a linear transformation $\boldsymbol{T}$ within a set of allowable transformations $\mathcal{G}$, such that $\boldsymbol{X}_i^\phi = \boldsymbol{X}_j^\phi \boldsymbol{T}$. Much of shape analysis literature focuses on the special case where $\mathcal{G}$ is the special orthogonal group $\mathcal{SO}(p) = \{\boldsymbol{R} \in \mathbb{R}^{p \times p} \mid \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}, \det(\boldsymbol{R}) = 1\}$, meaning that $\boldsymbol{X}_i^\phi$ and $\boldsymbol{X}_j^\phi$ are equivalent if there is a p-dimensional rotation (without reflection) that relates them. Standard shape analysis further considers each $\boldsymbol{X}_i^\phi$ to be a mean-centered ($(\boldsymbol{X}_i^\phi)^\top \boldsymbol{1} = \boldsymbol{0}$) and normalized ($\|\boldsymbol{X}_i^\phi\|_F = 1$) version of the raw landmark locations held in $\boldsymbol{X}_i \in \mathbb{R}^{m \times p}$ (an assumption that we will relax). Defining $\mathbb{S}^{m \times p}$ to be the hypersphere of $m \times p$ matrices with unit Frobenius norm, we see that $\boldsymbol{X}_i^\phi \in \mathbb{S}^{m \times p}$. In this context, $\boldsymbol{X}_i^\phi$ is called a "pre-shape." By removing rotations from a pre-shape, $[\boldsymbol{X}_i^\phi] = \{\boldsymbol{S} \in \mathbb{S}^{m \times p} \mid \boldsymbol{S} \sim \boldsymbol{X}_i^\phi\}$ for pre-shape $\boldsymbol{X}_i^\phi$, we recover its "shape."

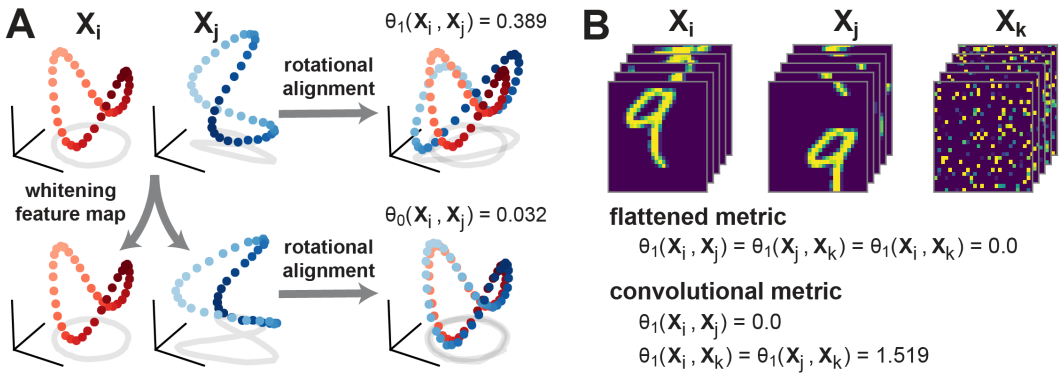

Figure 2: *(A)* Schematic illustration of metrics with rotational invariance (top), and linear invariance (bottom). Red and blue dots represent a pair of network representations $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, which correspond to m points in n-dimensional space. *(B)* Demonstration of convolutional metric on toy data. Flattened metrics (e.g. [4, 7]) that ignore convolutional layer structure treat permuted images ($\boldsymbol{X}_k$, right) as equivalent to images with coherent spatial structure ($\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, left and middle). A convolutional metric, Eq. (11), distinguishes between these cases while still treating $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ as equivalent (obeying translation invariance).

To quantify dissimilarity in network representations, we generalize this notion of shape to include other feature mappings (beyond mean-centering and normalization) and alignment operations (beyond rotations without reflection). The minimal distance within the feature space, after optimizing over alignments, defines a metric under suitable conditions (Fig. 2A). This results in a broad variety of *generalized shape metrics* (see also, ch. 18 of [14]), which fall into two categories as formalized by the pair of propositions below.

Proposition 1. *Let $\boldsymbol{X}_i^\phi \in \mathbb{R}^{m\times p}$, and let $\mathcal{G}$ be a group of linear isometries on $\mathbb{R}^{m\times p}$. Then,*

$$d(\boldsymbol{X}_i^\phi, \boldsymbol{X}_j^\phi) = \min_{\boldsymbol{T}\in\mathcal{G}} \|\boldsymbol{X}_i^\phi - \boldsymbol{X}_j^\phi \boldsymbol{T}\| \tag{4}$$

defines a metric, where $\boldsymbol{X}_i^\phi \sim \boldsymbol{X}_j^\phi$ if and only if there is a $\boldsymbol{T} \in \mathcal{G}$ such that $\boldsymbol{X}_i^\phi = \boldsymbol{X}_j^\phi \boldsymbol{T}$.

Proposition 2. *Let $\boldsymbol{X}_i^\phi \in \mathbb{S}^{m\times p}$, and let $\mathcal{G}$ be a group of linear isometries on $\mathbb{S}^{m\times p}$. Then,*

$$\theta(\boldsymbol{X}_i^\phi, \boldsymbol{X}_j^\phi) = \min_{\boldsymbol{T}\in\mathcal{G}} \arccos\left\langle \boldsymbol{X}_i^\phi, \boldsymbol{X}_j^\phi \boldsymbol{T}\right\rangle \tag{5}$$

defines a metric, where $\boldsymbol{X}_i^\phi \sim \boldsymbol{X}_j^\phi$ if and only if there is a $\boldsymbol{T} \in \mathcal{G}$ such that $\boldsymbol{X}_i^\phi = \boldsymbol{X}_j^\phi \boldsymbol{T}$.

Proofs are provided in Supplement B. The essential difference between d and θ is that the former uses the Euclidean metric to measure distance in the feature space, while the latter uses angular distance (geodesic distance on a sphere). The former typically enables easier computations, but the latter has appealing theoretical properties, including a close connection to CCA that we outline below.

Two key conditions appear in these propositions. First, $\mathcal{G}$ must be a *group* of functions. This means $\mathcal{G}$ is as a set that contains the identity function, is closed under composition ($\boldsymbol{T}_1\boldsymbol{T}_2 \in \mathcal{G}$ for any $\boldsymbol{T}_1 \in \mathcal{G}$ and $\boldsymbol{T}_2 \in \mathcal{G}$), and whose elements are invertible by other members of the set (if $\boldsymbol{T} \in \mathcal{G}$ then $\boldsymbol{T}^{-1} \in \mathcal{G}$). Second, every $\boldsymbol{T} \in \mathcal{G}$ must be an *isometry*, meaning that $\|\boldsymbol{X}_i^\phi - \boldsymbol{X}_j^\phi\| = \|\boldsymbol{X}_i^\phi \boldsymbol{T} - \boldsymbol{X}_j^\phi \boldsymbol{T}\|$ for all $\boldsymbol{T} \in \mathcal{G}$ and all elements of the feature space. On $\mathbb{R}^{m\times p}$ and $\mathbb{S}^{m\times p}$, all linear isometries are orthogonal transformations. Further, the set of orthogonal transformations, $\mathcal{O}(p) = \{\boldsymbol{Q} \in \mathbb{R}^{p\times p} : \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}\}$, defines a well-known group. Thus, the condition that $\mathcal{G}$ is a group of isometries is equivalent to $\mathcal{G}$ being a subgroup of $\mathcal{O}(p)$—i.e., a subset of $\mathcal{O}(p)$ satisfying the group axioms.

Intuitively, the condition that $\mathcal{G}$ is a group ensures that the alignment procedure is symmetric—i.e. it is equivalent to consider transforming $\boldsymbol{X}_i^\phi$ to match $\boldsymbol{X}_j^\phi$, or vice versa. Further, the condition that each $\boldsymbol{T} \in \mathcal{G}$ is an isometry ensures that the properties of the underlying metric space (Euclidean or angular distance) are preserved.

These propositions define a broad class of metrics as we enumerate below. For simplicity, we assume that $n_i = n_j = n$ in the examples below, with the understanding that a PCA or zero-padding preprocessing step has been performed in the case of dissimilar network sizes.

Permutation invariance The most stringent notion of representational similarity is to demand that neurons are one-to-one matched across networks. If we set $\boldsymbol{X}_i^{\phi} = \boldsymbol{X}_i$ for all i, then:

$$d_{\mathcal{P}}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \min_{\boldsymbol{\Pi} \in \mathcal{P}} \|\boldsymbol{X}_i - \boldsymbol{X}_j \boldsymbol{\Pi}\| \tag{6}$$

defines a metric by Proposition 1, since the set of permutation matrices, $\mathcal{P}(n)$ is a subgroup of $\mathcal{O}(n)$. To evaluate this metric we must optimize over the set of neuron permutations to align the two networks. This can be reformulated (see Supplement C) as a linear assignment problem, which is a fundamental problem in combinatorial optimization (see [25] for review). Exploiting an algorithm due to Jonker and Volgenant [26, 27] we can solve this problem in $O(n^3)$ time. The overall runtime for evaluating Eq. (6) is $O(mn^2 + n^3)$, since we must evaluate $\boldsymbol{X}_i^{\top} \boldsymbol{X}_j$ to formulate the assignment problem.

Rotation scaling invariance Let $\boldsymbol{C} = \boldsymbol{I}_m - (1/m)\mathbf{1}\mathbf{1}^{\top}$ denote an $m \times m$ *centering matrix*, and consider the feature mapping which mean-centers the columns, $\boldsymbol{X}_i^{\phi} = \boldsymbol{C}\boldsymbol{X}_i$. Then,

$$d_1(\boldsymbol{X}_i^{\phi}, \boldsymbol{X}_j^{\phi}) = \min_{\boldsymbol{Q} \in \mathcal{O}} \|\boldsymbol{X}_i^{\phi_1} - \boldsymbol{X}_j^{\phi_1} \boldsymbol{Q}\| \tag{7}$$

defines a metric by Proposition 1, and is equivalent to the *Procrustes size-and-shape distance* with reflections [14]. Further,

$$\theta_1(\boldsymbol{X}_i^{\phi}, \boldsymbol{X}_j^{\phi}) = \min_{\boldsymbol{Q} \in \mathcal{O}} \ \arccos \frac{\langle \boldsymbol{X}_i^{\phi_1}, \boldsymbol{X}_j^{\phi_1} \boldsymbol{Q} \rangle}{\|\boldsymbol{X}_i^{\phi_1}\| \|\boldsymbol{X}_j^{\phi_1}\|} \tag{8}$$

defines a metric by Proposition 2, and is equivalent to the Riemannian distance on Kendall's shape space [14]. To evaluate Eqs. (7) and (8), we must optimize over the set of orthogonal matrices to find the best alignment. This also maps onto a fundamental optimization problem known as the *orthogonal Procrustes problem* [28, 29], which can be solved in closed form in $O(n^3)$ time. As in the permutation-invariant metric described above, the overall runtime is $O(mn^2 + n^3)$.

Linear invariance Consider a partial whitening transformation, parameterized by $0 \leq \alpha \leq 1$:

$$\boldsymbol{X}^{\phi_\alpha} = \boldsymbol{C}\boldsymbol{X}(\alpha \boldsymbol{I} + (1-\alpha)(\boldsymbol{X}^{\top}\boldsymbol{C}\boldsymbol{X})^{-1/2}) \tag{9}$$

Note that $\boldsymbol{X}^{\top}\boldsymbol{C}\boldsymbol{X}$ is the empirical covariance matrix of $\boldsymbol{X}$. Thus, when $\alpha = 0$, Eq. (9) corresponds to ZCA whitening [30], which intuitively removes invertible linear transformations from the representations. When $\alpha = 1$, Eq. (9) reduces to the mean-centering feature map used above.

This feature space leads to a metric that is related to CCA. First, let $1 \geq \rho_1 \geq \ldots \geq \rho_n \geq 0$ denote the singular values of $(\boldsymbol{X}_i^{\phi_\alpha})^{\top}(\boldsymbol{X}_j^{\phi_\alpha})/\|\boldsymbol{X}_i^{\phi_\alpha}\|\|\boldsymbol{X}_j^{\phi_\alpha}\|$. Using orthogonal alignments, we see:

$$\theta_\alpha(\boldsymbol{X}_i^{\phi}, \boldsymbol{X}_j^{\phi}) = \min_{\boldsymbol{Q} \in \mathcal{O}} \ \arccos \frac{\langle \boldsymbol{X}_i^{\phi_\alpha}, \boldsymbol{X}_j^{\phi_\alpha} \boldsymbol{Q} \rangle}{\|\boldsymbol{X}_i^{\phi_\alpha}\| \|\boldsymbol{X}_j^{\phi_\alpha}\|} = \arccos(\tfrac{1}{n} \textstyle\sum_\ell \rho_\ell) \tag{10}$$

defines a metric by Proposition 2. The values $\rho_1, \ldots, \rho_n$ correspond to canonical correlation coefficients when $\alpha = 0$ [31] and ridge regularized canonical correlations when $\alpha > 0$ [32] (see Supplement C for further notes). Past works have suggested to use the average canonical correlation as a similarity measure [4, 5]. This heuristic corresponds to $\cos \theta_0(\boldsymbol{X}_i, \boldsymbol{X}_j)$, but applying $\arccos$ is needed to satisfy the triangle inequality. Since the covariance is often ill-conditioned or singular in practice, setting $\alpha > 0$ to regularize the calculation is typically necessary.

Nonlinear invariances We discuss feature maps that enable nonlinear notions of equivalence, and which relate to kernel CCA [33], in Supplement C.

2.4 Metrics for convolutional layers

In deep networks for image processing, each convolutional layer produces a $h \times w \times c$ array of activations, whose axes respectively correspond to image height, image width, and channels (number of convolutional filters). If stride-1 circular convolutions are used, then applying a circular

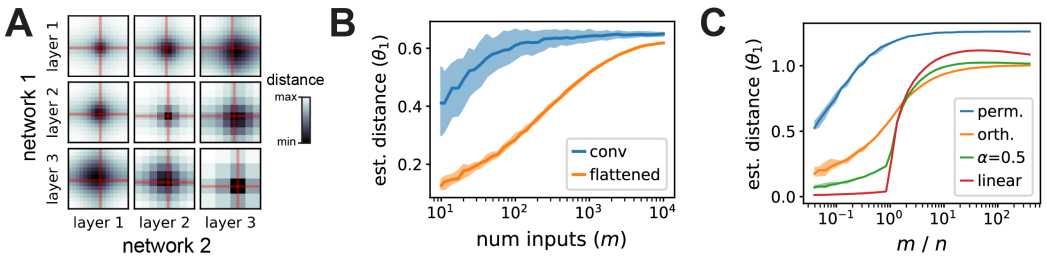

Figure 3: *(A)* Each heatmap shows a brute-force search over the shift parameters along the width and height dimensions of a pair of convolutional layers compared across two networks. The optimal shifts are typically close to zero (red lines). *(B)* Impact of sample size, m, on flattened and convolutional metrics with orthogonal invariance. The convolutional metric approaches its final value faster than the flattened metric, which is still increasing even at the full size of the CIFAR-10 test set ($m = 10^4$). *(C)* Impact of sample density, m/n, on metrics invariant to permutation, orthogonal, regularized linear ($\alpha = 0.5$), and linear transformations. Shaded regions mark the 10$^{\text{th}}$ and 90$^{\text{th}}$ percentiles across shuffled repeats. Further details for all simulations are provided in Supplement E.

shift along either spatial dimension produces the same shift in the layer's output. It is natural to reflect this property, known as translation equivariance [22], in the equivalence relation on layer representations. Supposing that the feature map preserves the shape of the activation tensor, we have $\boldsymbol{X}_k^\phi \in \mathbb{R}^{m \times h \times w \times c}$ for $k \in 1, \ldots, K$. Letting $\mathcal{S}(n)$ denote the group of n-dimensional circular shifts (a subgroup of the permutation group) and '$\otimes$' denote the Kronecker product, we propose:

$$\boldsymbol{X}_i^\phi \sim \boldsymbol{X}_j^\phi \iff \text{vec}(\boldsymbol{X}_i^\phi) = (\boldsymbol{I} \otimes \boldsymbol{S}_1 \otimes \boldsymbol{S}_2 \otimes \boldsymbol{Q})\text{vec}(\boldsymbol{X}_j^\phi) \tag{11}$$

for some $\boldsymbol{S}_1 \in \mathcal{S}(h)$, $\boldsymbol{S}_2 \in \mathcal{S}(w)$, $\boldsymbol{Q} \in \mathcal{O}(c)$, as the desired equivalence relation. This relation allows for orthogonal invariance across the channel dimension but only shift invariance across the spatial dimensions. The mixed product property of Kronecker products, $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AB} \otimes \boldsymbol{CD}$, ensures that the overall transformation maintains the group structure and remains an isometry. Figure 2B uses a toy dataset (stacked MNIST digits) to show that this metric is sensitive to differences in spatial activation patterns, but insensitive to coherent spatial translations across channels. In contrast, metrics that ignore the convolutional structure (as in past work [4, 7]) treat very different spatial patterns as identical representations.

Evaluating Eq. (11) requires optimizing over spatial shifts in conjuction with solving a Procrustes alignment. If we fit the shifts by an exhaustive brute-force search, the overall runtime is $O(mh^2w^2c^2 + hwc^3)$, which is costly if this calculation is repeated across a large collection of networks. In practice, we observe that the optimal shift parameters are typically close to zero (Fig. 3A). This motivates the more stringent equivalence relation:

$$\boldsymbol{X}_i^\phi \sim \boldsymbol{X}_j^\phi \iff \text{vec}(\boldsymbol{X}_i^\phi) = (\boldsymbol{I} \otimes \boldsymbol{I} \otimes \boldsymbol{I} \otimes \boldsymbol{Q})\text{vec}(\boldsymbol{X}_j^\phi) \quad \text{for some } \boldsymbol{Q} \in \mathcal{Q}, \tag{12}$$

which has a more manageable runtime of $O(mhwc^2 + c^3)$. To evaluate the metrics implied by Eq. (12), we can simply reshape each $\boldsymbol{X}_k^\phi$ from a $(m \times h \times w \times c)$ tensor into a $(mhw \times c)$ matrix and apply the Procrustes alignment procedure as done above for previous metrics. In contrast, the "flattened metric" in Fig. 2B reshapes the features into a $(m \times hwc)$ matrix, resulting in a more computationally expensive alignment that runs in $O(mh^2w^2c^2 + h^3w^3c^3)$ time.

2.5 How large of a sample size is needed?

An important issue, particularly in neurobiological applications, is to determine the number of network inputs, m, and neurons, n, that one needs to measure to accurately infer the distance between two network representations [11]. Reasoning about these questions rigorously requires a probabilistic perspective of shape distances, which we outline in Supplement D. Intuitively, looser equivalence relations involve a more flexible class of alignment operations, and thus require more sampled inputs to prevent overfitting. Figure 3B-C show that this intuition holds in practice for data from deep convolutional networks. Metrics with looser equivalence relations—the "flattened" metric in panel B, or e.g. the linear metric in panel C—converge slower to a stable estimate as m is increased.

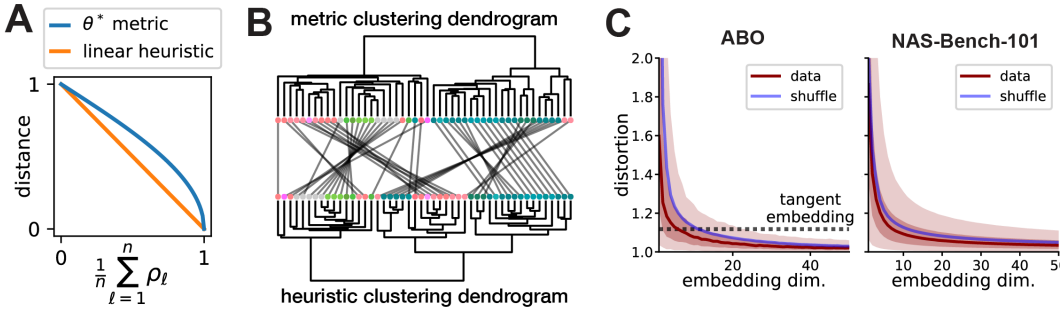

Figure 4: (A) Comparison of metric and linear heuristic. (B) Metric and linear heuristic produce discordant hierarchical clusterings of ABO dataset. Leaves represent brain areas that are clustered by representational similarity (see Fig. 1C), colored by Allen reference atlas, and ordered to maximize dendrogram similarities of adjacent leaves. In the middle, grey lines connect leaves corresponding to the same brain region across the two dendrograms. (C) ABO and NAS-Bench-101 datasets can be accurately embedded into Euclidean spaces. Dark red line shows median distortion. Light red shaded region corresponds to 5th to 95th percentiles of distortion, dark red shaded corresponds to interquartile range. The mean distortion of a null distribution over representations (blue line) was generated by shuffling the m inputs independently in each network.

2.6 Modeling approaches and conceptual insights

Generalized shape metrics facilitate several new approaches and perspectives. In many cases, the set of K neural representations can be visualized as K points on a smooth manifold (see Fig. 1). This holds rigorously due to the *quotient manifold theorem* [34] so long as $\mathcal{G}$ is not a finite set (e.g. corresponding to permutation) and all matrices are full rank in the feature space. This geometric intuition can be made even stronger when $\mathcal{G}$ corresponds to a connected manifold, such as $\mathcal{SO}(p)$. In this case, it can be shown that the geodesic distance between two neural representations coincides with the metrics we defined in Propositions 1 and 2 (see Supplement C, and [14]). This result extends to the well-documented manifold structure of *Kendall's shape space* [23].

Viewing neural representations as points on a manifold is not a purely theoretical exercise—several statistical models can be adapted to manifold-valued data (e.g. principal geodesic analysis [35] provides a generalization of PCA), and additional adaptions are an area of active research [36]. However, there is generally no simple connection between these curved geometries and the flat geometries of Euclidean or Hilbert spaces [37].[1] Unfortunately, the majority of off-the-shelf machine learning tools are incompatible with the former and require the latter. Thus, we can resort to a heuristic approach: the set of K representations can be embedded into a Euclidean space that approximately preserves the pairwise shape distances. One possibility, employed widely in shape analysis, is to embed points in the tangent space of the manifold at a reference point [40, 41]. Another approach, which we demonstrate below with favorable results, is to optimize the embedding directly [24, 42].

3 Applications and Results

We analyzed two large-scale public datasets spanning neuroscience (Allen Brain Observatory, ABO; Neuropixels - visual coding experiment; [20]) and deep learning (NAS-Bench-101; [21]). We constructed the ABO dataset by pooling recorded neurons from $K = 48$ anatomically defined brain regions across all sessions; each $\boldsymbol{X}_k \in \mathbb{R}^{m \times n}$ was a dimensionally reduced matrix holding the neural responses (summarized by $n = 100$ principal components) to $m = 1600$ movie frames (120 second clip, "natural movie three"). The full NAS-Bench-101 dataset contains 423,624 architectures; however, we analyze a subset of $K = 2000$ networks for simplicity. In this application each $\boldsymbol{X}_k \in \mathbb{R}^{m \times n}$ is a representation from a specific network layer, with $(m, n) \in \{(32^2 \times 10^5, 128), (16^2 \times 10^5, 256), (8^2 \times 10^5, 512), (10^5, 512)\}$. Here, n corresponds to the number of channels and m is the product of the number of test set images (10^5) and the height and width dimensions of the convolutional layer—i.e., we use equivalence relation in Eq. (12) to evaluate dissimilarity.

[1] However, see [38] for a conjectured relationship and [39] for a result in the special case of 2D shapes.

Triangle inequality violations can occur in practice when using existing methods. As mentioned above, a dissimilarity measure based on the mean canonical correlation, $1 - \sum_\ell \rho_\ell / n$, has been used in past work [5, 9]. We refer to this as the "linear heuristic." A slight reformulation of this calculation, $\arccos\left(\sum_\ell \rho_\ell / n\right)$, produces a metric that satisfies the triangle inequality (see Eq. (10)). Figure 4A compares these calculations as a function of the average (regularized) canonical correlation: one can see that $\arccos(\cdot)$ is approximately linear when the mean correlation is near zero, but highly nonlinear when the mean correlation is near one. Thus, we reasoned that triangle inequality violations are more likely to occur when K is large and when many network representations are close to each other. Both ABO and NAS-Bench-101 datasets satisfy these conditions, and in both cases we observed triangle inequality violations by the linear heuristic with full regularization ($\alpha = 1$): 17/1128 network pairs in the ABO dataset had at least one triangle inequality violation, while 10128/100000 randomly sampled network pairs contained violations in the NAS-Bench-101 Stem layer dataset. We also examined a standard version of RSA that quantifies similarity via Spearman's rank correlation coefficient [6]. Similar to the results above, we observed violations in 14/1128 pairs of networks in the ABO dataset.

Overall, these results suggest that generalized shape metrics correct for triangle inequality violations that do occur in practice. Depending on the dataset, these violations may be rare (~1% occurrence in ABO) or relatively common (~10% in the Stem layer of NAS-Bench-101). While we do not expect these violations to invalidate the qualitative conclusions of past work, these differences can produce quantitative discrepancies in downstream analyses. For example, the dendrograms produced by hierarchical clustering differ depending on whether one uses the linear heuristic or the shape distance (~85.1% dendrogram similarity as quantified by the method in [43]; see Fig. 4B).

Neural representation metric spaces can be approximated by Euclidean spaces. Having established that neural representations can be viewed as elements in a metric space, it is natural to ask if this metric space is, loosely speaking, "close to" a Euclidean space. We used standard multidimensional scaling methods (SMACOF, [24]; implementation in [44]) to obtain a set of embedded vectors, $\boldsymbol{y}_i \in \mathbb{R}^L$, for which $\theta_1(\boldsymbol{X}_i^\phi, \boldsymbol{X}_j^\phi) \approx \|\boldsymbol{y}_i - \boldsymbol{y}_j\|$ for $i, j \in 1, \ldots, K$. The embedding dimension L is a freely chosen hyperparameter. This embedding problem admits multiple formulations and optimization strategies [42], which could be systematically explored in future work. Our simple approach already yields promising results: in both datasets, we find that moderate embedding dimensions ($L \approx 20$) is sufficient to produce high-quality embeddings. Following past work [45], we quantify the embedding distortions multiplicatively as:

$$\max\left(\theta_1(\boldsymbol{X}_i^\phi, \boldsymbol{X}_j^\phi)/\|\boldsymbol{y}_i - \boldsymbol{y}_j\|;\ \|\boldsymbol{y}_i - \boldsymbol{y}_j\|/\theta_1(\boldsymbol{X}_i^\phi, \boldsymbol{X}_j^\phi)\right) \tag{13}$$

for each pair of networks $i, j \in 1, \ldots K$. Plotting the distortions as a function of L (Fig. 4C), we see that they rapidly decrease, such that 95% of pairwise distances are distorted by, at most, ~5% (ABO data) or 10% (NAS-Bench-101) for sufficiently large L. Past work [9] has used multidimensional scaling heuristically to visualize collections of network representations in $L = 2$ dimensions. Our results here suggest that such a small value of L, while being amenable to visualization, results in a highly distorted embedding. It is noteworthy that the situation improves dramatically when L is even modestly increased. While we cannot visualize these higher-dimensional vector embeddings, we can use them as features for downstream modeling tasks. This is well-motivated as an approximation to performing model inference in the true metric space that characterizes neural representations [45].

Anatomical structure and hierarchy is reflected in ABO representations. We can now collect the L-dimensional vector embeddings of K network representations into a matrix $\boldsymbol{Z} \in \mathbb{R}^{K \times L}$. We expect the distance between any two rows, $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|$, to closely reflect the distance between network representations i and j in shape space. We applied PCA to $\boldsymbol{Z}$ to visualize the $K = 48$ brain regions and found that anatomically related brain regions indeed were closer together in the embedded space (Fig. 5A): cortical and sub-cortical regions are separated along PC 1, and different layers of the same region (e.g. layers 2/3, 4, 5, and 6a of VISp) are clustered together. As expected from the distortions shown in Fig. 4C, performing multidimensional scaling directly ($L = 2$, as done in [9]) results in a qualitatively different outcome (see Supplement E). Additionally, we used $\boldsymbol{Z}$ to fit an ensembled kernel regressor (see Supplement E) to predict an anatomical hierarchy score (defined in [46]) from the embedded vectors. Overall, these results demonstrate that the geometry of the learned embedding is scientifically interpretable and can be exploited for novel analyses, such as nonlinear regression. To our best knowledge, the fine scale anatomical parcellation used here is novel in the context of representational similarity studies. The heightened scale and complexity of these data motivate the use of methods grounded in metric spaces that are amenable to detailed geometric interpretations.

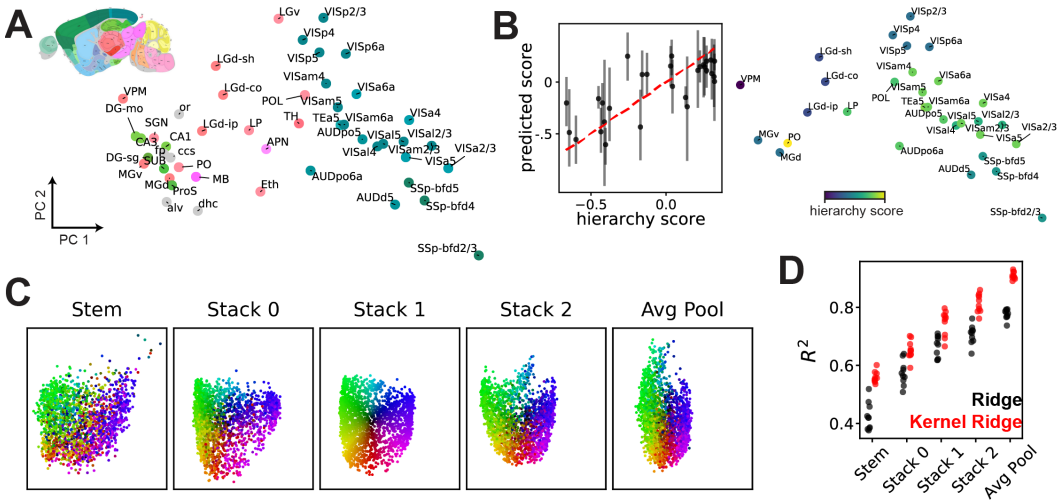

Figure 5: (A) PCA visualization of representations across 48 brain regions in the ABO dataset. Areas are colored by the reference atlas (see inset), illustrating a functional clustering of regions that maps onto anatomy. (B) *Left*, kernel regression predicts anatomical hierarchy [46] from embedded representations (see Supplement E). *Right*, PCA visualization of 31 areas labeled with hierarchy scores. (C) PCA visualization of 2000 network representations (a subset of NAS-Bench-101) across five layers, showing global structure is preserved across layers. Each network is colored by its position in the "Stack 1" layer (the middle of the architecture). (D) Embeddings of NAS-Bench-101 representations are predictive of test set accuracy, *even in very early layers*.

NAS-Bench-101 representations show persistent structure across layers. Since we collected representations across five layers in each deep network, the embedded representation vectors form a set of five $K \times L$ matrices, $\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3, \boldsymbol{Z}_4, \boldsymbol{Z}_5\}$. We aligned these embeddings by rotations in $\mathbb{R}^L$, and then performed PCA to visualize the $K = 2000$ network representations from each layer in a common low-dimensional space. We observe that many features of the global structure are remarkably well-preserved—two networks that are close together in the `Stack1` layer are assigned similar colors in Fig. 5C, and are likely to be close together in the other four layers. This preservation of representational similarity across layers suggests that even early layers contain signatures of network performance, which we expect to be present in the `AvgPool` layer. Indeed, when we fit ridge and RBF kernel ridge regressors to predict test set accuracy from representation embeddings, we see that even early layers support moderately good predictions (Fig. 5D). This is particularly surprising for the `Stem` layer. This is the first layer in each network, and its architecture is identical for all networks. Thus, the differences that are detected in the `Stem` layer result only from differences in backpropagated gradients. Again, these results demonstrate the ability of generalized shape metrics to incorporate neural representations into analyses with greater scale (K corresponding to thousands of networks) and complexity (nonlinear kernel regression) than has been previously executed.

4 Conclusion and Limitations

We demonstrated how to scale investigations into neural representations to large collections of networks by grounding analyses in proper metric spaces. An important limitation of our work, as well as the past works we build upon, is the possibility that representational geometry may only be loosely tied to the underlying mechanisms of network function [9]. On the other hand, analyses of representational geometry provide intermediate-level insights into network function [47]. Further, as we showed, these analyses can be systematically and quantitatively studied across thousands of networks—a much larger scale than is typically considered. Our work exploited connections to previously unrecognized shape analysis literature [12–14], while suggesting only modest modifications to existing similarity measures like CCA (see Fig. 4A). Further, several of the metrics we utilize can be viewed as geodesic distances on Riemannian manifolds [23]. Future work would ideally exploit methods that are rigorously adapted to such manifolds, which are being actively developed [36]. Nonetheless, we found that optimized Euclidean embeddings, while only approximate, provide a practical off-the-shelf solution for large-scale surveys of neural representations.

References

[1] David GT Barrett, Ari S Morcos, and Jakob H Macke. "Analyzing biological and artificial neural networks: challenges with opportunities for synergy?" *Current Opinion in Neurobiology* 55 (2019). Machine Learning, Big Data, and Neuroscience, pp. 55–64.

[2] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". *Proceedings of the National Academy of Sciences* 111.23 (2014), pp. 8619–8624.

[3] Santiago A. Cadena, Fabian H. Sinz, Taliah Muhammad, Emmanouil Froudarakis, Erick Cobos, Edgar Y. Walker, Jake Reimer, Matthias Bethge, Andreas Tolias, and Alexander S. Ecker. "How well do deep neural networks trained on object recognition characterize the mouse visual system?" *NeurIPS Workshop Neuro AI* (2019).

[4] 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*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 6076–6085.

[5] Ari Morcos, Maithra Raghu, and Samy Bengio. "Insights on representational similarity in neural networks with canonical correlation". *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 5727–5736.

[6] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. "Representational similarity analysis - connecting the branches of systems neuroscience". *Frontiers in Systems Neuroscience* 2 (2008), p. 4.

[7] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. "Similarity of Neural Network Representations Revisited". Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 3519–3529.

[8] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. "Algorithms for learning kernels based on centered alignment". *The Journal of Machine Learning Research* 13.1 (2012), pp. 795–828.

[9] Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. "Universality and individuality in neural dynamics across large populations of recurrent networks". *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 15629–15641.

[10] Thao Nguyen, Maithra Raghu, and Simon Kornblith. *Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network Representations Vary with Width and Depth*. 2020.

[11] Jianghong Shi, Eric Shea-Brown, and Michael Buice. "Comparison Against Task Driven Artificial Neural Networks Reveals Functional Properties in Mouse Visual Cortex". *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 5764–5774.

[12] Christopher G. Small. *The statistical theory of shape*. Springer series in statistics. New York: Springer, 1996.

[13] David George Kendall, Dennis Barden, Thomas K Carne, and Huiling Le. *Shape and shape theory*. New York: Wiley, 1999.

[14] Ian L. Dryden and Kantilal Mardia. *Statistical shape analysis with applications in R*. Chichester, UK Hoboken, NJ: John Wiley & Sons, 2016.

[15] Peter N Yianilos. "Data structures and algorithms for nearest neighbor search in general metric spaces". *Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms*. 1993, pp. 311–321.

[16] Sanjoy Dasgupta and Philip M Long. "Performance guarantees for hierarchical clustering". *Journal of Computer and System Sciences* 70.4 (2005), pp. 555–569.

[17] Saaid Baraty, Dan A. Simovici, and Catalin Zara. "The Impact of Triangular Inequality Violations on Medoid-Based Clustering". *Foundations of Intelligent Systems*. Ed. by Marzena Kryszkiewicz, Henryk Rybinski, Andrzej Skowron, and Zbigniew W. Raś. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 280–289.

[18] Fei Wang and Jimeng Sun. "Survey on distance metric learning and dimensionality reduction in data mining". *Data Mining and Knowledge Discovery* 29.2 (2015), pp. 534–564.

[19] C. Chang, W. Liao, Y. Chen, and L. Liou. "A Mathematical Theory for Clustering in Metric Spaces". *IEEE Transactions on Network Science and Engineering* 3.1 (2016), pp. 2–16.

[20] Joshua H. Siegle et al. "Survey of spiking in the mouse visual system reveals functional hierarchy". *Nature* 592.7852 (2021), pp. 86–92.

[21] Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, and Frank Hutter. "NAS-Bench-101: Towards Reproducible Neural Architecture Search". *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 7105–7114.

[22] Taco Cohen and Max Welling. "Group Equivariant Convolutional Networks". *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 2016, pp. 2990–2999.

[23] David G. Kendall. "Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces". *Bulletin of the London Mathematical Society* 16.2 (1984), pp. 81–121.

[24] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[25] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, 2012.

[26] R Jonker and A Volgenant. "A shortest augmenting path algorithm for dense and sparse linear assignment problems". *Computing* 38.4 (1987), pp. 325–340.

[27] David F. Crouse. "On implementing 2D rectangular assignment algorithms". *IEEE Transactions on Aerospace and Electronic Systems* 52.4 (2016), pp. 1679–1696.

[28] Peter H. Schönemann. "A generalized solution of the orthogonal procrustes problem". *Psychometrika* 31.1 (1966), pp. 1–10.

[29] J. C. Gower and Garmt B. Dijksterhuis. *Procrustes problems*. Oxford New York: Oxford University Press, 2004.

[30] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. "Optimal whitening and decorrelation". *The American Statistician* 72.4 (2018), pp. 309–314.

[31] Harold Hotelling. "Relations Between Two Sets of Variates". *Biometrika* 28.3/4 (1936), pp. 321–377.

[32] Hrishikesh D Vinod. "Canonical ridge and econometrics of joint production". *Journal of econometrics* 4.2 (1976), pp. 147–166.

[33] P. L. Lai and C. Fyfe. "Kernel and Nonlinear Canonical Correlation Analysis". *International Journal of Neural Systems* 10.05 (2000). PMID: 11195936, pp. 365–377.

[34] John M. Lee. *Introduction to smooth manifolds*. 2nd ed. Graduate texts in mathematics 218. New York ; London: Springer, 2013.

[35] P. Thomas Fletcher and Sarang Joshi. "Riemannian geometry for the statistical analysis of diffusion tensor data". *Signal Processing* 87.2 (2007). Tensor Signal Processing, pp. 250–262.

[36] Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Daniel Brooks, Bernhard Kainz, Claire Donnat, Susan Holmes, and Xavier Pennec. "Geomstats: A Python Package for Riemannian Geometry in Machine Learning". *Journal of Machine Learning Research* 21.223 (2020), pp. 1–9.

[37] Aasa Feragen, Francois Lauze, and Soren Hauberg. "Geodesic Exponential Kernels: When Curvature and Linearity Conflict". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[38] Aasa Feragen and Søren Hauberg. "Open Problem: Kernel methods on manifolds and metric spaces. What is the probability of a positive definite geodesic exponential kernel?" *29th Annual Conference on Learning Theory*. Ed. by Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir. Vol. 49. Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR, 2016, pp. 1647–1650.

[39] Sadeep Jayasumana, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. "A Framework for Shape Analysis via Hilbert Space Embedding". *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2013.

[40] Ian L Dryden and Kanti V Mardia. "Multivariate shape analysis". *Sankhyā: The Indian Journal of Statistics, Series A* (1993), pp. 460–480.

[41] F. James Rohlf. "Shape Statistics: Procrustes Superimpositions and Tangent Spaces". *Journal of Classification* 16.2 (1999), pp. 197–223.

[42] Akshay Agrawal, Alnur Ali, and Stephen Boyd. "Minimum-Distortion Embedding". *arXiv* (2021).

[43] Alexander J. Gates, Ian B. Wood, William P. Hetrick, and Yong-Yeol Ahn. "Element-centric clustering comparison unifies overlaps and hierarchy". *Scientific Reports* 9.1 (2019), p. 8574.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[45] Leena Chennuru Vankadara and Ulrike von Luxburg. "Measures of distortion for machine learning". *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[46] Julie A. Harris et al. "Hierarchical organization of cortical and thalamic connectivity". *Nature* 575.7781 (2019), pp. 195–202.

[47] Jess B. Hamrick and Shakir Mohamed. "Levels of Analysis for Machine Learning". *Proceedings of the ICLR 2020 Workshop on Bridging AI and Cognitive Science*. 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See final section on conclusions and limitations
   (c) Did you discuss any potential negative societal impacts of your work? [No] We cannot foresee any negative impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] They are included in the supplement.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] They are provided in the supplement. The vast majority of models were pre-trained, with details provided in [21].
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See, e.g., Fig. 5B
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Supplement E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See, e.g., our citations to [20, 21, 44]
   (b) Did you mention the license of the assets? [N/A] We do not believe this is needed for our use cases.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We have included a small code package in the supplement
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] No assets involve data collected from human subjects and the assets are all freely shared
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]