CENTERED KERNEL DISTANCE: DIMENSION-FREE ERROR BOUNDS FOR COMPARING NEURAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The asymptotic error of representational similarity metrics such as CKA and RSA has been well-characterized, but there are no realistic bounds on how estimates of representational similarity *simultaneously* depend on the number of sampled neurons (N) and experimental conditions (M). We study the natural metric counterpart of CKA and RSA, which we call the Centered Kernel Distance (CKD). We show that CKD can be studied via an unbiased estimator, whereas a large collection of similarity measures including CKA have no unbiased estimator. Under mild assumptions, we show that estimation error of CKD decays in proportion to $1/M + 1/N$. Strikingly, the rate in this bound does not depend on the dimensionality of the space of experimental conditions. Thus, similarity estimation does not suffer from the curse of dimensionality, which plagues similar problems in high-dimensional statistics. We verify our results empirically in random shallow neural networks, trained neural networks, and macaque visual cortex data. Ultimately, this paper advances a framework for studying the error of similarity estimators, and our theoretical and empirical predictions produce specific prescriptions on how to most efficiently estimate the similarity between multiple neural systems.

1 INTRODUCTION

Comparative analysis of neural representations is an active topic of research in neuroscience (Kriegeskorte & Diedrichsen, 2019) and AI (Sucholutsky et al., 2023; Klabunde et al., 2025). In neuroscience, initiatives such as BrainScore or the Algonauts competition apply similarity analyses to understand how computational models align with neural data (Schrimpf et al., 2018; Gifford et al., 2025). In machine learning, similarity analysis plays an important role in model stitching (Bansal et al., 2021), and more generally in studying how different training and architectural decisions impact learned representations (Kornblith et al., 2019; Raghu et al., 2021; Huh et al., 2024). Practitioners often quantify similarity via scalar scores such as centered kernel alignment (CKA; Kornblith et al., 2019) and the closely related framework of representational similarity analysis (RSA; Kriegeskorte et al., 2008)—see Williams (2024) for a precise connection between the two.

In practice, representational similarity scores are estimated from limited data. Even in artificial network models, it is impossible to exhaustively sample most input spaces (*e.g.*, natural images). In biological systems, we must additionally contend with sub-sampling neural dimensions since experimental recordings only capture a small fraction of neurons within a given brain region (Stevenson & Kording, 2011). Furthermore, biological responses are noisy and variable from trial-to-trial. Prior work has separately analyzed the effect of subsampling stimulus inputs (Pospisil et al., 2024) and subsampling neurons (Chun et al., 2025; Kang et al., 2025) in representational similarity measures. To our knowledge, a non-asymptotic theory that simultaneously accounts for subsampling along both axes (stimuli and neurons) has not been previously outlined.

Here, we analyze the **Centered Kernel Distance (CKD)** which we define as the distance metric naturally associated with Centered Kernel Alignment (CKA; Kornblith et al., 2019). The metric itself is not new; Harvey et al. (2024) analyzed it under a different name, and gave it operational meaning by showing CKD equals the average distance between decoded signals over random tasks. We focus our attention on statistical estimation of this metric and reveal several favorable properties:

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

- We show that the square of CKD admits an unbiased estimator, $\widehat{\mathrm{CKD}}^2$, as a three-sample U-statistic *where inputs as well as neurons are subsampled*. Alternative measures including CKA and the squared Procrustes distance are not unbiasedly estimable (see Appendix D.1).
- We derive non-asymptotic bounds on $\mathbb{V}\mathrm{ar}[\widehat{\mathrm{CKD}}^2]$ in terms of the number of sampled inputs M and neurons N_X and N_Y. Writing $\epsilon := 1/M + 1/N_X + 1/N_Y$, the variance scales as $\mathrm{CKD}^2 \cdot \epsilon + \epsilon^2$. The bound is *signal-adaptive*, tightening as the two representations converge. Interestingly, the dimension of the input space does not appear in this bound. Hence, there is no curse of dimensionality, and CKD^2 can be accurately estimated over complex domains such as natural images even in experiments with limited trials and neuron counts.
- Metrics satisfying the triangle inequality enable rigorous comparison of multiple neural systems at once (e.g. three or more animal subjects or artificial network models; see Williams et al., 2021). Recognizing that CKD is built from an inner product between kernels, we leverage classical results (see Dokmanic et al., 2015) to perform closed-form metric repair and low-dimensional embedding.

2 Problem Setup and Background

2.1 Sampling model

We model neural responses as functions or "tuning curves" mapping inputs (e.g. images) onto activation level. Formally, let $\mathcal{F}$ and $\mathcal{G}$ denote two stochastic processes representing distributions over tuning curves, and let P denote a distribution over network inputs, which live in a space with some intrinsic dimension D. Intuitively, differences in the geometry of neural representations stem from underlying differences in $\mathcal{F}$ and $\mathcal{G}$, measured according to the distribution of inputs specified by P. Throughout, we write

$$k(u_1, u_2) = \mathbb{E}_{f\sim\mathcal{F}}\left[f(u_1)f(u_2)\right] \qquad \text{and} \qquad h(u_1, u_2) = \mathbb{E}_{g\sim\mathcal{G}}\left[g(u_1)g(u_2)\right] \tag{1}$$

for the covariance operators of $\mathcal{F}$ and $\mathcal{G}$.

We analyze an experiment in which N_X and N_Y tuning curves are randomly sampled from each system, along with M inputs fed to both:

$$f_1, \ldots, f_{N_X} \sim \mathcal{F}, \quad g_1, \ldots, g_{N_Y} \sim \mathcal{G}, \quad \text{and} \quad u_1, \ldots, u_M \sim P, \tag{2}$$

independently. We thus observe $M \times N_X$ and $M \times N_Y$ random matrices $X_{mn} = f_n(u_m)$ and $Y_{mn} = g_n(u_m)$. Crucially, *we assume both axes—neurons and inputs—are randomly sampled*, and our focus is to elucidate how these two sources of sampling error jointly impact estimates of representational similarity. Note that the M inputs are shared between the two systems, while the neurons are drawn independently. While i.i.d. sampling of biological neurons is not necessarily possible in practice, the data generating process in eq. (2) captures the essence of how neural sampling is treated when testing experimental hypotheses.

2.2 Centred Kernel Alignment and Centered Kernel Distance

Let $\widehat{\boldsymbol{K}}(\boldsymbol{X}) = \frac{1}{N_X}\boldsymbol{X}\boldsymbol{X}^\top$ and $\widehat{\boldsymbol{H}}(\boldsymbol{Y}) = \frac{1}{N_Y}\boldsymbol{Y}\boldsymbol{Y}^\top$ denote the empirical linear kernel matrices, normalized by the number of sampled neurons in each system. To prevent notational clutter, we treat these matrices as implicit functions, writing $\widehat{\boldsymbol{K}}$ in place of $\widehat{\boldsymbol{K}}(\boldsymbol{X})$ and $\widehat{\boldsymbol{H}}$ in place of $\widehat{\boldsymbol{H}}(\boldsymbol{Y})$.

Centered Kernel Alignment (CKA; Cortes et al. 2012) is a popular framework for quantifying similarity in hidden layer representations (Kornblith et al., 2019). Writing $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{M}\boldsymbol{1}\boldsymbol{1}^\top$ for the $M \times M$ centering matrix, the plug-in linear CKA score is

$$\mathrm{CKA}_{\mathrm{pi}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{S_{\mathrm{pi}}(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{S_{\mathrm{pi}}(\boldsymbol{X}, \boldsymbol{X})\, S_{\mathrm{pi}}(\boldsymbol{Y}, \boldsymbol{Y})}} \quad \text{where} \quad S_{\mathrm{pi}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{M^2}\,\mathrm{Tr}\left[\widehat{\boldsymbol{K}}\boldsymbol{C}\widehat{\boldsymbol{H}}\boldsymbol{C}\right]. \tag{3}$$

Here S_{pi} is the plug-in **Hilbert–Schmidt Independence Criterion** score (HSIC; Gretton et al. 2005), which targets the *Hilbert–Schmidt inner product* between the centered covariance functions,

$$\lim_{\min(M, N_X, N_Y)\to\infty} S_{\mathrm{pi}}(\boldsymbol{X}, \boldsymbol{Y}) = \langle \bar{k}, \bar{h} \rangle_{\mathrm{HS}} = \mathbb{E}_{u,u'\sim P}\left[\, \bar{k}(u, u')\, \bar{h}(u, u') \,\right], \tag{4}$$

where $\bar{k}$ and $\bar{h}$ are the doubly centered kernel functions

$$\bar{k}(u_1, u_2) = k(u_1, u_2) - \mathbb{E}_{a\sim P}\left[k(a, u_2)\right] - \mathbb{E}_{b\sim P}\left[k(u_1, b)\right] + \mathbb{E}_{a,b\sim P}\left[k(a, b)\right], \tag{5}$$

$$\bar{h}(u_1, u_2) = h(u_1, u_2) - \mathbb{E}_{a\sim P}\left[h(a, u_2)\right] - \mathbb{E}_{b\sim P}\left[h(u_1, b)\right] + \mathbb{E}_{a,b\sim P}\left[h(a, b)\right]. \tag{6}$$

Here we focus on a related quantity, which we call the **Centered Kernel Distance (CKD)**. Whereas CKA normalizes the HSIC inner product, CKD is the *distance* that inner product induces:

$$\mathrm{CKD}^2 := \left\|\bar{k} - \bar{h}\right\|_{\mathrm{HS}}^2 = \langle \bar{k}, \bar{k}\rangle_{\mathrm{HS}} + \langle \bar{h}, \bar{h}\rangle_{\mathrm{HS}} - 2\langle \bar{k}, \bar{h}\rangle_{\mathrm{HS}}. \tag{7}$$

We focus on estimating the *squared* distance throughout since it is unbiasedly estimable (see section 3). Substituting eq. (4) into eq. (7), the natural plug-in estimator is

$$\mathrm{CKD}^2_{\mathrm{pi}}(\boldsymbol{X}, \boldsymbol{Y}) = S_{\mathrm{pi}}(\boldsymbol{X}, \boldsymbol{X}) + S_{\mathrm{pi}}(\boldsymbol{Y}, \boldsymbol{Y}) - 2\, S_{\mathrm{pi}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{8}$$

which is consistent, since $\mathrm{CKD}^2_{\mathrm{pi}} \to \mathrm{CKD}^2$ as $\min(M, N_X, N_Y) \to \infty$ by applying the large numbers along each sampling axis. Notably, the order of the limits does not matter for HSIC or CKD, but *does* matter for CKA; jointly taking $M, N_X, N_Y \to \infty$ in CKA produces different behavior than taking one axis to infinity first (see Appendix D.2 and Chun et al., 2025). This already hints that CKD is a more nicely behaved quantity than CKA.

In Appendix B.1, we show that CKD can be viewed as a special case of a maximal mean discrepancy (MMD; Gretton et al., 2012) with a quadratic kernel. This places CKD within a well-developed body of theory, which we expect to be useful beyond the results developed here. Harvey et al. (2024) also previously studied the same quantity as CKD and showed that it can be interpreted as the average difference between linearly decoded signals. Finally, CKD is closely related to Representational Similarity Analysis (RSA; Kriegeskorte et al. 2008). In particular, Proposition 1 in Williams (2024) implies that the CKD between $\widehat{\boldsymbol{K}}$ and $\widehat{\boldsymbol{H}}$ is proportional to the distance between doubly centered representational dissimilarity matrices (RDMs) with a squared Euclidean base distance. Thus, all of our results on CKD immediately extend to an equivalent form of RSA.

The present paper is *not* interested in demonstrating a general preference for CKD over any other metric. Rather, we seek to precisely characterize how easy it is to estimate this metric with subsampled inputs as well as subsampled feature dimensions (i.e., subsampled neurons). We show that CKD is statistically well-behaved, motivating its use in datasets with limited samples.

2.3 Bias-Variance Decomposition and U-statistics

Our narrative relies on the classical bias-variance decomposition of point estimators. Given a target estimand of interest, θ, and an estimator, $\widehat{\theta}$, the decomposition of the mean squared error is

$$\mathbb{E}\left[\left(\theta - \widehat{\theta}\right)^2\right] = \underbrace{\left(\theta - \mathbb{E}\left[\widehat{\theta}\right]\right)^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right]\right)^2\right]}_{\text{variance}} \tag{9}$$

While the bias of several similarity metrics have been previously characterized (Pospisil et al., 2024; Chun et al., 2025), their variances remain underexplored. We aim to resolve this issue in CKD^2.

A key step of our analysis is to show that CKD^2 can be estimated as a *U-statistic* (see Lee, 1990, for a textbook reference). At a high level, the goal of a U-statistic is to estimate a parameter that can be written as $\theta = \mathbb{E}[\phi(z_1, \ldots, z_k)]$, where ϕ is some nonlinear function that is symmetric in its inputs and $z_1, \ldots, z_k$ are independent samples from some distribution P. Given $n \geq k$ independent samples from P, a one-sample, order-k U-statistic estimator is

$$\widehat{\theta} = \binom{n}{k}^{-1} \sum_{\boldsymbol{i}\in\mathcal{I}_{n,k}} \phi(z_{i_1}, \ldots, z_{i_k}) \tag{10}$$

where $\mathcal{I}_{n,k}$ is the set of all length k subsets of unique indices from $\{1, \ldots, n\}$. Since each multi-index, $\boldsymbol{i} \in \mathcal{I}_{n,k}$, has no repeated elements, the expectation of each term in the sum is θ. Hence, the U-statistic, being an average of these terms, is also unbiased $\mathbb{E}[\widehat{\theta}] = \theta$. Moreover, among all unbiased estimators it is generally the optimal one—in the absence of additional assumptions about P, the Lehmann–Scheffé theorem implies that the U-statistic estimator has minimal variance. Song et al. (2012) derived a U-statistic estimator for HSIC under the assumption of a fixed feature space. Chun et al. (2025) extended this estimator to account for sampling feature dimensions (i.e. sampling neural axes). We leverage these constructions heavily in our analysis.

3 USEFUL PROPERTIES OF CKD

3.1 CKD^2 ADMITS AN UNBIASED ESTIMATOR

Many popular measures of representational similarity are incompatible with unbiased estimation. Indeed, in Appendix D.1, we show that no unbiased estimators exist for the squared Procrustes distance or for CKA. In both cases, the bias can be mitigated close to zero (often at the cost of increased variance), but it cannot be entirely eliminated. See Pospisil et al. (2024) for a discussion of bias in the Procrustes distance and Chun et al. (2025) for a discussion of bias in CKA.

The plug-in estimator $\mathrm{CKD}_{\mathrm{pi}}$ from Equation (8) is a linear combination of three HSIC estimators, each of which is biased. The bias arises from two sources. First, S_{pi} averages over *all* pairs of inputs, including the coinciding pairs. Second, the self-terms $S_{\mathrm{pi}}(\boldsymbol{X}, \boldsymbol{X})$ and $S_{\mathrm{pi}}(\boldsymbol{Y}, \boldsymbol{Y})$ additionally average over pairs of coinciding neurons. In both cases, a pair of coinciding samples—the same input twice, or the same neuron twice—compares a quantity to itself, and so contributes a spuriously large value (upward bias). We propose an unbiased estimator of CKD by combining several existing approaches for bias-correction in HSIC.

To remove the first source, we use a *stimulus-corrected* HSIC estimator due to Song et al. (2012):

$$S_{\mathrm{sc}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{M(M-3)} \left[\mathrm{Tr}\left[\tilde{\boldsymbol{K}}\tilde{\boldsymbol{H}}\right] + \frac{(\mathbf{1}^\top \tilde{\boldsymbol{K}}\mathbf{1})(\mathbf{1}^\top \tilde{\boldsymbol{H}}\mathbf{1})}{(M-1)(M-2)} - \frac{2}{M-2}\mathbf{1}^\top \tilde{\boldsymbol{K}}\tilde{\boldsymbol{H}}\mathbf{1} \right], \tag{11}$$

where $\tilde{\boldsymbol{K}}$ and $\tilde{\boldsymbol{H}}$ are versions of $\widehat{\boldsymbol{K}}$ and $\widehat{\boldsymbol{H}}$ with their diagonals set to zero. This estimator is unbiased in the setting of fixed feature dimensions (i.e. sampling M inputs acrossed fixed neural dimensions), but in our setting of interest the self-terms $S_{\mathrm{sc}}(\boldsymbol{X}, \boldsymbol{X})$ and $S_{\mathrm{sc}}(\boldsymbol{Y}, \boldsymbol{Y})$ incur a bias as described above, but the cross term $S_{\mathrm{sc}}(\boldsymbol{X}, \boldsymbol{Y})$ remains unbiased.

To remove the source of bias in the self-terms, we use a *stimulus-and-neuron-corrected* HSIC estimator due to Chun et al. (2025). We can write this estimator as:

$$S_{\mathrm{sanc}}(\boldsymbol{X}) = \frac{1}{N_X(N_X-1)} \sum_{n \neq n'} S_{\mathrm{sc}}(\boldsymbol{x}_n \boldsymbol{x}_n^\top, \boldsymbol{x}_{n'} \boldsymbol{x}_{n'}^\top) \tag{12}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_X} \in \mathbb{R}^M$ are the columns of $\boldsymbol{X}$. Intuitively, S_{sanc} skips all terms in S_{sc} in which the same neuron appears twice in order to remove the upward bias. A direct calculation of S_{sanc} as written above requires $O(N_X^2)$ calls to S_{sc}, but in Appendix B we describe an approach that scales linearly with the number of neurons. Putting these components together, we arrive at our estimator:

$$\widehat{\mathrm{CKD}}^2 = S_{\mathrm{sanc}}(\boldsymbol{X}) + S_{\mathrm{sanc}}(\boldsymbol{Y}) - 2S_{\mathrm{sc}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{13}$$

3.2 THE VARIANCE OF $\widehat{\mathrm{CKD}}^2$ CAN BE EXPLICITLY BOUNDED WITH A FAVORABLE RATE

Our results rely only on an assumption that neural responses have finite kurtosis. This is guaranteed *in vivo* (a neuron cannot fire infinitely) and *in silico* (values overflow).

Theorem 3.1 (Variance Bound, Appendix B)**.** *There exists a constant, $C > 0$, depending only on the maximum kurtosis and scale of individual neuron responses such that*

$$\mathbb{V}\mathrm{ar}(\widehat{\mathrm{CKD}}^2) \leq C\left[\mathrm{CKD}^2\left(\frac{1}{M} + \frac{1}{N_X} + \frac{1}{N_Y}\right) + \left(\frac{1}{M} + \frac{1}{N_X} + \frac{1}{N_Y}\right)^2\right] \tag{14}$$

for the estimator given in Equation (13)*.*

We provide the proof and a precise characterization of the constant C in Appendix B.4. We derive a similar scaling law for HSIC, as well as the respective biased plug-in estimators, with slightly different constants (Appendix A, empirical results alongside CKA in Section D.3). The proof of Theorem 3.1 follows from a construction of $\widehat{\mathrm{CKD}}^2$ as a 3-sample U-statistic, followed by standard strategies for decomposing and upper bounding the variance of a U-statistic.

Estimating the variance by bootstrapping. The constant C in eq. (14) depends on the maximum kurtosis of neural responses. In practice, this constant is almost always conservative. Typically,

the variance still decays at the rate proportional to the right hand side theorem 3.1 but with a more favorable constant. We recommend that practitioners empirically estimate variance of $\widehat{\mathrm{CKD}}^2$ by bootstrapping. Bootstrap estimates of standard errors are thoroughly characterized in the context of U-statistics; one can show that they are consistent for non-degenerate U-statistics and characterize their convergence rates (see Lee, 1990, Ch. 5).

Dimension-independence. Let D denote the intrinsic dimension of the input space—for instance, the dimension of the manifold on which natural images are supported. Many analyses of neural population geometry are cursed by D. Estimating the kernel k itself, or any other function on $\mathcal{U} \times \mathcal{U}$, is a nonparametric regression problem. For these problems, the number of samples required to reach a fixed error grows exponentially in D (see minimax lower bounds in Györfi et al., 2002). Our results show that comparing two representations is a fundamentally easier problem—the rate in eq. (14) does not degrade as the input dimension increases.

3.3 Decorrelated noise does not add bias and preserves $\mathbb{V}\mathrm{ar}[\widehat{\mathrm{CKD}}^2]$ scaling

Neural responses are noisy in biological systems—spike counts can differ dramatically across nominally identical trials (see e.g. Goris et al. 2014). Noise is also sometimes present in artificial networks (e.g. dropout layers), and past work has introduced specialized metrics to compare the geometry of noise across neural systems (Duong et al., 2023; Nejatbakhsh et al., 2025). Here, we focus on comparing the geometry of the average response from limited measurements. In particular, how does the scale of noise impact our estimate of CKD?

To address this question, we modify the sampling model of Section 2.1 to allow for multiple presentations of the same stimulus input. Let $L \geq 1$ denote the number of repetitions, so that the response of neuron n to input i on repetition ℓ is given by $f_n(u_i) + \xi_{ni\ell}$ where $f_n \sim \mathcal{F}$ is a sampled tuning curve, $u_i \sim P$ is a sampled input, and $\xi_{ni\ell}$ denotes mean zero random noise with finite variance.

If noise is independent across neurons, the U-statistic estimator $\widehat{\mathrm{CKD}}^2$ turns out to remain unbiased even when $L = 1$ (see Proposition C.1 in Appendix C). This is surprising in light of prior work (e.g., Walther et al., 2016) which carefully designed procedures based on $L \geq 2$ repetitions to debias estimates of representational similarity from noisy measurements. In biological experiments, the number of repetitions, L, and the number of unique stimuli, M, are inversely coupled to each other due to constraints on the experiment's duration. The ability to use $L = 1$ repetition therefore allows one to sample a greater number of inputs, M, decreasing the variance of the CKD^2 estimate.

Furthermore, under the assumption of decorrelated noise, the variance bound in Theorem 3.1 applies essentially verbatim, with the constant C depending on the maximum kurtosis of the observed (noise-corrupted) responses rather than the kurtosis of the underlying tuning curves (see Appendix C.4). In fact, the constants appearing in the first-order expansion of the variance are completely insensitive to trial-to-trial noise (see Proposition C.2 in Appendix C). Thus, the asymptotic rate of convergence for $\mathbb{V}\mathrm{ar}[\widehat{\mathrm{CKD}}^2]$ is exactly the same as the noiseless case.

A limitation of this analysis is that it assumes the absence of noise correlations. That is, for two neurons n and n', the above results assume $\mathbb{C}\mathrm{ov}[\xi_{ni\ell}, \xi_{n'i\ell}] = 0$ across all trials. In practice, this assumption is not strictly satisfied—neural responses are generally correlated from trial-to-trial and the strength of these correlations can be stimulus dependent (Kohn & Smith, 2005; Ponce-Alvarez et al., 2013). As we document in Appendix C, these correlations introduce bias in $\widehat{\mathrm{CKD}}^2$. We note this as a limitation, and leave a detailed investigation to future work. In particular, a fully satisfactory answer would require relaxing our independence assumptions in eq. (2), which considerably complicates the picture.

3.4 Closed form metric repair for multi-systems analysis

We have thus far considered estimating the distance between a single pair of neural systems. One is often interested in extending these analyses across multiple systems at once—e.g. across network architectures, layers, or animal subjects in a neuroscience experiment—in order to elucidate how systematic variation in structure of models induces variation in their function and, conversely, how specific changes in the value of similarity measures indicate something about the underlying structure of models (Lange et al., 2023; He et al., 2026). When comparing a collection of T neural

systems, the $T \times T$ matrix of pairwise distance scores Δ is the object of primary interest. When this matrix is constructed with respect to an underlying *metric space*, it contains rich information about the space of representations (e.g. whether systems are clustered into discrete groups or distributed uniformly; see Williams et al., 2021). Taken one step further, one can measure geodesic paths and angles between systems to trace how factors such as network depth or data augmentation move a representation through this space (Lange et al., 2023; He et al., 2026).

Formally, let $\bar{k}_1, \ldots, \bar{k}_T$ denote the centered kernels of the T systems, let $\boldsymbol{G}$ be the $T \times T$ Gram matrix based on the HSIC inner product. Then,

$$G_{st} = \left\langle \bar{k}_s, \bar{k}_t \right\rangle_{\mathrm{HS}}, \qquad \Delta_{st} = G_{ss} + G_{tt} - 2G_{st}, \qquad \text{for all } t, s \in \{1, \ldots, T\}. \tag{15}$$

Given observations $\boldsymbol{X}_i \in \mathbb{R}^{M \times N_t}$ for $t \in \{1, \ldots, T\}$, we can form an entrywise unbiased estimate $\widehat{\boldsymbol{G}}$ of $\boldsymbol{G}$ using the elements developed in Section 3.1, where $\widehat{G}_{tt} = S_{\mathrm{sanc}}(\boldsymbol{X}_t, \boldsymbol{X}_t)$ and $\widehat{G}_{st} = S_{\mathrm{sc}}(\boldsymbol{X}_s, \boldsymbol{X}_t)$ for $s \neq t$. Since $\boldsymbol{G}$ is a Gram matrix, it is positive semidefinite. However, due to sampling noise, $\widehat{\boldsymbol{G}}$ is generally not positive semidefinite. The corresponding estimate of CKD distances, given by $\widehat{D}_{st} = \widehat{G}_{ss} + \widehat{G}_{tt} - 2\widehat{G}_{st}$, can have negative entries and triangle inequality violations. We can repair this by projecting $\boldsymbol{G}$ onto the feasible cone $\mathcal{S}_+^T$ of $T \times T$ positive semidefinite matrices, which fortuitously can be done in closed form (Higham, 1988):

$$\boldsymbol{G}^\star \;=\; \mathrm{argmin}_{\boldsymbol{G} \in \mathcal{S}_+^T} \left\| \boldsymbol{G} - \widehat{\boldsymbol{G}} \right\|_F^2 = \boldsymbol{U} \max(\Lambda, 0) \boldsymbol{U}^\top, \tag{16}$$

where $\boldsymbol{U}\Lambda\boldsymbol{U}^\top$ denotes the eigendecomposition of $\widehat{\boldsymbol{G}}$ and $\max(\Lambda, 0)$ is applied elementwise to clip negative eigenvalues to zero. Projection onto a convex set containing the truth is non-expansive, so

$$\left\| \boldsymbol{G}^\star - \boldsymbol{G} \right\|_F \;\leq\; \left\| \widehat{\boldsymbol{G}} - \boldsymbol{G} \right\|_F, \tag{17}$$

for every realization of data. The price is that $\boldsymbol{G}^\star$ is no longer element-wise unbiased since the projection operation is nonlinear. Thus, metric repair accepts a small bias in exchange for a larger reduction in variance. An improved estimate of the $T \times T$ matrix of CKD values has elements $\Delta^*_{st} = G^*_{ss} + G^*_{tt} - 2G^*_{st}$.

3.5 Multidimensional scaling gives a Euclidean embedding of neural systems in closed form

The eigendecomposition that repairs $\boldsymbol{G}$ also provides an embedding. Writing $\boldsymbol{V} = \boldsymbol{U} \max(\Lambda, 0)^{1/2}$, we have $\boldsymbol{G}^\star = \boldsymbol{V}\boldsymbol{V}^\top$, so the rows $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T$ of $\boldsymbol{V}$ are explicit coordinates satisfying

$$\left\| \boldsymbol{v}_s - \boldsymbol{v}_t \right\|_2^2 \;=\; D^\star_{st} \qquad \text{for all } s, t \in \{1, \ldots, T\}. \tag{18}$$

This is the classical multidimensional scaling (MDS) construction (Gower, 1966); see Dokmanic et al. (2015) for a modern treatment of Euclidean distance matrices.

It is worth emphasizing that the embedding in Equation (18) is exact, rather than approximate. Thus, the coordinates held in $\boldsymbol{V}$ can be directly passed to essentially any method for clustering, dimensionality reduction, regression, or classification. Alternative metrics including the Procrustes distance and $\cos^{-1}$ (CKA) satisfy the triangle inequality but do not admit an inner product. Analyses are then restricted either to approximate MDS embeddings (see Borg & Groenen, 2005) or to methods operating on raw distances, such as hierarchical clustering and nearest-neighbor prediction. Alternative metrics also do not support closed form solutions for metric repair, as done in Section 3.4.

4 Applications and Experiments

4.1 Validation on Random Neural Networks

We first confirm our theoretical predictions in one layer neural networks with random weights (see Appendix G for construction details). For simplicity, we compare networks of equal width, $N = N_X = N_Y$, but comment on the case of $N_X \neq N_Y$ in Appendix E. We report the normalized variance of the U-statistic estimator, $\mathbb{V}\mathrm{ar}[\widehat{\mathrm{CKD}}^2]/\mathrm{CKD}^2$, where the numerator is estimated across multiple simulations (different random networks) and the denominator is given in closed form based

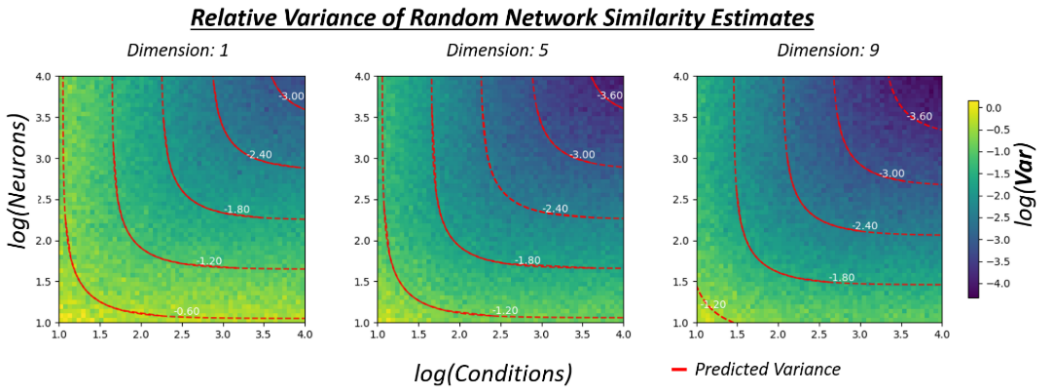

Figure 1: Heatmap of empirical variance in estimating CKD between random neural networks for $d = 1, 5, 9$. Red lines show theoretical prediction from eq. (14).

on prior work (Bietti & Bach 2021). Equation 14 predicts that this quantity decays at a rate proportional $\mathcal{O}(\frac{1}{M} + \frac{1}{N})$ so long as the ground truth distance is not too small: $\mathrm{CKD}^2 \geq \frac{1}{M} + \frac{1}{N}$ up to the multiplicative constants.

Figure 1 shows simulations varying the number of sampled inputs, M, and number of sampled neurons, $N = N_X = N_Y$. In red, we show our predicted rate of decay at a rate of $\mathcal{O}(\frac{1}{M} + \frac{1}{N})$, which show good numerical agreement across a range of input dimensions: the gaps between level curves remains constant across dimensions and the level curves in each heatmap are L-shaped.

4.2 Validation on Trained Deep Artificial Networks and Biological Data

We consider the error in estimating similarities in deep neural networks by training deep convolutional networks on the ImageNet dataset (Deng et al., 2009) and evaluating the variance of their pairwise similarity estimates. For details on architecture and training, see Appendix I.

In Fig. 2A, we verify that our theoretical results still hold when comparing trained deep networks with one another; namely, $\mathbb{V}\mathrm{ar}(\mathrm{CKD})$ still maintains consistently-spaced L-shaped level curves on the heatmap. While we cannot explicitly modify the dimensionality of the input space, we can study this by proxy by presenting different distributions of test images. Evaluating the variance in similarity estimates on test images from the ImageNet dataset (i.e. in-distribution) versus images from the MacaqueITBench dataset (i.e. out-of-distribution; see Appendix H for details) reveals similar scaling for CKD. Hence, in line with our theoretical results, the difficulty in estimating the similarity between two models is largely independent of the dataset over which evaluation occurs or changes in the underlying similarity estimate.

To validate our predictions on biological data, in Fig. 2(B) we plot similar heatmaps when comparing across regions between two monkeys in the THINGS Ventral Stream Dataset (Papale et al. (2025); see Appendix I.3 for sampling details). Again, $\mathbb{V}\mathrm{ar}(\mathrm{CKD})$ follows evenly-spaced L-shaped level curves. We use the different regions as a functional proxy for studying the effect of dimensionality, with the assumption that the representations in different portions of the visual hierarchy likely differ in dimensionality. The scale of variance in all three region-to-region comparisons is consistent, confirming our theoretical predictions that the variance in measuring similarity between neural regions is independent of the specific region identity. In Appendix F, we show that this effect extends even when comparing different regions to one another across monkeys (e.g. V1 to IT), or within an individual monkey.

Last, we use the model-to-model comparisons to verify the signal-adaptivity of our bound in Fig. 2C, where we show close adherence of empirical behavior to the predictions made by our signal-adaptive theory. Notably, when $\frac{1}{M} + \frac{1}{N}$ is sufficiently small, scaling decelerates.

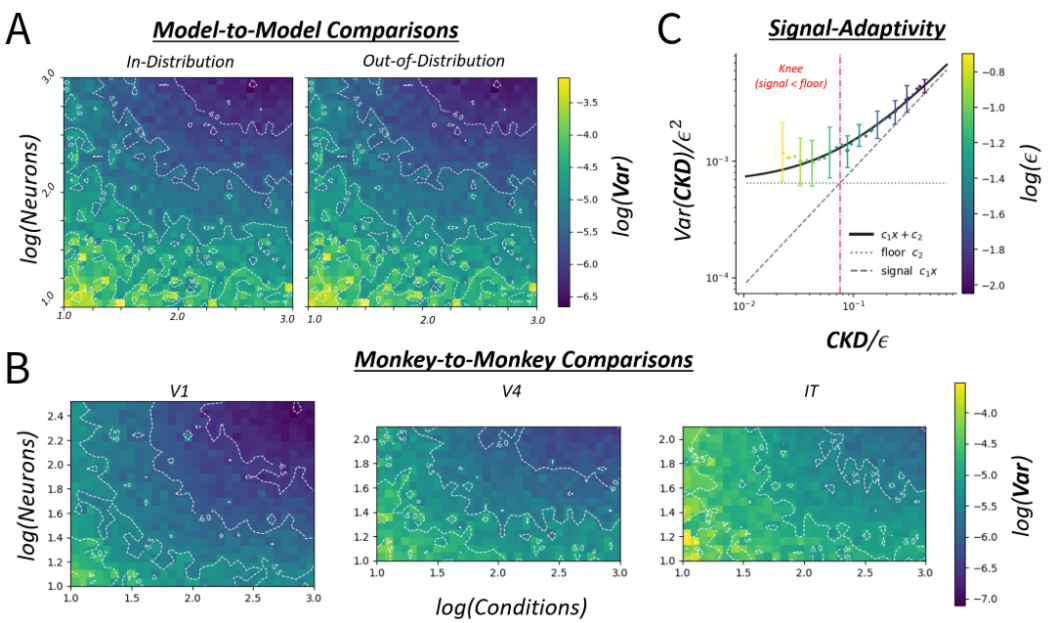

Figure 2: Variance of similarity estimates in deep networks and monkeys. **(A)** Models presented in-distribution versus out-of-distribution test images. Variance scales independent of dataset choice. **(B)** Variance of CKD across three brain regions of two monkeys scales independently from region identity. Multi-units are considered distinct "neurons." **(C)** Representative sample of signal-adaptivity of variance (using model-to-model comparisons), with decreased rate of variance reduction when $\frac{1}{M} + \frac{1}{N}$ small.

4.3 Similarity Estimates across Multiple Systems

Next, we simultaneously study the relationships between representations formed by 3 brain regions replicated across 2 monkey subjects (Papale et al., 2025) and 4 trained convolutional neural networks. We might expect brain regions V1, V4, and IT to have distinct representations because they fall along different parts of the visual hierarchy. However, different sampling regimes may make evaluating this hypothesis impossible. Furthermore, we don't know the exact variability that different neural regions and models exhibit.

In Fig. 3A we produce low-dimensional Euclidean embeddings using multidimensional scaling to visualize the effect of subsampling on the estimated distances among systems (see procedure in Appendix J). Overlaps in ellipsoids indicate probable scenarios where false equivalences may be drawn between two neural systems, while highly elongated ellipsoids indicate probable circumstances where differences between systems would be falsely overstated.

When M, N are small (i.e. $M = 10, N = 10$ in the top-left corner), estimates of distance are noisy and the estimated relationships between neural systems are variable depending on the subsample. Increasing sampling to $M = 1000, N = 33$ significantly reduces this uncertainty, and further by sampling at $M = 1000, N = 100$. Remarkably, however, we can attain *even better* estimates of relationships even when significantly undersampling the number of stimuli ($M = 100$) by simply mildly increasing the number of recorded neurons ($N = 100$). This is in line with our prior prescription that the most efficient way to reduce error in estimating similarity is to keep $M \approx N$, and that oversampling only M but not N (and vice versa) does not significantly improve one's estimate.

This result holds in pre-projection space as well. In Figure 3B, we analyze Kendall's τ measured when subsampling across multiple models. This measures the divergence between the estimated rankings of similarity between multiple models with the true rankings of similarity between models in the pre-projection distance estimates. A larger Kendall's τ indicates greater alignment of similarity ranking between subsampled and true measurements, and thus a more accurate read on the pairwise relationships between models. Similar to our previous results, a large Kendall's τ is at-

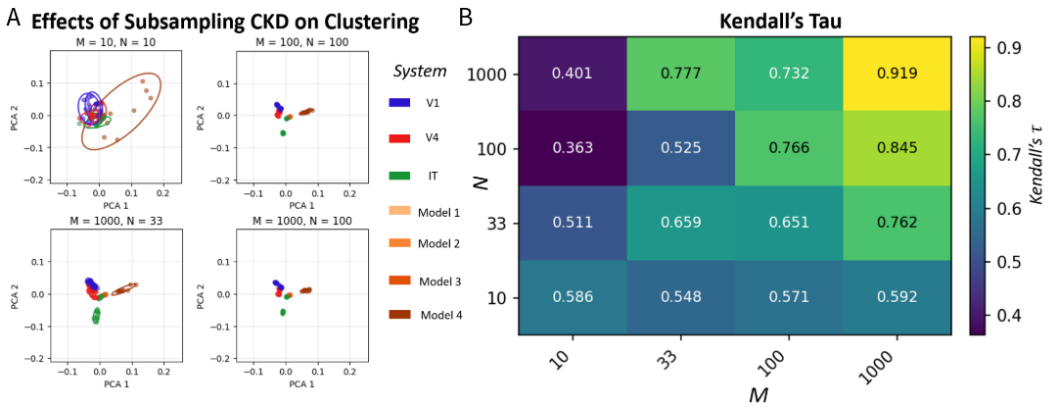

Figure 3: Comparisons between monkeys and similarity analysis between multiple neural systems simultaneously. (**A**) CKD between 4 deep models and 6 neural recordings from 3 different brain areas. Each ellipse denotes an estimated region in PC space of a single neural system's random subsampling at certain values of M and N. (**B**) Estimates of Kendall's Tau on different degrees of subsampling, measuring the error in rankings of similarity between all subjects.

tained when $M = N = 100$, achieving approximately equal or better performance than when only either M or N has a degree of magnitude more samples for marginal trade-off along the other axis.

5 Discussion

We have shown that the CKD exhibits a range of favorable statistical properties with respect to sampling of neurons and of stimuli: it can be estimated without bias, its variance decays with a dimension-independent rate, and it enables principled metric repair and embedding. To our knowledge, no such theory exists to account for the simultaneously subsampling of both stimuli M and neurons N, and our work represents a first step towards resolving this gap in the literature.

Our results on CKD derive from its strong analytic tractability, which other more common measures such as CKA lack (see Appendix D). While CKD is the natural metric associated with CKA, the latter includes a normalization to renders it invariant to scale. Though seemingly a minor difference, this has significant impact on CKA's utilization (lack of unbiased estimator) and theory (difficulties in asymptotic analysis).

Our results have important implications for designing future experiments. For example, we focused on measuring similarity with respect to a single distribution of network inputs, $u_1, \ldots, u_M \sim P$. In practice, one might want to measure similarity across multiple input distributions (e.g. natural images, textures, and artificial gratings). How many can we reasonably fit into a single experiment without cutting our statistical power too thin? These theoretical results providea prescription for *pre-experiment* parametrization to maximize data efficiency. For example, while bootstrapping is a common approach to increase the utility of data *post hoc*, bootstrapping would still be constrained ill-parameterized experiments (e.g. for $M = 1000, N = 10$). Finally, our theory demonstrates that one's ability to estimate CKD from limited samples does not rapidly degrade as a function of task complexity. That is, estimation does not suffer from a curse of dimensionality, unlike related estimation problems in high-dimensional geometry, thus underwriting confidence in similarity estimates for complex stimuli spaces.

Limitations. While CKD's lack of normalization provides it powerful properties for analysis, its sensitivity to the scale of a neural representation may be argued to be an undesirable property (see e.g. Kornblith et al., 2019). Incorporating normalization into analyses of CKD and related measures is an important direction of future research. Simultaneously, more work is required to characterize and improve estimation of metrics like CKA that cannot be formulated as U-statistics.

6 AI DISCLOSURE

In this work, we used generative AI tools (LLMs; Claude Opus 5.5, GPT 5.6 Sol) substantially in developing the mathematical results and providing feedback on research methodology. The authors derived an initial bound on the same order for HSIC was derived without the assistance of LLMs, and the inspiration to formulate our claims via U-statistics was also done without LLMs. LLMs were used to assist in learning the deeper mathematical properties of U-statistics as an accompanying guide to Lee (1990). Afterwards, LLMs assisted in formulating precise statements of technical claims for generalization after the authors had specified the quantity to be analyzed and the kind of result sought (a bound on the same order as the initial approach) and in drafting and revising the written proofs. The authors heavily revised and edited all mathematical proofs and extended the LLM-generated output to cover background material and broader context.

The conceptual framing of the work is due to the human authors. This includes the motivating problem of how similarity estimates depend jointly on the number of sampled neurons and stimuli, the choice to study the Centered Kernel Distance and its unbiased U-statistic estimator, and the decision of which properties to establish. The interpretation of all theoretical and empirical results, including their implications for experimental design in neuroscience and machine learning, is likewise due to the human authors. We did not use generative AI tools to propose hypotheses or design experiments, but LLMs were used to generate code for simulations and aid in formatting figures. Translation and qualitative data analysis are not applicable to this work.

We have reviewed all AI-assisted work. All AI-assisted derivations were checked line-by-line by all the authors. The resulting bounds were independently checked against numerical simulations in random networks, trained networks, and macaque recordings (Section 4); similarly, AI-generated code was checked against our manual implementations and validated by the numerical experiments in these simulations. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 225–236, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

Alberto Bietti and Francis Bach. Deep equals shallow for ReLU networks in kernel regimes. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=aDjoksTpXOP.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

Chanwoo Chun, Abdulkadir Canatar, SueYeon Chung, and Daniel D Lee. Estimating neural representation alignment from sparsely sampled inputs and features. *arXiv preprint arXiv:2502.15104*, 2025.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012.

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

Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

Lyndon Duong, Jingyang Zhou, Josue Nassar, Jules Berman, Jeroen Olieslagers, and Alex H Williams. Representational dissimilarity metric spaces for stochastic neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xjb563TH-GH.

Alessandro Gifford et al. The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies. *arXiv preprint arXiv:2501.00504*, 2025.

Robbe LT Goris, J Anthony Movshon, and Eero P Simoncelli. Partitioning neuronal variability. *Nature neuroscience*, 17(6):858–865, 2014.

John C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In *International Conference on Algorithmic Learning Theory (ALT)*, volume 3734 of *Lecture Notes in Computer Science*, pp. 63–77. Springer, 2005.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

Sarah E. Harvey, Brett W. Larsen, and Alex H. Williams. Duality of Bures and Shape Distances with Implications for Comparing Neural Representations, November 2023. URL http://arxiv.org/abs/2311.11436. arXiv:2311.11436 [stat].

Sarah E. Harvey, David Lipshutz, and Alex H. Williams. What Representational Similarity Measures Imply about Decodable Information, November 2024. URL http://arxiv.org/abs/2411.08197. arXiv:2411.08197 [stat].

Tianxiao He, Alex H Williams, and Sarah E Harvey. How data augmentation shapes neural representations. *arXiv preprint arXiv:2605.15306*, 2026.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Alexis H. Rockter, ¡bos¿5.0 Zheng, Chris I. Baker, and John-Dylan Haynes. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023. doi: 10.7554/eLife.82580. URL https://doi.org.

Nicholas J Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear algebra and its applications*, 103:103–118, 1988.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Hyunmo Kang, Abdulkadir Canatar, and SueYeon Chung. Spectral analysis of representational similarity with limited neurons. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=sfTtFZXINg.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Adam Kohn and Matthew A. Smith. Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *Journal of Neuroscience*, 25(14):3661–3673, 2005.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte and Jörn Diedrichsen. Peeling the onion of brain representations. *Annual review of neuroscience*, 42(1):407–432, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Richard D Lange, Devin Kwok, Jordan K Matelsky, Xinyue Wang, David Rolnick, and Konrad P Kording. Deep networks as paths on the manifold of neural representations. In *Proceedings of the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, volume 221 of *Proceedings of Machine Learning Research*. PMLR, 2023.

A. J. Lee. *U-Statistics: Theory and Practice*, volume 110 of *Statistics: Textbooks and Monographs*. Marcel Dekker, 1990.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2nd edition, 1998.

Spandan Madan, Will Xiao, Mingran Cao, Hanspeter Pfister, Margaret Livingstone, and Gabriel Kreiman. Benchmarking out-of-distribution generalization capabilities of dnn-based encoding models for the ventral visual cortex. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a27eda1ce7c14f4875f15cf3d1737cc5-Abstract-Datasets_and_Benchmarks_Track.html.

Amin Nejatbakhsh, Victor Geadah, Alex Williams, and David Lipshutz. Comparing noisy neural population dynamics using optimal transport distances. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 32970–32996, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/51484744337f4bf5fea0e4dd92ddab0b-Paper-Conference.pdf.

Paolo Papale, Feng Wang, Matthew W. Self, and Pieter R. Roelfsema. An extensive dataset of spiking activity to reveal the syntax of the ventral stream. *Neuron*, 113(3), 2025. doi: 10.1016/j.neuron.2024.12.003.

Adrián Ponce-Alvarez, Alexander Thiele, Thomas D. Albright, Gene R. Stoner, and Gustavo Deco. Stimulus-dependent variability and noise correlations in cortical MT neurons. *Proceedings of the National Academy of Sciences*, 110(32):13162–13167, 2013. doi: 10.1073/pnas.1300098110.

Dean A Pospisil, Brett W Larsen, Sarah E Harvey, and Alex H Williams. Estimating shape distances on neural representations with limited samples. In *International Conference on Learning Representations (ICLR)*, 2024.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128, 2021.

Martin Schrimpf, Jonas Kubilius, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2018. doi: 10.1101/407007. URL https://www.biorxiv.org/content/10.1101/407007v1.

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

Ian H Stevenson and Konrad P Kording. How advances in neural recording affect data analysis. *Nature Neuroscience*, 14(2):139–142, 2011.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Christopher J Cueva, Erin Grant, Iris Groen, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200, 08 2016. doi: 10.1016/j.neuroimage.2015.12.012.

Alex H Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 10–23. PMLR, 2024.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott W Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 4738–4750, 2021.

A U-STATISTIC ESTIMATORS OF HSIC

We begin with a formalization of HSIC as a U-statistic. In the process, we implicitly define a new estimator for HSIC that is unbiased with respect to both M and N. Prior work (Song et al., 2012; Chun et al., 2025) have constructed debiasing efforts for only one of these axes at a time.

Following the construction of such an estimator, we then follow standard arguments on the variance of U-statistics (Lee, 1990) to derive a dimension-free upper bound. In Appendix B, the techniques applied for HSIC will be re-applied, albeit in a more complex fashion, for the CKD estimator we study in the main body of this paper; for those unfamiliar with the theory of U-statistics, this is a useful primer. Arguments as to the effects of noise on our estimator are elaborated upon in Appendix C.

A.1 A FINITE-SAMPLE ESTIMATOR

Plugging eqs. (5) and (6) into the formula for HSIC and expanding leads to a sixteen term expression with many cancellations. Ultimately, the expression simplifies to

$$\langle \bar{k}, \bar{h} \rangle_{\mathrm{HS}} = \mathbb{E}_{u_1,u_2,u_3,u_4}[k(u_1,u_2)h(u_1,u_2) + k(u_1,u_2)h(u_3,u_4) - 2k(u_1,u_2)h(u_1,u_3)] \tag{19}$$

which is equivalent to Equation 3 in Song et al. (2012). Thus, the HSIC estimand can be represented as an expectation over four independent samples, $u_1, u_2, u_3, u_4 \sim P$, of an expression based on the uncentered kernel functions.

Here, we are interested in comparing kernels that are defined by neural tuning curve distributions, $\mathcal{F}$ and $\mathcal{G}$, over the space of functions mapping $\mathcal{U} \mapsto \mathbb{R}$. Specifically, we define

$$k(u_1,u_2) = \mathbb{E}_{f\sim\mathcal{F}}[f(u_1)f(u_2)] \tag{20}$$
$$h(u_1,u_2) = \mathbb{E}_{g\sim\mathcal{G}}[g(u_1)g(u_2)] \tag{21}$$

Now plugging these equations into eq. (19), we obtain

$$\langle \bar{k}, \bar{h} \rangle_{\mathrm{HS}} = \mathbb{E}_{f,g,u_1,u_2,u_3,u_4} \begin{bmatrix} f(u_1)f(u_2)g(u_1)g(u_2) \\ + f(u_1)f(u_2)g(u_3)g(u_4) \\ - 2f(u_1)f(u_2)g(u_1)g(u_3) \end{bmatrix} \tag{22}$$

Let Π denote the set of all $4!$ permutations of $(1,2,3,4)$, and define

$$\phi(f,g,u_1,u_2,u_3,u_4) = \frac{1}{4!} \sum_{(i,j,\ell,p)\in\Pi} \begin{bmatrix} f(u_i)f(u_j)g(u_i)g(u_j) \\ + f(u_i)f(u_j)g(u_\ell)g(u_p) \\ - 2f(u_i)f(u_j)g(u_i)g(u_\ell) \end{bmatrix} \tag{23}$$

Note that ϕ is symmetric in its final four arguments (i.e. invariant to permuting the order of u_1, u_2, u_3, and u_4). It is additionally symmetric in its first two arguments, $\phi(f,g;\cdot) = \phi(g,f;\cdot)$, since swapping f and g induces a permutation over the input indices which is averaged out.

We assume that we have access to N independently sampled neural tuning curves from each system, $f_1,\ldots,f_{N_X} \sim \mathcal{F}$ and $g_1,\ldots,g_{N_Y} \sim \mathcal{G}$, alongside M independently sampled inputs, $u_1,\ldots,u_M \sim P_{\mathcal{U}}$. Then,

$$S_{\mathrm{sc}} = \frac{1}{N_X N_Y} \binom{M}{4}^{-1} \sum_{n=1}^{N_X} \sum_{n'=1}^{N_Y} \sum_{i<j<\ell<p} \phi(f_n, g_{n'}, u_i, u_j, u_\ell, u_p) \tag{24}$$

is an unbiased estimator of HSIC_∞. Indeed, it is a *three-sample U-statistic* of degree $(1,1,4)$: order one in each of the $\mathcal{F}$ and $\mathcal{G}$ samples, and order four in the inputs. Note that here we investigate specifically the stimulus-corrected estimator S_{sc}. Neuron-corrections for bias is necessary when the underlying distributions $\mathcal{F},\mathcal{G}$ are identical (e.g. comparing data from two sessions in the same animal-region). Fortunately, our approach in this section can be similarly applied to a U-statistic for S_{sanc}.

A.2 THE VARIANCE AS A SUM OVER COLLISIONS

The properties of U-statistics have been extensively characterized (Lee, 1990). In particular, we can write down an explicit expression for the variance of the estimator. To do this it is convienent to package a set of inputs into a single symbol:

$$\sigma = \{f_n, g_{n'}, u_i, u_j, u_\ell, u_p\} \tag{25}$$

Then eq. (24) becomes

$$S_{\text{sc}} = \frac{1}{N_X N_Y} \binom{M}{4}^{-1} \sum \phi(\sigma) \tag{26}$$

where the sum is understood to range over all unique tuples. Thus, we can write the variance of the estimator as

$$\mathbb{V}\text{ar}[S_{\text{sc}}] = \mathbb{V}\text{ar}\left[\frac{1}{N_X N_Y} \binom{M}{4}^{-1} \sum \phi(\sigma)\right] = \frac{1}{N_X^2 N_Y^2} \binom{M}{4}^{-2} \sum_{\sigma,\sigma'} \mathbb{C}\text{ov}(\phi(\sigma), \phi(\sigma')) \tag{27}$$

where the sums over σ and σ' independently span over all unique tuples.

Next, we invoke a standard identity for U-statistic covariances (Section 1.4, Theorem 3, Lee, 1990):

$$\mathbb{C}\text{ov}(\phi(\sigma), \phi(\sigma')) = \mathbb{V}\text{ar}_{\sigma\cap\sigma'}\Big[\,\mathbb{E}\big[\,\phi(\sigma)\,\big|\,\sigma\cap\sigma'\,\big]\,\Big]. \tag{28}$$

Intuitively, this relation states that the covariance between $\phi(\sigma)$ and $\phi(\sigma')$ grows with the size of their overlapping elements $\sigma\cap\sigma'$. The inner expectation denotes partial averaging—all variables that are *not overlapping* are averaged out, and the variance is computed with to respect to the remaining terms. When $\sigma \cap \sigma' = \emptyset$ (no overlap), the inner expectation is with respect to all variables in σ and evaluates to a constant; thus, we have:

$$\sigma \cap \sigma' = \emptyset \quad \Rightarrow \quad \mathbb{C}\text{ov}(\phi(\sigma), \phi(\sigma')) = 0. \tag{29}$$

In the other extreme, when $S = S'$ (full overlap), nothing is averaged out in the inner expectation and we recover $\mathbb{C}\text{ov}(\phi(\sigma), \phi(\sigma')) = \mathbb{V}\text{ar}[\phi(\sigma)]$.

To proceed, we enumerate all possible overlap patterns between σ and σ' in the sum appearing in eq. (27). We use the term "collison" to denote the event that one of the variables is part of the overlap and we introduce three index variables to account for these collisions:

$$\alpha \in \{0,1\} \;=\; \#\text{ of collisions in } \mathcal{F} \text{ samples} \tag{30}$$

$$\beta \in \{0,1\} \;=\; \#\text{ of collisions in } \mathcal{G} \text{ samples} \tag{31}$$

$$\gamma \in \{0,1,2,3,4\} \;=\; \#\text{ of collisions in input samples} \tag{32}$$

For example, $\sigma = (f_1, g_1, u_1, u_2, u_3, u_4)$ and $\sigma' = (f_1, g_2, u_2, u_3, u_4, u_5)$ corresponds to a overlap pattern of $(\alpha, \beta, \gamma) = (1, 0, 3)$; one collision in the $\mathcal{F}$ neural sample, no collisions in the $\mathcal{G}$ neural sample, and three collisions in the input samples. Note that the *position* of the four input samples can be ignored since ϕ is a symmetric function by our construction in eq. (23).

Now let $C_{\alpha\beta\gamma}$ denote the number of collisions of each type occuring in eq. (27), and let

$$V_{\alpha\beta\gamma} = \mathbb{C}\text{ov}(\phi(\sigma), \phi(\sigma')) = \mathbb{V}\text{ar}_{\sigma\cap\sigma'}\Big[\,\mathbb{E}\big[\,\phi(\sigma)\mid\sigma\cap\sigma'\,\big]\,\Big], \tag{33}$$

denote the variance of each corresponding term. Note that the variance only depends on the number of collisions in each block of variables due to the i.i.d. assumption of samples and the blockwise symmetry of ϕ. Thus, we can compactly re-express the variance of the estimator as:

$$\mathbb{V}\text{ar}[S_{\text{sc}}] = \frac{1}{N_X^2 N_Y^2} \binom{M}{4}^{-2} \sum_{\alpha,\beta,\gamma} C_{\alpha\beta\gamma} V_{\alpha\beta\gamma} \tag{34}$$

where the sum ranges over all $2 \times 2 \times 5 = 20$ possible collision patterns. Note that, for large N and M, the vast majority of terms will have no collisions, and these terms do not contribute to the variance since $V_{000} = 0$ as already established in eq. (29).

A.3 COUNTING COLLISIONS

We now derive an explicit expression for $C_{\alpha\beta\gamma}$ using some basic combinatorial calculations. For each block of variables ($\mathcal{F}$ neurons, $\mathcal{G}$ neurons, and inputs), both σ and σ' draw a subset of fixed size ν without replacement from a population of P elements ($\nu = 1$, $P = N$ for each neuron sample; $\nu = 4$, $P = M$ for the inputs). Conditioned on whichever ν elements are drawn by σ, the overlap follows a hypergeometric distribution—i.e. the number of times one of the ν elements in σ is drawn by σ' for its own ν draws without replacement. The number of ordered pairs of subsets that overlap in exactly c elements is therefore

$$\underbrace{\binom{P}{\nu}}_{S\text{'s subset}} \quad \underbrace{\binom{\nu}{c}\binom{P-\nu}{\nu-c}}_{\text{hypergeometric weight for } S'}, \tag{35}$$

where $\binom{\nu}{c}$ counts the marked elements that σ' reuses and $\binom{P-\nu}{\nu-c}$ the remaining ones it draws from the unmarked population; dividing through by $\binom{P}{\nu}$ recovers the hypergeometric probability mass function. Multiplying eq. (35) across the three samples—with $(\nu, P, c) = (1, N, \alpha)$ for $\mathcal{F}$, $(1, N, \beta)$ for $\mathcal{G}$, and $(4, M, \gamma)$ for the inputs—yields

$$C_{\alpha\beta\gamma} = \underbrace{\binom{N_X}{1}\binom{1}{\alpha}\binom{N_X-1}{1-\alpha}}_{\mathcal{F}} \underbrace{\binom{N_Y}{1}\binom{1}{\beta}\binom{N_Y-1}{1-\beta}}_{\mathcal{G}} \underbrace{\binom{M}{4}\binom{4}{\gamma}\binom{M-4}{4-\gamma}}_{\text{inputs}}. \tag{36}$$

This simplifies simply to

$$C_{\alpha\beta\gamma} = N_X(N_X-1)^{1-\alpha}\, N_Y(N_Y-1)^{1-\beta} \binom{M}{4}\binom{4}{\gamma}\binom{M-4}{4-\gamma}. \tag{37}$$

It is helpful to normalize $C_{\alpha\beta\gamma}$ by the number of terms in the sum. Thus, we have

$$\mathbb{V}\text{ar}[S_{\text{sc}}] = \sum_{\alpha,\beta,\gamma} \widetilde{C}_{\alpha\beta\gamma} V_{\alpha\beta\gamma} \tag{38}$$

where

$$\widetilde{C}_{\alpha\beta\gamma} = \frac{1}{N_X^2 N_Y^2}\binom{M}{4}^{-2} C_{\alpha\beta\gamma} = \underbrace{\frac{(N_X-1)^{1-\alpha}}{N_X}}_{\mathcal{F}} \underbrace{\frac{(N_Y-1)^{1-\beta}}{N_Y}}_{\mathcal{G}} \underbrace{\frac{\binom{4}{\gamma}\binom{M-4}{4-\gamma}}{\binom{M}{4}}}_{\text{inputs}}. \tag{39}$$

Each factor is the probability mass function of a hypergeometric distribution, so $\tilde{C}_{\alpha\beta\gamma}$ is the probability that two independently drawn tuples collide in pattern (α, β, γ); in particular $\sum_{\alpha,\beta,\gamma} \tilde{C}_{\alpha\beta\gamma} = 1$, and the variance is the expected collision covariance, $\mathbb{V}\text{ar}[S_{\text{sc}}] = \mathbb{E}[V_{\alpha\beta\gamma}]$.

Next, we assume that the random variables $f(u)$ (over $f \sim \mathcal{F}$) and $g(u)$ (over $g \sim \mathcal{G}$) are sub-Gaussian. This assumption is desired because it guarantees that their fourth moments are bounded, $\mathbb{E}_f[f(u)^4] \le C\sigma_{\mathcal{F}}^4$ and $\mathbb{E}_g[g(u)^4] \le C\sigma_{\mathcal{G}}^4$, uniformly in u (overloading notation somewhat, and definition $\sigma_{\mathcal{F}}^4, \sigma_{\mathcal{G}}^4$ as the kurtosis of their respective distributions). Note that this is a looser restriction than bounded firing rates, which are guaranteed in both biological and computational networks, so we do not restrict ourselves of consideration in any practical settings. We show that the variance decreases at a rate proportional to $1/M + 1/N$ with a multiplicative constant that is independent of the dimensionality of the inputs or smoothness of the tuning curves.

A.4 VARIANCE BOUND ON HSIC

First, we note that

$$V_{\alpha\beta\gamma} \le V_{114} = \mathbb{V}\text{ar}[\phi(\sigma)], \tag{40}$$

for all collision patterns indexed by (α, β, γ). This is because $V_{\alpha\beta\gamma} = \mathbb{V}\text{ar}_{\sigma\cap\sigma'}\left[\mathbb{E}[\phi(\sigma) \mid \sigma \cap \sigma']\right]$ averages out non-shared arguments in the inner expectation, and such partial averaging can only decrease the variance.

The core idea is to bound $V_{114} = \mathbb{V}\mathrm{ar}[\phi(\sigma)] \leq \mathbb{E}[\phi(\sigma)^2] = \|\phi\|_2^2$, where $\|X\|_2 := \mathbb{E}[X^2]^{1/2}$ denotes the L^2 norm of a scalar random variable. Using the triangle inequality associated with this norm, on both the average over permutations in eq. (23) and the three terms with coefficients $(1, 1, -2)$, we have

$$\|\phi\|_2 \leq 4 \max_{a \neq b,\ c \neq d} \left\| f(u_a) f(u_b)\, g(u_c) g(u_d) \right\|_2, \tag{41}$$

where the maximum is over the finitely many index patterns appearing in ϕ (the indices a, b need not be distinct from c, d). Unlike a maximum over permutations inside the norm, this is a maximum of fixed numbers, so no worst-case permutation needs to be selected.

Now fix any such pattern and condition on the inputs. Due to independence of $f \sim \mathcal{F}$ and $g \sim \mathcal{G}$, we have:

$$\mathbb{E}_{f,g}\left[f(u_a)^2 f(u_b)^2\, g(u_c)^2 g(u_d)^2 \right] = \mathbb{E}_f[f(u_a)^2 f(u_b)^2] \cdot \mathbb{E}_g[g(u_c)^2 g(u_d)^2] \tag{42}$$

Each term is bounded by Cauchy–Schwarz, which we can then bound due to the sub-Gaussian assumption on the tuning curves:

$$\mathbb{E}_f[f(u_a)^2 f(u_b)^2] \leq \sqrt{\mathbb{E}_f[f(u_a)^4]\, \mathbb{E}_f[f(u_b)^4]} \leq C\sigma_{\mathcal{F}}^4, \tag{43}$$

and likewise $\mathbb{E}_g[g(u_c)^2 g(u_d)^2] \leq C\sigma_{\mathcal{G}}^4$. As these bounds hold for every realization of the inputs, averaging over the inputs gives $\|f(u_a) f(u_b)\, g(u_c) g(u_d)\|_2^2 \leq C^2 \sigma_{\mathcal{F}}^4 \sigma_{\mathcal{G}}^4$, so we conclude that

$$V_{114} = \mathbb{V}\mathrm{ar}[\phi(\sigma)] \leq \mathbb{E}[\phi(\sigma)^2] \leq 16\, C^2 \sigma_{\mathcal{F}}^4\, \sigma_{\mathcal{G}}^4 = C'\, \sigma_{\mathcal{F}}^4\, \sigma_{\mathcal{G}}^4 \tag{44}$$

for a universal constant C (absorbing the numerical factors).

Now, starting from eq. (38) and the fact that $V_{000} = 0$, we have

$$\mathbb{V}\mathrm{ar}[S_{\mathrm{sc}}] = \sum_{(\alpha,\beta,\gamma) \neq 0} \widetilde{C}_{\alpha\beta\gamma} V_{\alpha\beta\gamma} \tag{45}$$

$$= \frac{1}{N_X} V_{100} + \frac{1}{N_Y} V_{010} + \frac{16}{M} V_{001} + \mathcal{O}\left(\left(\frac{1}{N_X} + \frac{1}{N_Y} + \frac{1}{M} \right)^2 \right) \tag{46}$$

$$\lesssim \left(\frac{1}{N_X} + \frac{1}{N_Y} + \frac{16}{M} \right) V_{114} = C\, \sigma_{\mathcal{F}}^4\, \sigma_{\mathcal{G}}^4 \left(\frac{1}{N_X} + \frac{1}{N_Y} + \frac{16}{M} \right) \tag{47}$$

B Theoretical results on CKD^2

B.1 Maximum Mean Discrepancy

CKD is a special case of the maximum mean discrepancy (MMD; Gretton et al., 2012) with quadratic kernel $\kappa(f, g) = \langle \bar{f}, \bar{g} \rangle^2$, where $\bar{f}, \bar{g}$ represent the mean-subtracted (i.e. centered) tuning curves and the inner product is $L_2(P_{\mathcal{U}})$. Given any kernel κ, its MMD is defined as

$$\mathrm{MMD}^2(\mathcal{F}, \mathcal{G}) = \mathbb{E}\Big[\kappa(f, f')\Big] + \mathbb{E}\Big[\kappa(g, g')\Big] - 2\mathbb{E}\Big[\kappa(f, g)\Big]$$

with tuning curves $f, f' \sim \mathcal{F}, g, g' \sim \mathcal{G}$ drawn independently from their respective distributions. Hence, we get $\langle \bar{f} \otimes \bar{f}, \bar{g} \otimes \bar{g} \rangle_{HS} = \langle \bar{f}, \bar{g} \rangle^2 = \kappa(f, g)$ and $\mathrm{MMD}^2 = \mathrm{CKD}^2$.

B.2 The CKD Estimator

The main difference with the CKD estimator is the two self-terms S_{sanc} involved. The self terms require a small modification: $\langle \bar{k}, \bar{k} \rangle$ is estimated by feeding ϕ *two independent neurons from the same system*,

$$\mathbb{E}_{f_1, f_2 \sim \mathcal{F},\, u_{1:4}}\left[\phi(f_1, f_2, u_{1:4}) \right] = \langle \bar{k}, \bar{k} \rangle, \tag{48}$$

which holds because ϕ is unbiased for $\langle \bar{k}, \bar{h} \rangle$ whenever its two neuron arguments are independent.

Collecting the three pieces, and symmetrizing the cross term over which neuron is drawn from each system, define the kernel

$$\psi(f_1, f_2, g_1, g_2, u_{1:4}) = \phi(f_1, f_2, u_{1:4}) + \phi(g_1, g_2, u_{1:4}) - \frac{1}{2}\sum_{a=1}^{2}\sum_{b=1}^{2}\phi(f_a, g_b, u_{1:4}), \tag{49}$$

which satisfies $\mathbb{E}[\psi] = \langle \bar{k}, \bar{k}\rangle - 2\langle \bar{k}, \bar{h}\rangle + \langle \bar{h}, \bar{h}\rangle = \mathrm{CKD}^2$ and is symmetric within each block of arguments (the two $\mathcal{F}$-neurons, the two $\mathcal{G}$-neurons, and the four inputs). Given $N \geq 2$ neurons from each system and M inputs, the corresponding U-statistic is

$$\widehat{\mathrm{CKD}}^2 = \binom{N}{2}^{-2}\binom{M}{4}^{-1}\sum_{n_1<n_2}\sum_{n_1'<n_2'}\sum_{i<j<\ell<p}\psi(f_{n_1}, f_{n_2}, g_{n_1'}, g_{n_2'}, u_i, u_j, u_\ell, u_p), \tag{50}$$

an unbiased estimator of CKD^2. It is a three-sample U-statistic of degree $(2, 2, 4)$: order two in each neuron sample—one higher than for HSIC, since each self term reuses its system—and order four in the inputs.

We can efficiently compute our estimator $\widehat{\mathrm{CKD}}^2$ in practice by leveraging the expression of S_{sc} in eq. (11), due to Song et al. (2012). In particular, every term appearing on the right hand side of eq. (11) is separately linear in $\widehat{\boldsymbol{K}}$ and $\widehat{\boldsymbol{H}}$. Thus, setting the two inputs to the estimator equal to $\widehat{\boldsymbol{K}} = N_X^{-1}\sum_n \boldsymbol{x}_n\boldsymbol{x}_n^\top$ and rearranging, we get:

$$S_{\mathrm{sanc}}(\boldsymbol{X}, \boldsymbol{X}) = \frac{1}{N_X^2}\sum_{n=1}^{N_X}\sum_{n'=1}^{N_X} S_{\mathrm{pi}}(\boldsymbol{x}_n\boldsymbol{x}_n^\top, \boldsymbol{x}_{n'}\boldsymbol{x}_{n'}^\top). \tag{51}$$

We need to remove the diagonal terms where $n = n'$. Subtracting off these terms and rescaling to achieve the proper normalization we get

$$S_{\mathrm{sanc}}(\boldsymbol{X}, \boldsymbol{X}) = \frac{1}{N_X(N_X-1)}\left(N_X^2 S_{\mathrm{sanc}}(\widehat{\boldsymbol{K}}, \widehat{\boldsymbol{K}}) - \sum_n S_{\mathrm{sanc}}(\boldsymbol{x}_n\boldsymbol{x}_n^\top, \boldsymbol{x}_n\boldsymbol{x}_n^\top)\right) \tag{52}$$

To efficiently calculate S_{sanc} between two rank one matrices, we do not form the $M \times M$ matrices explicitly. Instead, one can verify that

$$S_{\mathrm{sanc}}(\boldsymbol{v}\boldsymbol{v}^\top, \boldsymbol{v}\boldsymbol{v}^\top) = \frac{1}{M(M-3)}\left[(z_2^2 - z_4) + \frac{(z_1^2 - z_2)^2}{(M-1)(M-2)} - \frac{2\left(z_1^2 z_2 + z_4 - 2z_1 z_3\right)}{M-2}\right], \tag{53}$$

for any vector $\boldsymbol{v} \in \mathbb{R}^M$ and $z_1 = \sum_i v_i$, $z_2 = \sum_i v_i^2$, $z_3 = \sum_i v_i^3$, and $z_4 = \sum_i v_i^4$. Thus, computing S_{sanc} on two identical rank-one matrices requires only $O(M)$ calculations. Repeating this calculation across all neurons in eq. (52) requires only $O(MN_X)$ computation.

B.3 The variance as a sum over collisions

The variance decomposes exactly as in eqs. (27) to (39), now for the degree-$(2, 2, 4)$ kernel ψ. The only change is that the two neuron samples now contribute degree two, so the collision indices run over

$$\alpha, \beta \in \{0, 1, 2\}, \qquad \gamma \in \{0, 1, 2, 3, 4\}, \tag{54}$$

giving a $3 \times 3 \times 5$ table of patterns. Writing $V_{\alpha\beta\gamma}$ for the projection variance of each pattern (with $V_{000} = 0$ as before),

$$\mathbb{V}\mathrm{ar}[\widehat{\mathrm{CKD}}^2] = \sum_{\alpha,\beta,\gamma}\tilde{C}_{\alpha\beta\gamma}V_{\alpha\beta\gamma}, \tag{55}$$

where $\tilde{C}_{\alpha\beta\gamma}$ is again a product of three hypergeometric probabilities, now with degree $d = 2$ for each neuron sample,

$$\tilde{C}_{\alpha\beta\gamma} = \frac{\binom{2}{\alpha}\binom{N_X-2}{2-\alpha}}{\binom{N_X}{2}} \cdot \frac{\binom{2}{\beta}\binom{N_Y-2}{2-\beta}}{\binom{N_Y}{2}} \cdot \frac{\binom{4}{\gamma}\binom{M-4}{4-\gamma}}{\binom{M}{4}}. \tag{56}$$

The leading single-collision probabilities again follow the hypergeometric mean overlap d^2/P of two degree-d draws from a population of size P: a single neuron collision occurs with probability $\tilde{C}_{100} \to 4/N_X, \tilde{C}_{010} \to 4/N_Y$ (degree $d = 2$), and a single input collision with $\tilde{C}_{001} \to 16/M$ (degree $d = 4$). The neuron rate is thus raised from the $1/N$ of the preceding section to $4/N$ (for both N_X and N_Y.

B.4 Proof of theorem 3.1

For each neuron, write $\bar{f} = \mathbb{E}_u[f(u)]$ and $\bar{g} = \mathbb{E}_u[g(u)]$ for its stimulus-averaged response. Then consider the expansion of $\mathbb{E}_{u,u'}[(f(u) - \bar{f})(f(u') - \bar{f})\, h(u, u')]$, given by

$$\mathbb{E}_{u,u'}[f(u)f(u')h(u,u')] + \bar{f}^2\, \mathbb{E}_{u,u'}[h(u,u')] - 2\bar{f}\, \mathbb{E}_{u,u'}[f(u)h(u,u')],$$

where the cross terms combine by the symmetry of h. Substituting $h(u, u') = \mathbb{E}_g[g(u)g(u')]$ and $\bar{f} = \mathbb{E}_u[f(u)]$ into the right hand side and pulling out the expectation, we get:

$$\mathbb{E}_{g,\, u_{1:4}}\big[f(u_1)f(u_2)g(u_1)g(u_2) + f(u_1)f(u_2)g(u_3)g(u_4) - 2f(u_1)f(u_2)g(u_1)g(u_3)\big] = \mathbb{E}[\phi(\sigma) \mid f],$$

where $\sigma = (f_n, g_{n'}, u_i, u_j, u_\ell, u_p)$ denotes a set of inputs to ϕ, as in Appendix A. The equality holds because the inputs are i.i.d., so the permutation average in eq. (23) does not change the expectation.

Since $f(u) - \bar{f}$ has zero mean over u, the centering terms of $\bar{h}$ integrate against it to zero, so h may be replaced by $\bar{h}$:

$$\mathbb{E}[\phi(\sigma) \mid f] = \mathbb{E}_{u,u'}\big[(f(u) - \bar{f})(f(u') - \bar{f})\, \bar{h}(u,u')\big] = \langle \bar{k}_f, \bar{h} \rangle, \tag{57}$$

where $\bar{k}_f(u,u') = (f(u) - \bar{f})(f(u') - \bar{f})$ is the rank-one centered kernel of the single neuron f. The same holds when the other neuron is drawn from any distribution, with $\bar{h}$ replaced by that distribution's centered kernel.

Turning to the CKD estimator, we use $\Sigma = (f_{n_1}, f_{n_2}, g_{n'_1}, g_{n'_2}, u_i, u_j, u_\ell, u_p)$ to denote a set of inputs to the U-statistic kernel ψ, defined in eq. (49). We write Σ and Σ' as a given pair of these inputs, analogous to how σ and σ' denote a pair of inputs to HSIC U-statistic kernel ϕ.

Consider V_{100}, the covariance of $\psi(\Sigma)$ and $\psi(\Sigma')$ when they share exactly one $\mathcal{F}$-neuron f; by eq. (33), this equals $\mathbb{V}\mathrm{ar}_f\big[\mathbb{E}[\psi(\Sigma) \mid f]\big]$. We condition ψ on f and average over the remaining neurons and inputs, applying eq. (57) to each of the three terms of eq. (49). The self-$\mathcal{F}$ term projects onto $\bar{k}$ (the population kernel of the second $\mathcal{F}$-neuron), the self-$\mathcal{G}$ term does not involve f at all, and among the four cross terms only the two that contain f retain it:

$$\begin{aligned}
\mathbb{E}[\psi(\Sigma) \mid f] &= \underbrace{\langle \bar{k}_f, \bar{k} \rangle}_{\phi(f,f_2)} + \underbrace{\langle \bar{h}, \bar{h} \rangle}_{\phi(g_1,g_2)} - \frac{1}{2}\Big(\underbrace{2\langle \bar{k}_f, \bar{h} \rangle}_{f_a = f} + \underbrace{2\langle \bar{k}, \bar{h} \rangle}_{f_a = f_2} \Big) \\
&= \langle \bar{k}_f,\, \bar{k} - \bar{h} \rangle + \langle \bar{h}, \bar{h} \rangle - \langle \bar{k}, \bar{h} \rangle \\
&= \langle \bar{k}_f, \bar{k} - \bar{h} \rangle + \text{const.}
\end{aligned} \tag{58}$$

The terms constant in f do not affect the variance, leaving only the gap-dependent part. This projection is *linear* in $\bar{k} - \bar{h}$, so by Cauchy–Schwarz

$$V_{100} = \mathbb{V}\mathrm{ar}_f\big[\langle \bar{k}_f, \bar{k} - \bar{h} \rangle\big] \le \mathbb{E}_f\big[\|\bar{k}_f\|^2\big]\, \|\bar{k} - \bar{h}\|^2 = \mathbb{E}_f\big[\|\bar{k}_f\|^2\big]\, \mathrm{CKD}^2, \tag{59}$$

and likewise for V_{010} with the roles of the $\mathcal{F}$- and $\mathcal{G}$-neurons exchanged. For a shared input, the three terms of ψ instead complete the square:

$$\mathbb{E}[\psi(\Sigma) \mid u] = \frac{1}{2}\mathbb{E}_{u'}\big[(\bar{k} - \bar{h})(u,u')^2\big] + \text{const.} \tag{60}$$

This is quadratic in $\bar{k} - \bar{h}$, so that $V_{001} = \frac{1}{4}\mathbb{V}\mathrm{ar}_u\big[\mathbb{E}_{u'}(\bar{k} - \bar{h})(u,u')^2\big]$.

We now bound each contribution explicitly in terms of the *scale* s and *kurtosis* κ of the neural responses by

$$s^2 = \max\Big\{ \sup_u \mathbb{E}_f\big[(f(u) - \bar{f})^2\big],\ \sup_u \mathbb{E}_g\big[(g(u) - \bar{g})^2\big] \Big\}, \tag{61}$$

$$\kappa = \max\left\{ \sup_u \frac{\mathbb{E}_f[(f(u)-\bar{f})^4]}{\mathbb{E}_f[(f(u)-\bar{f})^2]^2},\ \sup_u \frac{\mathbb{E}_g[(g(u)-\bar{g})^4]}{\mathbb{E}_g[(g(u)-\bar{g})^2]^2} \right\}, \tag{62}$$

where $f \sim \mathcal{F}$ and $g \sim \mathcal{G}$. Thus s^2 is the largest, over stimuli, mean-square deviation of a neuron's response from its own stimulus-averaged rate, and $\kappa \geq 1$ is finite whenever the responses have finite kurtosis. Two consequences are used below. First, $\mathbb{E}_f[(f(u)-\bar{f})^4] \leq \kappa s^4$ for every u, and likewise for g. Second, since $\bar{k}(u,u') = \mathbb{E}_f[(f(u)-\bar{f})(f(u')-\bar{f})]$, Cauchy–Schwarz gives $|\bar{k}(u,u')| \leq s^2$, and likewise $|\bar{h}(u,u')| \leq s^2$.

By eq. (59),

$$V_{100} \leq \mathbb{E}_f\Big[\big(\mathbb{E}_u(f(u)-\bar{f})^2\big)^2\Big]\,\mathrm{CKD}^2 \leq \mathbb{E}_u\mathbb{E}_f\big[(f(u)-\bar{f})^4\big]\,\mathrm{CKD}^2 \leq \kappa s^4\,\mathrm{CKD}^2,$$

and likewise $V_{010} \leq \kappa s^4\,\mathrm{CKD}^2$. For the input collision,

$$(\bar{k}-\bar{h})(u,u')^2 \leq 2\bar{k}(u,u')^2 + 2\bar{h}(u,u')^2 \leq 4s^4. \tag{63}$$

Therefore,

$$\begin{aligned} V_{001} &\leq \tfrac{1}{4}\,\mathbb{E}_u\Big[\big(\mathbb{E}_{u'}(\bar{k}-\bar{h})(u,u')^2\big)^2\Big] \\ &\leq \tfrac{1}{4}\Big(\sup_u \mathbb{E}_{u'}(\bar{k}-\bar{h})(u,u')^2\Big)\,\mathbb{E}_{u,u'}\big[(\bar{k}-\bar{h})(u,u')^2\big] \\ &\leq s^4\,\mathrm{CKD}^2 \end{aligned}$$

It remains to control the higher-order collision terms. As before, we use a crude bound: $\mathbb{V}\mathrm{ar}[\psi] = V_{224}$ is the maximal $V_{\alpha,\beta,\gamma}$ term since conditioning on fewer shared arguments can only reduce the variance of the projection. V_{224} is the covariance of $\psi(\Sigma)$ and $\psi(\Sigma')$ when $\Sigma = \Sigma'$. Adding a constant to either neuron's responses leaves ϕ unchanged, because the shift adds to $f(u_i)f(u_j)$ only terms depending on at most one of the two inputs, and the coefficients $(1,1,-2)$ cancel these under the permutation average. We may therefore evaluate ψ on the stimulus-centered responses $f-\bar{f}$ and $g-\bar{g}$, so that every summand of ϕ in eq. (23) has the form $a(u_i)a(u_j)\,b(u_\ell)b(u_p)$ for two *independent* centered neurons a, b. Conditioning on the inputs and applying Cauchy–Schwarz to each neuron separately gives $\mathbb{E}[a(u_i)^2a(u_j)^2\,b(u_\ell)^2b(u_p)^2] \leq (\kappa s^4)^2$.

The triangle inequality over the permutation average and the coefficients $(1,1,-2)$ then gives $\|\phi(a,b,\cdot)\|_2 \leq 4\kappa s^4$. Hence, replacing the terms in eq. (49),

$$\|\psi\|_2 \leq 4\kappa s^4 + 4\kappa s^4 + \tfrac{1}{2}\cdot 4\cdot 4\kappa s^4 = 16\,\kappa s^4,$$

and thus $V_{\alpha\beta\gamma} \leq \mathbb{V}\mathrm{ar}[\psi] \leq 256\,\kappa^2 s^8$.

We complete the proof by applying a counting argument on the higher order terms (i.e. Σ, Σ' pairs with at least *two* overlapping terms). For ease of notation, let $\epsilon := \frac{1}{M} + \frac{1}{N_X} + \frac{1}{N_Y}$. Next, let the number of overlapping terms be $T \geq 2$. Single-collision patterns occur with probability $\tilde{C}_{100} \leq 4/N_X$, $\tilde{C}_{010} \leq 4/N_Y$, and $\tilde{C}_{001} \leq 16/M$.

Next, note that a hypergeometric overlap of two size-δ draws from P items has $\mathbb{E}[X(X-1)] = \delta^2(\delta-1)^2/(P(P-1)) \leq 2(\mathbb{E}X)^2$. In our case and $\mathbf{1}\{T \geq 2\} \leq T(T-1)/2$, this gives

$$\Pr(T \geq 2) \leq \Big(\tfrac{4}{N_X} + \tfrac{4}{N_Y} + \tfrac{16}{M}\Big)^2 \leq 256\,\epsilon^2.$$

Combining these bounds,

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\widehat{\mathrm{CKD}}^2] &\leq \tfrac{4}{N_X}V_{100} + \tfrac{4}{N_Y}V_{010} + \tfrac{16}{M}V_{001} + \Pr(T \geq 2)\,\mathbb{V}\mathrm{ar}[\psi] \\ &\leq 16\,\kappa s^4\,\mathrm{CKD}^2\,\epsilon \;+\; 2^{16}\,\kappa^2 s^8\,\epsilon^2. \end{aligned} \tag{64}$$

Since $\kappa \geq 1$, this implies eq. (14) with $C = 2^{16}\kappa^2\max(s^4, s^8)$. The numerical constants have not been optimized.

C TRIAL-TO-TRIAL NOISE

In this section, we analyze how the U-statistic estimator of CKD^2 behaves when neural responses are corrupted by trial-to-trial noise. We assume that noise is independent across neurons and trials. Relaxing this assumption is nontrivial, and we therefore leave it to future work.

In this appendix section, we show the following. First, we show that the estimator remains unbiased (Section C.2). Second, we show that the estimator's variance is insensitive to first-order: the leading rate remains $1/M+1/N_X+1/N_Y$ and the constants V_{001}, V_{010}, and V_{100} that scale each term remain the same (Section C.3). Finally, we discuss how this picture changes if we relax the assumption that noise is decorrelated across neurons.

C.1 SAMPLING MODEL WITH NOISE

We denote noise variables by ξ (system X) and ζ (system Y). For a tuning curve f we write $\bar{f} := \mathbb{E}_u[f(u)]$.

Neurons and inputs are sampled as in Section 2.1: $f_1, \dots, f_{N_X} \sim \mathcal{F}$, $g_1, \dots, g_{N_Y} \sim \mathcal{G}$, and $u_1, \dots, u_M \sim P$, all independent. We collect these noiseless sampling variables into

$$\mathcal{D} := (f_{1:N_X},\, g_{1:N_Y},\, u_{1:M}).$$

Each input is presented $L \geq 1$ times. On repetition ℓ we observe

$$X_{ni\ell} = f_n(u_i) + \xi_{ni\ell}, \qquad Y_{ni\ell} = g_n(u_i) + \zeta_{ni\ell}. \tag{65}$$

We make the following assumptions.

- (N1) *Conditional mean zero.* $\mathbb{E}[\xi_{ni\ell} \mid \mathcal{D}] = 0$ and $\mathbb{E}[\zeta_{ni\ell} \mid \mathcal{D}] = 0$.
- (N2) *Conditional independence.* Conditional on $\mathcal{D}$, all noise variables $\{\xi_{ni\ell}\}$ and $\{\zeta_{ni\ell}\}$ are mutually independent across neurons, inputs, repetitions, and systems. Moreover, the conditional law of $\xi_{ni\ell}$ depends on $\mathcal{D}$ only through (f_n, u_i), and likewise for $\zeta_{ni\ell}$ through (g_n, u_i).
- (N3) *Bounded conditional variance.* $\mathrm{Var}(\xi_{ni\ell} \mid \mathcal{D}) = s_X^2(f_n, u_i) \leq \sigma_X^2$ and $\mathrm{Var}(\zeta_{ni\ell} \mid \mathcal{D}) = s_Y^2(g_n, u_i) \leq \sigma_Y^2$.

Only second moments of the noise are required and heteroscedastic noise is allowed. The estimator is computed from trial-averaged responses

$$\bar{X}_{ni} := \frac{1}{L}\sum_{\ell=1}^{L} X_{ni\ell} = f_n(u_i) + \bar{\xi}_{ni}, \qquad \bar{Y}_{ni} := \frac{1}{L}\sum_{\ell=1}^{L} Y_{ni\ell} = g_n(u_i) + \bar{\zeta}_{ni}. \tag{66}$$

The averaged noise $\bar{\xi}_{ni}$ satisfies (N1)–(N2) with conditional variance $s_X^2(f_n, u_i)/L$, and similarly for $\bar{\zeta}_{ni}$.

C.2 THE U-STATISTIC ESTIMATOR REMAINS UNBIASED UNDER NOISE

Recall from appendix A and appendix B that,

$$\phi(f, g; u_{1:4}) = \frac{1}{4!}\sum_{\pi \in S_4} \begin{bmatrix} f(u_{\pi_1})f(u_{\pi_2})g(u_{\pi_1}g(u_{\pi_2}) \\ + f(u_{\pi_1})f(u_{\pi_2})g(u_{\pi_3})g(u_{\pi_4}) \\ - 2f(u_{\pi_1})f(u_{\pi_2})g(u_{\pi_1})g(u_{\pi_3}) \end{bmatrix}, \tag{67}$$

$$\psi(f_1, f_2, g_1, g_2; u_{1:4}) = \phi(f_1, f_2; u_{1:4}) + \phi(g_1, g_2; u_{1:4}) - \frac{1}{2}\sum_{a=1}^{2}\sum_{b=1}^{2}\phi(f_a, g_b; u_{1:4}), \tag{68}$$

and $\widehat{\mathrm{CKD}}^2$ averages ψ over all index tuples

$$\Sigma = (n_1 < n_2;\ n_1' < n_2';\ i_1 < i_2 < i_3 < i_4).$$

With noisy data, ψ is evaluated on observed entries. That is, each $f_n(u_i)$ is replaced by $\bar{X}_{ni}$ and each $g_n(u_i)$ by $\bar{Y}_{ni}$. In this appendix section, we write $\widehat{\mathrm{CKD}}^2_{\mathrm{noisy}}$ to denote this estimator, and we write $\widehat{\mathrm{CKD}}^2_{\mathrm{nf}}$ (where nf is for "noise-free") to denote the estimator computed on the noise-free responses. The "noise-free" estimator is precisely the one previously analyzed in Appendix B.

Proposition C.1 (Unbiasedness under noise). *Assume (N1)–(N2) and the moment conditions of theorem 3.1. Then for every $L \geq 1$,*

$$\mathbb{E}\big[\widehat{\mathrm{CKD}}^2_{\mathrm{noisy}} \,\big|\, \mathcal{D}\big] = \widehat{\mathrm{CKD}}^2_{\mathrm{nf}}, \qquad \textit{and hence} \qquad \mathbb{E}\big[\widehat{\mathrm{CKD}}^2_{\mathrm{noisy}}\big] = \mathrm{CKD}^2. \tag{69}$$

Proof. Applying iterated expectations, $\mathbb{E}\big[\widehat{\mathrm{CKD}}^2_{\mathrm{noisy}}\big] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\xi,\zeta}\left[\widehat{\mathrm{CKD}}^2_{\mathrm{noisy}} \mid \mathcal{D}\right]\right]$, with the inner expectation averaging over the noise terms.

Conditional on $\mathcal{D}$, the averaged noise terms are independent and mean zero by (N1)–(N2). Expanding each response into signal plus noise, every term that contains a noise variable contains it only once, and so has conditional mean zero. Hence, every term averages like:

$$\mathbb{E}\left[\begin{array}{l} \bar{X}_{n_1 i_1}\bar{X}_{n_1 i_2}\bar{Y}_{n_2 i_1}\bar{Y}_{n_2 i_2} \\ +\,\bar{X}_{n_1 i_1}\bar{X}_{n_1 i_2}\bar{Y}_{n_2 i_3}\bar{Y}_{n_2 i_4} \\ -\,2\,\bar{X}_{n_1 i_1}\bar{X}_{n_1 i_2}\bar{Y}_{n_2 i_1}\bar{Y}_{n_2 i_3} \end{array} \,\middle|\, \mathcal{D}\right] = \begin{array}{l} f_{n_1}(u_{i_1})f_{n_1}(u_{i_2})g_{n_2}(u_1)g_{n_2}(u_2) \\ +\,f_{n_1}(u_1)f_{n_1}(u_2)g_{n_2}(u_3)g_{n_2}(u_4) \\ -\,2\,f_{n_1}(u_1)f_{n_1}(u_2)g_{n_2}(u_1)g_{n_2}(u_3), \end{array}$$

including for the self-terms since the noise is assumed independent across neurons and the self-term estimator only touches distinct neuron pairs. Thus, the inner expectation evaluates to $\mathbb{E}_{\xi,\zeta}\left[\widehat{\mathrm{CKD}}^2_{\mathrm{noisy}} \mid \mathcal{D}\right] = \widehat{\mathrm{CKD}}^2_{\mathrm{nf}}$ and the estimator is remains unbiased as claimed. □

C.3 The variance of the U-statistic estimator is insensitive to noise to first-order

Recall from Appendix B that two index tuples Σ and Σ' have collision pattern (α, β, γ) when they share α $\mathcal{F}$-neurons, β $\mathcal{G}$-neurons, and γ inputs. After incorporating noise, we need to additionally account for collisions over neuron-input pairs. For example, suppose Σ and Σ' both contain the $\mathcal{F}$-neuron f_1 and both contain the input u_2. Then the noisy response $\bar{X}_{12}$ enters both $\psi(\Sigma)$ and $\psi(\Sigma')$, and its noise is a source of covariance between them. If instead the two tuples share f_1 but no input, or share u_2 but no $\mathcal{F}$-neuron, then the two kernels are evaluated on disjoint sets of noisy responses, so the noise contributes nothing.

Accordingly, we now read Σ as the full collection of random variables that $\psi(\Sigma)$ depends on: its neurons, its inputs, and the averaged noise of its sixteen neuron-input pairs. Under this reading, $\Sigma \cap \Sigma'$ consists of the shared neurons, the shared inputs, and the noise of the shared neuron-input pairs, and our prior bound holds verbatim. To distinguish the two, we write

$$V^{\mathrm{noisy}}_{\alpha\beta\gamma} := \mathrm{Var}\Big(\mathbb{E}\big[\psi(\Sigma) \,\big|\, \Sigma \cap \Sigma'\big]\Big), \tag{70}$$

to denote the variance terms for $\widehat{\mathrm{CKD}}^2_{\mathrm{noisy}}$.

Collision probabilities depend only on the index tuples, not on the noise. As before,

$$\mathrm{Var}\big[\widehat{\mathrm{CKD}}^2_{\mathrm{noisy}}\big] = \sum_{\alpha,\beta,\gamma} \tilde{C}_{\alpha\beta\gamma}\, V^{\mathrm{noisy}}_{\alpha\beta\gamma}, \tag{71}$$

with the same hypergeometric weights $\tilde{C}_{\alpha\beta\gamma}$.

In Appendix A and Appendix B, we previously saw that the constants V_{001}, V_{010}, and V_{100} describe the estimator's variance to first-order, and that these terms asymptotically dominate in the large-sample limit. Here we show that these determinitive constants are insensitive to the trial-to-trial noise.

Proposition C.2 (Noise does not affect the first-order constants). *Assume (N1)–(N3). If a collision pattern (α, β, γ) has no shared neuron-input pairs, that is if $\alpha + \beta = 0$ or $\gamma = 0$, then*

$$V^{\mathrm{noisy}}_{\alpha\beta\gamma} = V_{\alpha\beta\gamma}.$$

In particular, this holds for the three first-order patterns,

$$V^{\mathrm{noisy}}_{100} = V_{100}, \qquad V^{\mathrm{noisy}}_{010} = V_{010}, \qquad V^{\mathrm{noisy}}_{001} = V_{001},$$

as well as for $V^{\mathrm{noisy}}_{000} = V_{000} = 0$. For every other pattern, $V^{\mathrm{noisy}}_{\alpha\beta\gamma} \geq V_{\alpha\beta\gamma}$.

Proof. Write A for the index part of the overlap, namely the shared neurons and shared inputs, so that $\Sigma \cap \Sigma'$ consists of A together with the noise of the shared neuron-input pairs. By the law of total variance,

$$V^{\text{noisy}}_{\alpha\beta\gamma} = \mathrm{Var}\Big(\mathbb{E}\big[\psi(\Sigma) \mid A\big]\Big) + \mathbb{E}\Big[\mathrm{Var}\Big(\mathbb{E}\big[\psi(\Sigma) \mid \Sigma \cap \Sigma'\big] \;\Big|\; A\Big)\Big]. \tag{72}$$

Consider the first term. Since A contains no noise variables, the expectation conditioned on A averages over all of the noise. Hence, by Proposition C.1, $\mathbb{E}[\psi(\Sigma) \mid A]$ returns the kernel evaluated on noiseless responses, and the variance of this quantity is precisely what we defined as $V_{\alpha\beta\gamma}$. The second term of eq. (72) is nonnegative, which gives the final claim.

Now suppose the pattern has no shared neuron-input pairs. Then $\Sigma \cap \Sigma' = A$, so the inner conditional expectation in the second term is a function of A alone. Conditioning on A leaves it constant, so its conditional variance is zero and the term vanishes. A neuron-input pair is shared only if its neuron and its input are both shared, so the number of shared pairs is $\alpha\gamma$ in system X and $\beta\gamma$ in system Y. Both vanish when $\alpha + \beta = 0$ or $\gamma = 0$, which covers $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, and $(0, 0, 0)$. □

C.4 The variance bound carries over under noise

Under (N1)–(N3), the proof of theorem 3.1 carries over with minimal modifications. Each noise variable is attached to a single (neuron, input) pair, so two terms of the U-statistic sum can therefore share noise only if they share both a neuron and an input, so the decomposition over collision patterns is unchanged. Indeed, Proposition C.2 above already established this and showed that the first-order behavior of $\mathbb{V}\mathrm{ar}[\widehat{\mathrm{CKD}}^2]$ is unaffected by noise.

The only step that changes is the bound on $\mathbb{V}\mathrm{ar}[\psi]$ that controls the higher-order collisions. That argument uses only the offset invariance of ϕ, the conditional independence of the two neurons given the inputs (which (N2) preserves), and fourth moments of the responses. It therefore holds with s and κ in eqs. (61) and (62) computed from the observed responses, i.e. with $f(u) - \bar{f}$ replaced by $f(u) + \bar{\xi} - \bar{f}$ (and likewise for g), where the expectations now also average over the noise at input u. Hence eq. (64) holds verbatim, with s and κ taken to be the scale and kurtosis of the noise-corrupted responses, provided the noise has finite fourth moments.

D What is different about CKA?

D.1 No Unbiased Estimators for CKA and Squared Procrustes Distance

We have seen that HSIC admits an unbiased U-statistic estimator, leading to an unbiased estimator of CKD^2. Here, we show that it is not possible to completely eliminate estimator bias in other popular representational similarity measures including CKA and the squared Procrustes distance.

CKA is a nonlinear combination of three HSIC terms, which are unbiasedly estimable. However, unbiasedness is not preserved under nonlinear transformations. The textbook example is that $1/p$ is not unbiasedly estimable from n Bernoulli trials for any n (see Lehmann & Casella, 1998, example 1.2 in Chapter 2). The proposition below constructs a similar argument for CKA, showing that it is not possible to achieve unbiased estimation across all problem instances.

Proposition D.1. *Let $\mathcal{U} = [0, 1]$ with P_U uniform, and fix $M \geq 1$ and $N_X, N_Y \geq 1$. There is a one-parameter family of tuning-curve distributions such that no estimator $\hat{T}$ of* CKA *computed from the observed response matrices, and having finite expectation at some interior parameter value, is unbiased on any subinterval of* $[0, 1]$.

Proof. Let $\varphi_1(u) = \sqrt{2}\cos(2\pi u)$ and $\varphi_2(u) = \sqrt{2}\cos(4\pi u)$, which are centered, bounded, and orthonormal in $L^2(P_U)$. Take $\mathcal{F}$ to be the point mass at φ_1, and let $\mathcal{G}$ place mass p on φ_1 and $1 - p$ on φ_2. That is, for $f \sim \mathcal{F}$, we have $f = \varphi_1$ with probability one, while for $g \sim \mathcal{G}$ we have $g = \varphi_1$ with probability p and $g = \varphi_2$ with probability $1 - p$.

Both curves are centered, so

$$\bar{k} = \varphi_1 \otimes \varphi_1, \qquad \bar{h}_p = p\,\varphi_1 \otimes \varphi_1 + (1 - p)\,\varphi_2 \otimes \varphi_2, \tag{73}$$

and it is easy to check that

$$\langle \bar{k}, \bar{h}_p \rangle = p, \qquad \|\bar{k}\|^2 = 1, \qquad \|\bar{h}_p\|^2 = p^2 + (1-p)^2 = 2p^2 - 2p + 1. \tag{74}$$

Thus we have constructed a family of problem instances where

$$\mathrm{CKA}_\infty = \frac{p}{\sqrt{2p^2 - 2p + 1}} \tag{75}$$

controlled by an arbitrary parameter $p \in [0, 1]$.

Crucially, the right hand side of eq. (75) is *not expressible as a polynomial* in p. Below, we show that every estimator has an expected value that *is a polynomial* in p. The two can therefore never exactly agree over any subinterval $p \in [0, 1]$, which is sufficient to conclude the proof.

Let ν denote the number of tuning curves drawn from $\mathcal{G}$ that are equal to φ_1. Given our construction, ν is a binomial random variable with N_Y trials and success probability p. Now let $\widehat{T}$ denote any estimator and let $t_\nu = \mathbb{E}[\widehat{T} \mid \nu]$ denote the conditional expectation of the estimator given ν. Crucially, t_ν is a constant with respect to p since the count variable ν contains all information about p that is available to any estimator. By the law of total expectation followed by the expression for the expected value of a binomial random variable we have:

$$\mathbb{E}[\widehat{T}] = \mathbb{E}_\nu[t_\nu] = \sum_{\nu=0}^{N_Y} \binom{N_Y}{\nu} p^\nu (1-p)^{N_Y - \nu}\, t_\nu. \tag{76}$$

This is a polynomial in p of degree at most N_Y, thus concluding the proof. □

The same construction shows that the squared Procrustes distance is also not unbiasedly estimable. Recall that, by the duality of Harvey et al. (2023), the squared Procrustes distance between normalized representations is the squared Bures distance between the centered kernel operators,

$$\mathrm{Proc}^2_\infty(\bar{k}, \bar{h}) \;=\; \mathrm{Tr}\,\bar{k} \;+\; \mathrm{Tr}\,\bar{h} \;-\; 2\,\mathrm{Tr}\left[\left(\bar{k}^{1/2}\bar{h}\,\bar{k}^{1/2}\right)^{1/2}\right]. \tag{77}$$

Proposition D.2. *Under the hypotheses of Proposition D.1, and on the same family of tuning-curve distributions, no estimator of $\mathcal{P}^2$ with finite expectation at some interior parameter value is unbiased on any subinterval of $[0, 1]$.*

Proof. Take $\mathcal{F}$ and $\mathcal{G}$ exactly as in the proof of Proposition D.1. Since φ_1 is a unit vector, $\bar{k} = \bar{k}^{1/2} = \varphi_1 \otimes \varphi_1$. Orthogonality of φ_1 and φ_2 then gives

$$\bar{k}^{1/2}\,\bar{h}_p\,\bar{k}^{1/2} \;=\; p\,\varphi_1 \otimes \varphi_1, \tag{78}$$

whose square root is $\sqrt{p}\,\varphi_1 \otimes \varphi_1$. Thus, the Bures fidelity term is

$$\mathrm{Tr}\left[\left(\bar{k}^{1/2}\bar{h}\,\bar{k}^{1/2}\right)^{1/2}\right] = \sqrt{p}, \tag{79}$$

while $\mathrm{Tr}\,\bar{k} = 1$ and $\mathrm{Tr}\,\bar{h}_p = p + (1-p) = 1$, so that

$$\mathrm{Proc}^2_\infty \;=\; 2 - 2\sqrt{p}. \tag{80}$$

The right-hand side of eq. (80) is not a polynomial in p. The rest of the proof carries over verbatim from proposition D.1: the expectation of any estimator is a polynomial in p, and therefore no estimator can be unbiased on this family of problem instances. □

The inability of CKA or Procrustes to be estimated with an unbiased estimator has implications for their variance as well. Crucially, because no unbiased estimator exists for either CKA or Procrustes, then neither measure can be constructed as a U-statistic. Hence, the approaches we take in Appendix A and Appendix B no longer apply. Indeed, empirical simulations in Section D.1 reveals that the variance of CKA behaves quite nonlinearly compared to the ordered theory we derive for HSIC and CKA.

D.2 THE CONSISTENCY OF THE PLUG-IN ESTIMATOR OF CKA DEPENDS ON THE ORDER OF LIMITS

In the main text, we noted that

$$\lim_{\min(M,N_X,N_Y)\to\infty} S_{\mathrm{pi}}(\boldsymbol{X},\boldsymbol{Y}) = \langle \bar{k}, \bar{h} \rangle_{\mathrm{HS}}, \tag{81}$$

and that in this sense the plug-in estimators of both HSIC and CKD are consistent irrespective of how one takes $M, N_X, N_Y \to \infty$. Here, we show that CKA is not in general consistent if one takes a proportional high-dimensional limit in which we take $M, N_X, N_Y \to \infty$ such that

$$\frac{M}{N_X} \to q_X \in (0,\infty) \quad \text{and} \quad \frac{M}{N_Y} \to q_Y \in (0,\infty) \tag{82}$$

then the limit of the plug-in estimator may deviate by an order-one amount from the desired limit depending on the properties of the population kernel. This result is known, for example, from Chun et al. (2025), but we give a self-contained though informal derivation here.

Consider, for simplicity, the case in which the tuning curves are Gaussian, such that the empirical kernel matrices $\widehat{\boldsymbol{K}}$ and $\widehat{\boldsymbol{H}}$ are Wishart. Expanding the definition, the plug-in estimator of the CKA is

$$\widehat{\mathrm{CKA}}_{\mathrm{pi}}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\widehat{\boldsymbol{K}}\boldsymbol{C}\widehat{\boldsymbol{H}})}{\sqrt{\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\widehat{\boldsymbol{K}}\boldsymbol{C}\widehat{\boldsymbol{K}})\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\widehat{\boldsymbol{H}}\boldsymbol{C}\widehat{\boldsymbol{H}})}}. \tag{83}$$

Now, using standard results on Wishart matrix moments, we have that the expectations of the three terms with respect to the Gaussian distribution of tuning curves are

$$\mathbb{E}_{f,g}\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\widehat{\boldsymbol{K}}\boldsymbol{C}\widehat{\boldsymbol{H}}) = \frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}\boldsymbol{H}), \tag{84}$$

$$\mathbb{E}_{f,g}\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\widehat{\boldsymbol{K}}\boldsymbol{C}\widehat{\boldsymbol{K}}) = \left(1+\frac{1}{N_X}\right)\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}\boldsymbol{K}) + \frac{M}{N_X}\left(\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\boldsymbol{K})\right)^2, \tag{85}$$

and analogously

$$\mathbb{E}_{f,g}\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\widehat{\boldsymbol{H}}\boldsymbol{C}\widehat{\boldsymbol{H}}) = \left(1+\frac{1}{N_X}\right)\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\boldsymbol{H}\boldsymbol{C}\boldsymbol{H}) + \frac{M}{N_X}\left(\frac{1}{M}\operatorname{Tr}(\boldsymbol{C}\boldsymbol{H})\right)^2. \tag{86}$$

Under standard regularity assumptions on the kernels k and h, the remaining traces are well-behaved, and moreover the three normalized traces concentrate to their expectations with respect to both neuron and stimulus sampling. This gives, somewhat schematically,

$$\widehat{\mathrm{CKA}}_{\mathrm{pi}}(\boldsymbol{X},\boldsymbol{Y}) \sim \frac{\mathrm{CKA}(\boldsymbol{K},\boldsymbol{H})}{\sqrt{(1+d_X)(1+d_Y)}}, \tag{87}$$

where we have defined the normalized effective dimensionalities

$$d_X = \frac{1}{N_X}\frac{(\operatorname{Tr}(\boldsymbol{C}\boldsymbol{K}))^2}{\operatorname{Tr}(\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}\boldsymbol{K})} \quad \text{and} \quad d_Y = \frac{1}{N_Y}\frac{(\operatorname{Tr}(\boldsymbol{C}\boldsymbol{H}))^2}{\operatorname{Tr}(\boldsymbol{C}\boldsymbol{H}\boldsymbol{C}\boldsymbol{H})}. \tag{88}$$

This is the primary high-dimensional bias result of Chun et al. (2025). Here, we have been quite informal: we have not specified the notion of asymptotic equivalence in $\sim$, nor the required regularity on the kernels. But, all of this could be made precise, and the qualitative result stands: if we consider a limit such that d_X and d_Y do not tend to zero sufficiently rapidly, then the plug-in estimator of the CKA acquires an order-one multiplicative bias. This is, importantly, dependent on the spectral properties of the kernel: roughly speaking if k and h are trace-class then d_X and d_Y should tend to zero even if M is taken to infinity proportionally with N_X and N_Y, while if their traces scale with M then M must be taken to infinity strictly more slowly. The latter is the usual high-dimensional case, while in the former the data have non-extensive intrinsic dimension.

The most extreme, and most illustrative, high-dimensional example is if we considered two independent, isotropic Gaussian clouds, *i.e.*, if $\boldsymbol{K} = \boldsymbol{H} = \mathbf{I}_M$. This corresponds to an infinitely-sharp

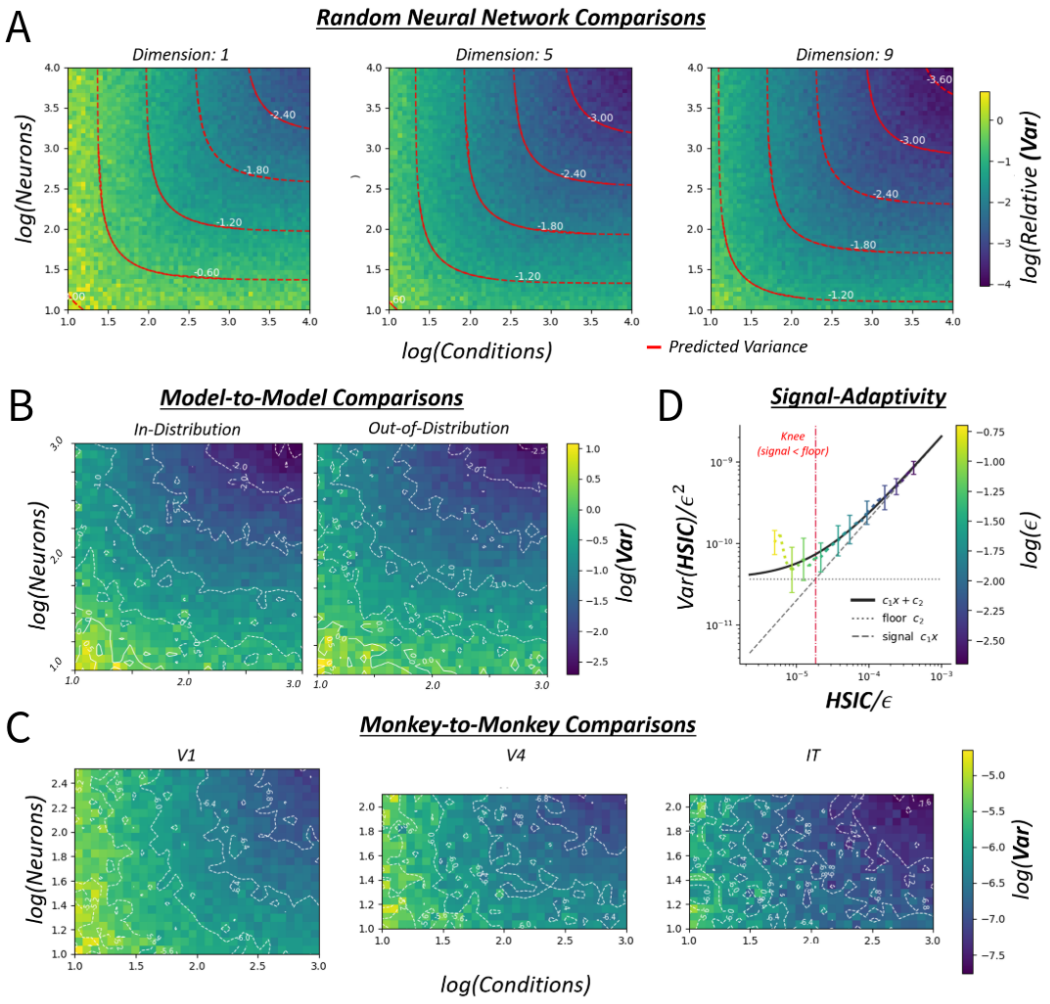

Figure 4: Reproduction of results for CKD in the context of HSIC. (**A**) Variance of estimating HSIC between random neural networks as in Section 4.1. (**B**) Variance of estimating HSIC between trained neural networks as in Section 4.2. (**C**) Variance of estimating HSIC between monkey regions as in Section 4.3. (**D**) Signal adaptivity of HSIC.

kernel that is one if two stimuli coincide and zero otherwise. In this case, we of course have that the CKA between $\boldsymbol{K}$ and $\boldsymbol{H}$ is

$$\mathrm{CKA}(\boldsymbol{K}, \boldsymbol{H}) = 1, \tag{89}$$

as they are the same. However,

$$d_X = \frac{M-1}{N_X} \sim q_X \quad \text{and} \quad d_Y = \frac{M-1}{N_Y} \sim q_Y, \tag{90}$$

so

$$\widehat{\mathrm{CKA}}_{\mathrm{pi}} \sim \frac{1}{\sqrt{(1+q_X)(1+q_Y)}}, \tag{91}$$

which may be substantially biased downward from one if q_X and q_Y are not small.

D.3 Experimental Results on HSIC and CKA

Here, we show the results for the behavior of $\mathbb{V}\mathrm{ar}(\mathrm{HSIC})$ in the context of the main experiments presented in the main body of the paper. The results presented here are for the unbiased S_{sc} estimator. Given that we derive a similar upper bound on variance scaling for HSIC as with CKD

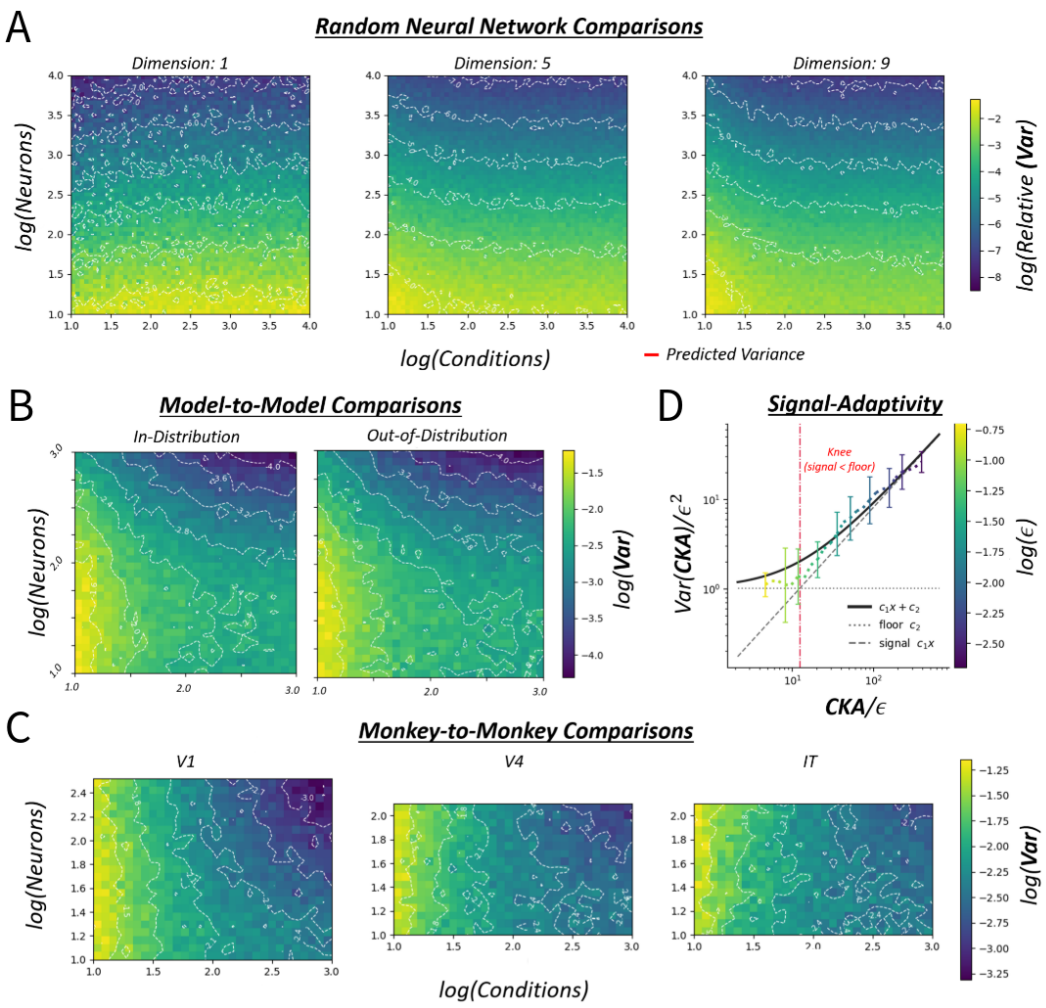

Figure 5: Reproduction of results for CKD in the context of CKA. Note scaling behavior differs wildly. (**A**) Variance of estimating CKA between random neural networks as in Section 4.1. (**B**) Variance of estimating CKA between trained neural networks as in Section 4.2. (**C**) Variance of estimating CKA between monkey regions as in Section 4.3. (**D**) Signal adaptivity of CKA.

in Appendix A, it is no surprise that we demonstrate similar qualitative and quantitative results for HSIC as in CKD.

Next, we present the results for the behavior of $\mathbb{V}\mathrm{ar}(\mathrm{CKA})$ in the context of the main experiments from the main body of the paper. CKA behaves markedly different in all these scenarios, where its dependence on changes in M and N vary wildly on both setting and magnitude.

Last, we plot the same trends but for CKA (Figure 6A-B), which produces similar qualitative results. This is suprising because CKA admits no unbiased estimator, and can diverge qualitatively from the predicted $\mathcal{O}(\frac{1}{M} + \frac{1}{N})$ scaling law when comparing one model to another. This suggests that while CKA admits no unbiased estimator, variance in similarity estimation in the top 2 principal components follows the established theoretical trend well, so our experimental prescriptions may hold for a larger class of similarity metrics.

Notably, the Kendall's τ for CKA are decreased relative to CKD (Figure 6C). Since CKA is scale-invariant while CKD is not, this suggests that the difference in scale across multiple neural regions and models may be important in understanding multi-system relationships. In other words, the scale invariance of CKA, when comparing multiple monkeys and models to one another, surprisingly hurts one's ability to determine the proper ordering of the their pairwise relationships in concert,

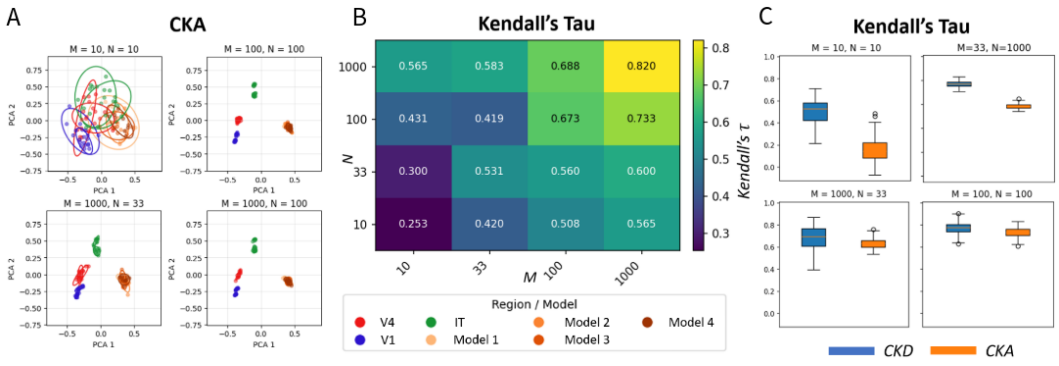

Figure 6: Clustering experiments for CKA. (**A**) Projection of estimated CKA between 4 deep models and 6 neural recordings from 3 different brain areas. (**B**) Estimates of Kendall's τ on different degrees of subsampling. (**C**) Comparisons between Kendall's τ in CKD with CKA.

even when comparing across models and neural data. Hence, the difference in scale across multiple neural regions and models may be important in understanding multi-system relationships.

E Variance Scaling when $N_X \neq N_Y$

In the main body of the paper, we present empirical verification of 3.1 in experiments where $N_X = N_Y$. For $N = N_X = N_Y$, we showed L-shaped level curves when plotting log-variance against $\log(M)$ and $\log(N)$, which are predicted by our theory of $\mathbb{V}\text{ar} \sim \mathcal{O}(\frac{1}{M} + \frac{1}{N})$.

However, in most practical neuroscience applications, $N_X \neq N_Y$. In Figure 7, we show variance scaling in this setting, fixing M to be sufficiently large as to be negligible under our theory. Similarly, our results in Figure 7(A) recapitulate the L-shaped curves from Section 4.1, verifying our theory. Reassuringly, comparing monkey data to monkey data in Figure 7B also follows theory quite well.

However, scaling on occasion can follow our theory, but there are systematic deviations from it. For example, while HSIC appears to follow L-shaped level curves for model-to-model comparisons in Figure 8, CKD is dominated by the neurons drawn from one of the two models. Our hypothesis for this effect is that when M is fixed and $N_X \neq N_Y$, the scale of the activations—and thus the constant factors in the bounds—become more important to the overall scaling behavior. We attempt to isolate this effect by normalizing the scale of activations to have unit norm, mapping activations $X \to \frac{X}{||X||_F}, Y \to \frac{Y}{||Y||_F}$, and then finding the variance of similarity estimates on the normalized activations.

This has a surprising effect on the variance scaling. The variance of HSIC and CKD flattens to $\mathcal{O}(\frac{1}{N_X N_Y})$, faster than our upper bound. We suspect this is due to vanishing leading constants $V_{001}, V_{010}, V_{100}$ in the U-statistics proof for statistical reasons. Notably, the scale of variance remains constant independent of region identity. This suggests that removing the effect of scaling can remove the overdependence of variance on one neural system over the other. In a neuroscience setting, where the constants cannot be well estimated until after the experiment, this improves *a priori* optimization of experimental parameters.

F Cross-Region Comparisons in Monkey Data

To further illustrate the region-independent nature of our theory, we extend our results from Section 4.3. In Figure 2B, we plotted variance heatmaps of similarity estimates between the same region across two monkeys. One may protest to our results' generality; two regions in the same species likely have similar underlying statistics. In Figures 9, 10, 11, 12, we show that our results can be generalized across arbitrary region-to-region estimation.

Almost universally, regardless of which region is compared to which region, regardless of whether we are comparing across monkeys or within monkeys, the variance estimates follow L-shaped curves

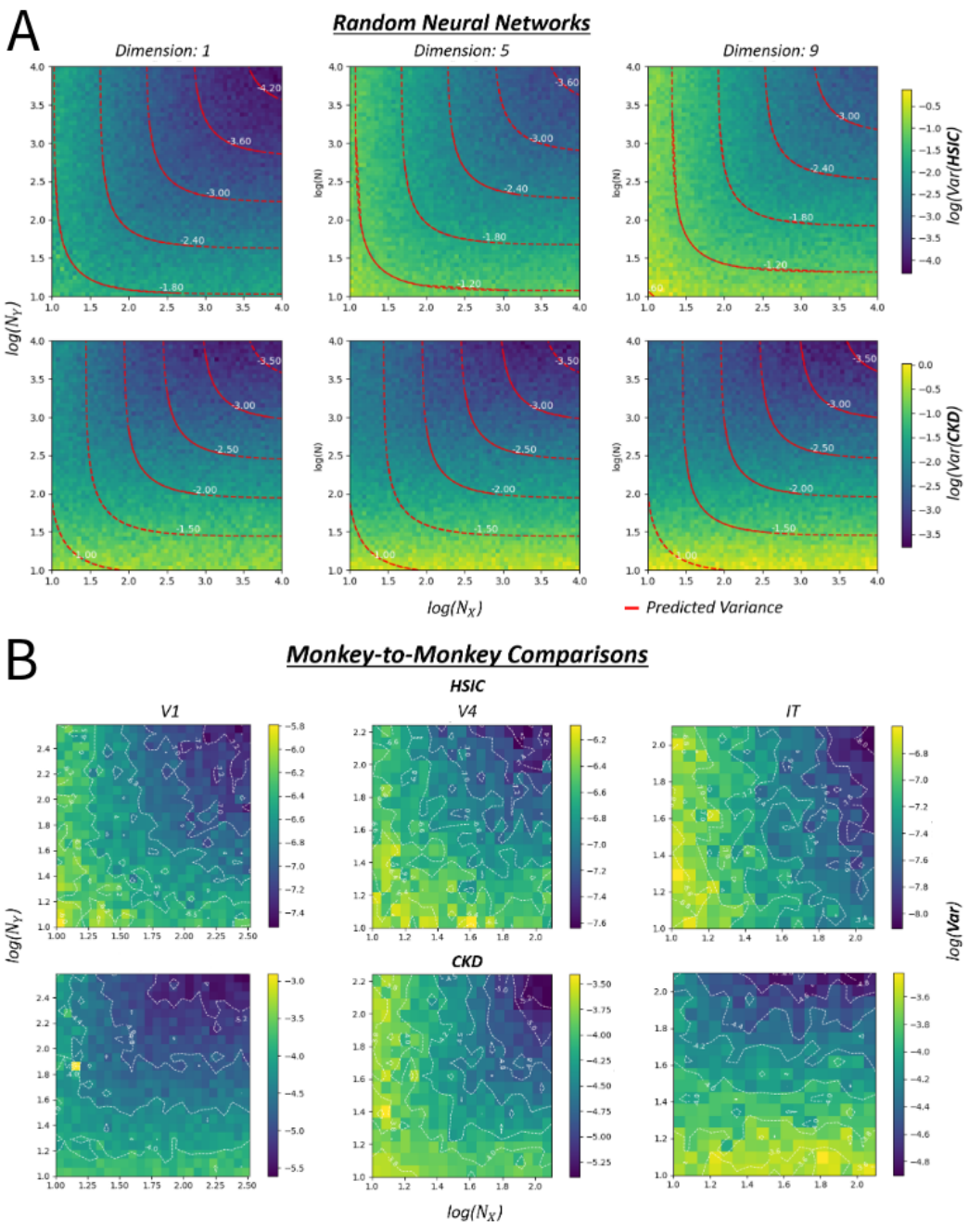

Figure 7: Heatmaps of variance scaling when $N_X \neq N_Y$. X-axes are $\log(N_X)$, Y-axes are $\log(N_Y)$. (**A**) Variance of estimating HSIC and CKD between random neural networks as in Section 4.1. $M = 10000$. (**B**) Monkey regions from Papale et al. (2025) as in 4.3. L-shaped level curves emerge in V1 and V4 for both HSIC and CKD, while IT comparisons loses some structure. $M = 1000$

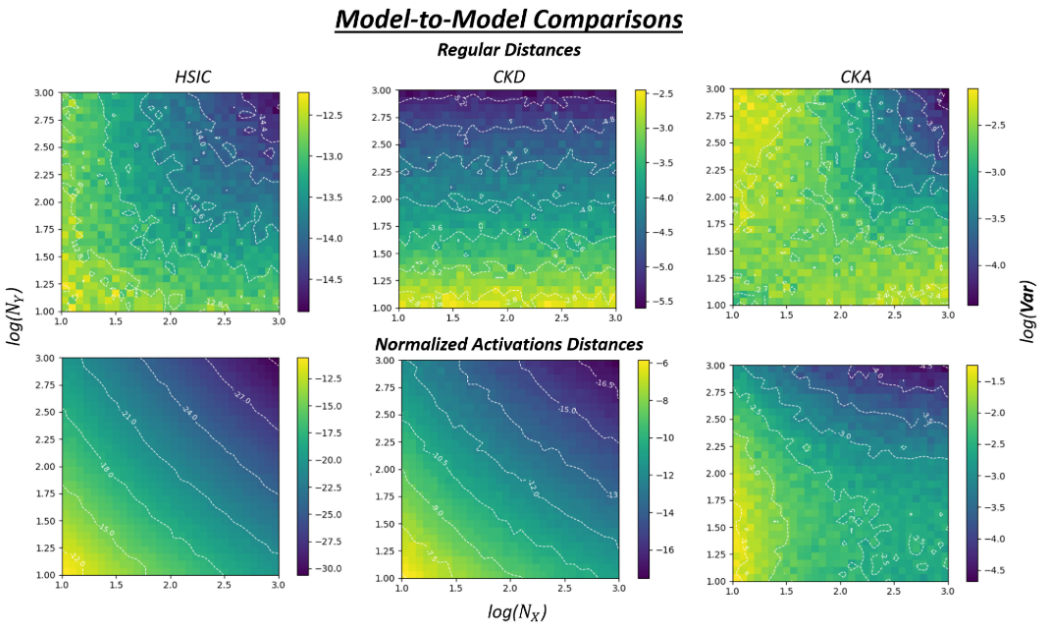

Figure 8: Heatmaps of variance scaling when $N_X \neq N_Y$. $M = 1000$ for trained deep networks as in Section 4.2. HSIC aligns with theory. CKD is dominated by the neurons in the model with the more severe noise filter applied during training, despite activations being collected when noise is turned off. CKA is symmetric but behaves differently from HSIC.

as predicted by our theory. Notably, the scale of relative variance appears inconstant for the most part regardless of the simulation type, as seen by traversing the corresponding heatmaps across the rows in each figure. Overall, these results reiterate that our theoretical predictions are reproduced in biological data, and serve as a prediction for how to efficiently collect neural data given a finite recording budget.

G Random Neural Network Construction

Consider two random neural networks k, h. While their true similarity can be measured by sampling infinitely many M or N, we want to study how subsampling affects our similarity estimate. We can compare the similarity between k and h by constructing representations X and Y by presenting M inputs and extracting activations from only N neurons, keeping the neuron identities constant through sampling.

Consider a neural network k with one hidden layer with weights; its outputs are arbitary as we are only concerned with representations in this untrained hidden layer. For

$$w \sim \mathcal{N}(\mu_k, \sigma_k), \text{weights}$$
$$u \sim \mathcal{N}(0, I_D), \text{inputs}$$

over M inputs and N neurons, we can collect the neural representation for k

$$X = \text{ReLU}(UW), U = (u_1^\top, .., u_M^\top) \in \mathbb{R}^{M \times D}, W = (w_1, ..., w_N) \in \mathbb{R}^{D \times N} \tag{92}$$

Similarly, we can construct neural representation for another shallow, random neural network h

$$Y = \text{ReLU}(UV), V = (v_1, ..., v_N) \in \mathbb{R}^{D \times N}, v \sim \mathcal{N}(\mu_h, \sigma_h)$$

Thus, we can compare the similarity between k and h by measuring the similarity between their neural representations X and Y for any similarity measure of our choosing.

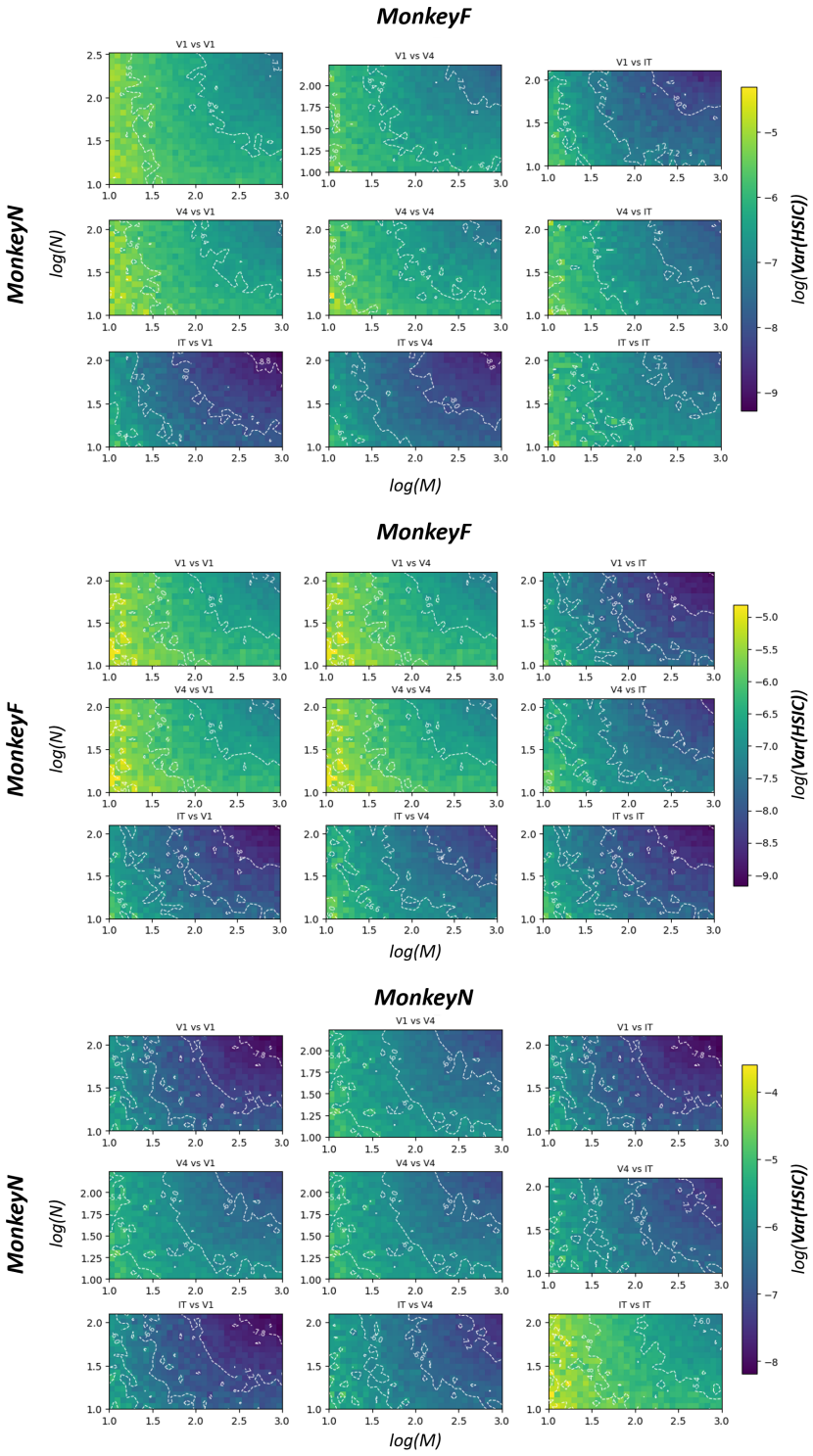

Figure 9: Variance of HSIC when subsampling across all regions between the two macaques available in the dataset. Within a set of heatmaps, rows (left-to-right) and columns (top-to-bottom) are ordered as V1, V4, IT. Subplots are ordered as MonkeyF-to-MonkeyN, MonkeyF-to-Monkey-F, and MonkeyN-to-MonkeyN.

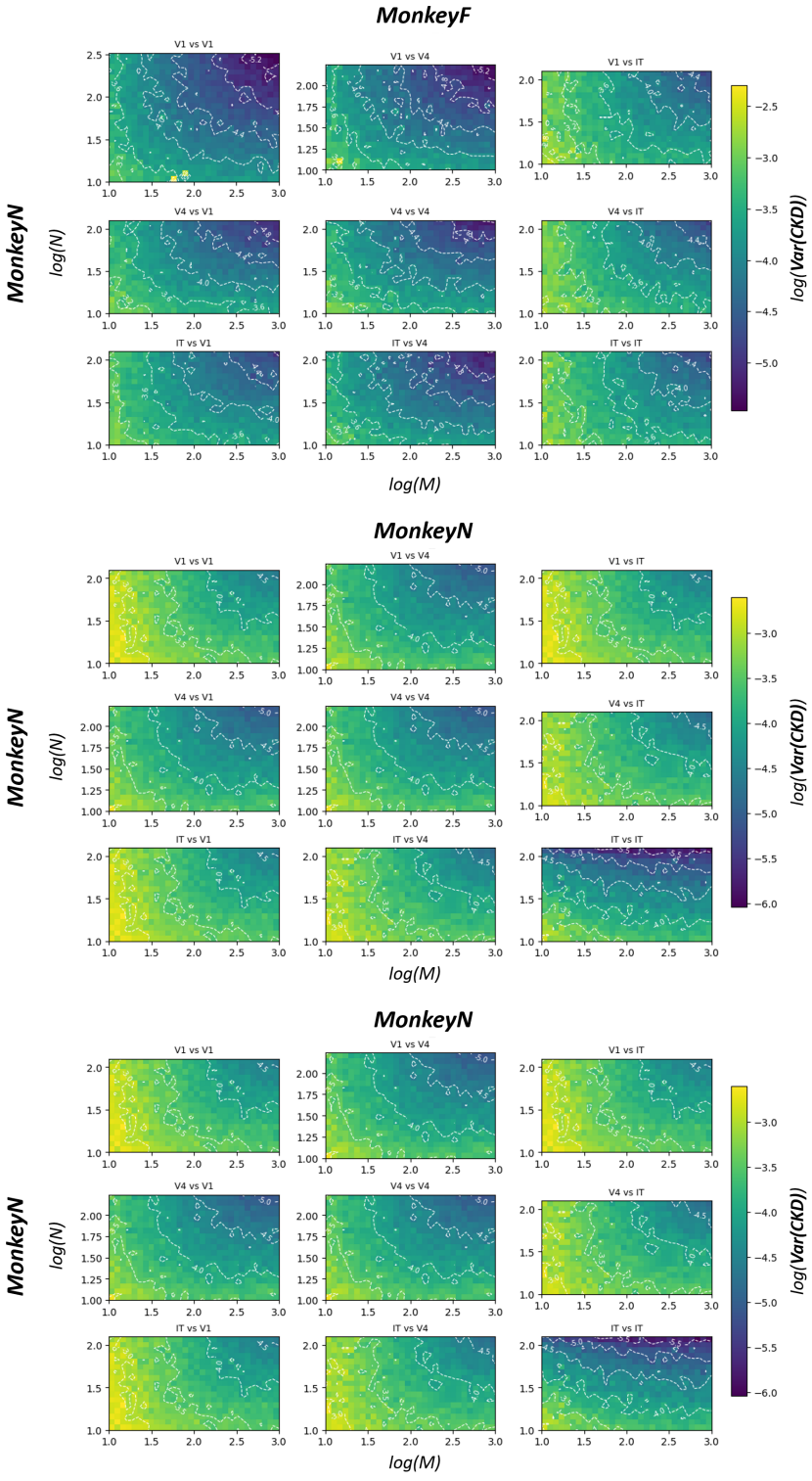

Figure 10: Variance of CKD when subsampling across all regions between the two macaques available in the dataset. Within a set of heatmaps, rows (left-to-right) and columns (top-to-bottom) are ordered as V1, V4, IT. Subplots are ordered as MonkeyF-to-MonkeyN, MonkeyF-to-Monkey-F, and MonkeyN-to-MonkeyN.

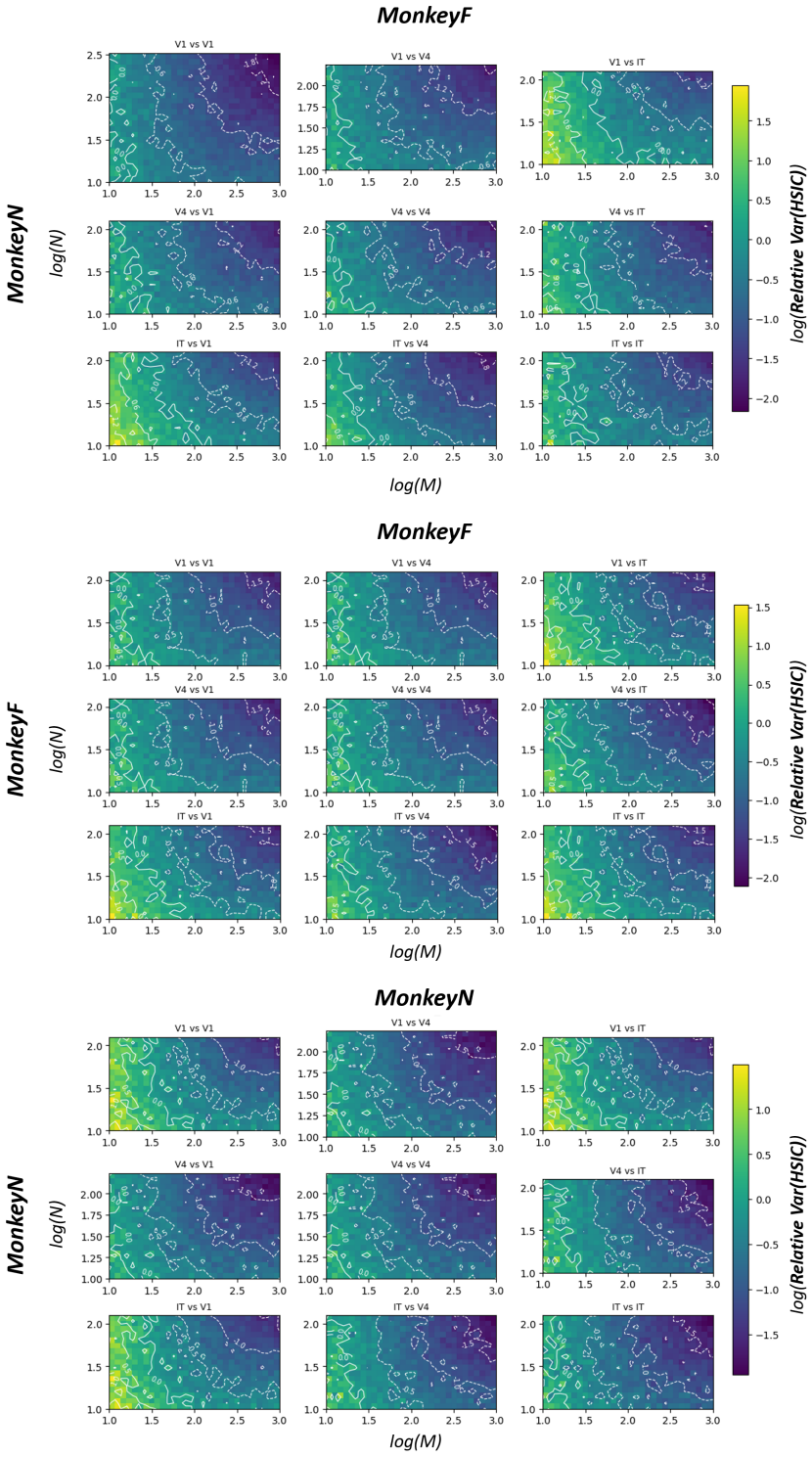

Figure 11: Relative variance of HSIC when subsampling across all regions between the two macaques available in the dataset. Within a set of heatmaps, rows (left-to-right) and columns (top-to-bottom) are ordered as V1, V4, IT. Subplots are ordered as MonkeyF-to-MonkeyN, MonkeyF-to-Monkey-F, and MonkeyN-to-MonkeyN.

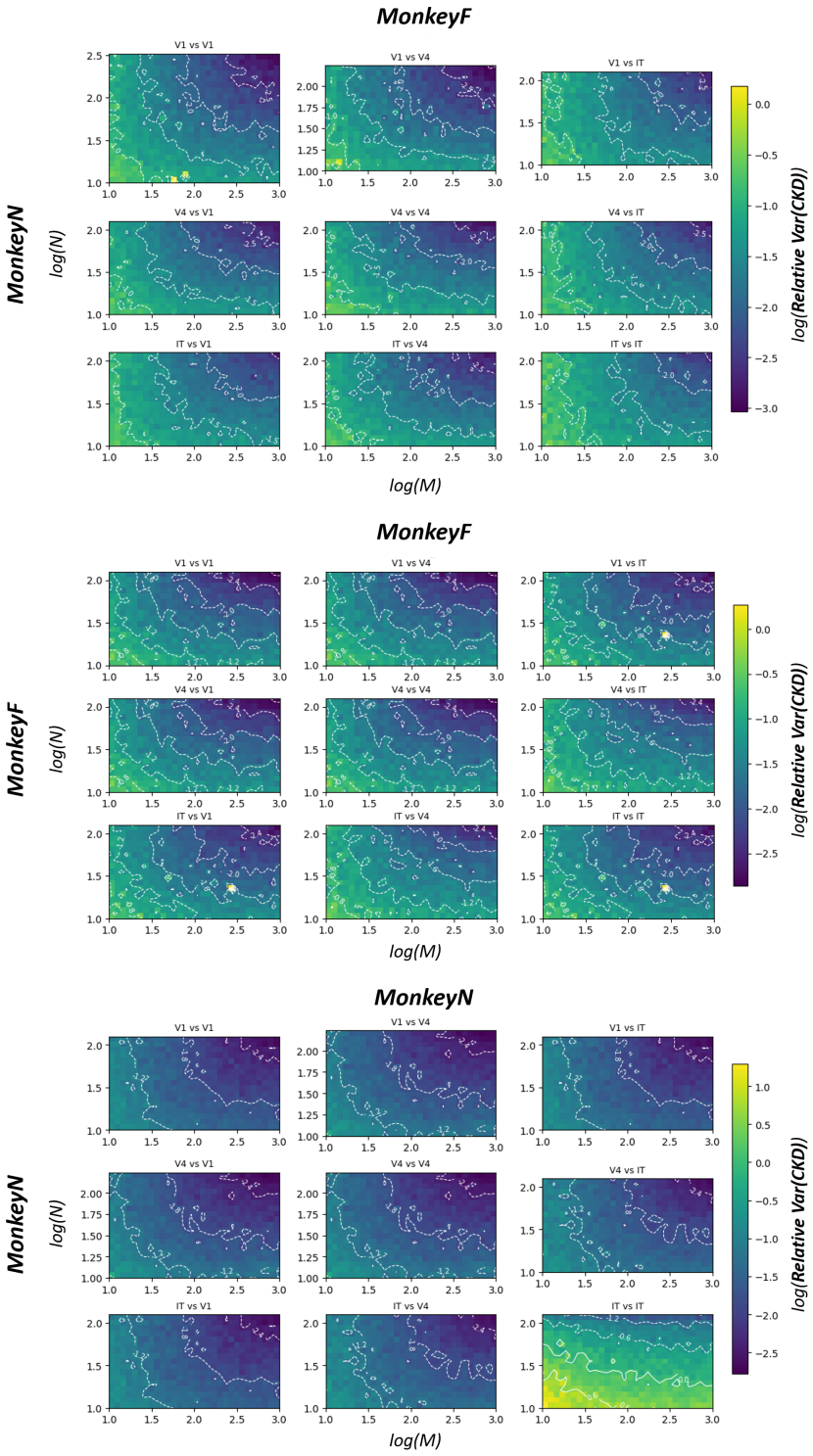

Figure 12: Relative variance of CKD when subsampling across all regions between the two macaques available in the dataset. Within a set of heatmaps, rows (left-to-right) and columns (top-to-bottom) are ordered as V1, V4, IT. Subplots are ordered as MonkeyF-to-MonkeyN, MonkeyF-to-Monkey-F, and MonkeyN-to-MonkeyN.



Figure 13: Representative images from the ImageNet dataset (Deng et al., 2009), the MacaqueIT-Bench dataset (Madan et al., 2024), and the THINGS Ventral Spiking Dataset (Papale et al., 2025). ImageNet and THINGS both comprise of naturalistic whole-scenes images. MacaqueITBench comprises of natural objects against a white background. Specific image categories intersect but are not identical between the three datasets.

H IMAGE DATASET DETAILS

We use three public datasets in this paper: the ImageNet dataset (Deng et al., 2009), the MacaqueIT-Bench dataset (Madan et al., 2024), and the THINGS Ventral Spiking Dataset (Papale et al., 2025) (Figure 13). The ImageNet dataset contains over 14 million naturalistic images with more than 20,000 unique semantic categories. We use the ImageNet dataset to test and validate our deep networks. Evaluation of performance and estimation of similarity between networks is done on a test set partition of the ImageNet dataset.

In Section 4.2, we evaluate the effects of in- versus out-of-distribution evaluation data for similarity estimates on our deep networks. We use the MacaqueITBench dataset as a proxy for out-of-distribution image data. The MacaqueITBench dataset consists of 8233 unique natural images. Since the deep models did not see these images during training or evaluation, we use the entire dataset as a pool of test images. The images in the MacaqueITBench dataset differ markedly from the ImageNet dataset; while both datasets comprise of naturalistic images, MacaqueITBench images consist of a natural object on a white background, while ImageNet images are an entire natural scene. The specific images and image categories also differ, so we use the MacaqueITBench dataset as a proxy for out-of-distribution natural images.

Last, we use both image and macaque spiking data from the THINGS Ventral Spiking Dataset (TVSD) in Section 4.3. Image data in TVSD was collected in the THINGS database (Hebart et al., 2023), consisting of 26,107 naturalistic visual images in whole scenes, similar to the ImageNet dataset. The TVSD dataset presented these images to two macaques, recording from V1, V4, and IT using chronic Utah array, consisting of between 200-512 neurons per region. The TVSD dataset is divided into a training set of 22,248 images presented 1 time each (12 images per category) and a test set of 100 images presented at 30 repetitions each. Because our theory suggests that trial-to-trial variability does not significantly impact similarity estimates, to increase the number of M we can work with, we use only images and data from the training set.

We threshold for mean reliability > 0.3 and use activations from the remaining neurons for comparison when subsampling; relatively few neurons are filtered out. Recordings are taken over 4 days for one macaque and 7 days for the other, with pseudorandomized image presentations across sessions, although all images are presented. In this paper, we treat the neurons recorded across sessions as identical. Hence, we construct six $22248 \times N_{TOTAL}$ activation matrices, and we randomly subsample rows (subsampling M) and columns (subsampling N) to evaluate similarity comparisons over empirical $X \in \mathbb{R}^{M \times N_X}$ matrices.

I Model Training Details

I.1 Model Architecture

We train self-supervised SimCLR models on ImageNet using an all-convolutional encoder with injected Gaussian noise after each convolution.

Let x be an input image. The encoder produces a representation h and the projection head produces the contrastive embedding z:

$$h = \text{AllConvNet}(x),$$
$$z = \text{projection_head}(h).$$

Contrastive learning loss (Normalized Temperature-scaled Cross-Entropy loss, also known as NT-Xent) is computed on z.

After every convolutional layer, activations are injected with Gaussian noise at different parametrizations of σ^2, such that $x \to x + \mathcal{N}(0,\ \sigma^2)$. In the main body of the paper, we focus on 5 models, with $\sigma^2 = 0, 0.01, 0.03, 0.09$. through Noise is sampled independently at each forward pass.

Table 1: AllConvNet encoder architecture. All convolutions use 3×3 kernels and padding 1. Channel width is `layer_dim` $= 96$. After `conv7` the feature map is 2×2 spatially; global average pooling (`adaptive_avg_pool2d` to 1×1) produces a 96-D representation.
Operations are all `Conv2d`, noise is always σ^2.

Layer	In $\to$ Out	Stride	Activation		
1	$3 \to 96$	1	ReLU		
2	$96 \to 96$	2	ReLU		
3	$96 \to 96$	2	ReLU		
4	$96 \to 96$	2	ReLU		
5	$96 \to 96$	2	ReLU		
6	$96 \to 96$	2	ReLU		
7	$96 \to 96$	2	—		
—	Adaptive avg. pool	$96 \times 2 \times 2 \to 96$	—	—	—

I.2 Training Regime

Table 2: Training hyperparameters and compute setup.

Setting	Value
Objective	SimCLR with NT-Xent (`lightly.loss.NTXentLoss`)
Optimizer	SGD, learning rate 0.06, momentum 0.9, weight decay 5×10^{-4}
LR schedule	`CosineAnnealingLR` over `epochs`
Epochs	2
Batch size	512
Input size	128×128 RGB

Training uses the ImageNet train split via `LightlyDataset`, producing two independently augmented views per image. Validation uses the ImageNet val split, resized to 128×128 and ImageNet-normalized, also subsampled at stride 2. Each batch is a pair of views (x_0, x_1) of the same images. The model computes $z_0 = \text{SimCLR}(x_0)$ and $z_1 = \text{SimCLR}(x_1)$, then minimizes NT-Xent between positive pairs across the batch.

I.3 Activation Extraction for Similarity Estimation

During evaluation of similarity estimates, we present M randomly sampled images from a dataset of choice (for specifics on the datasets see Appendix H). We sample N random neurons from a

desired layer (V1, V4, or IT in monkeys; second-to-last layer consisting of $96 \times 4 \times 4 = 1536$ neurons post-ReLU without additive Gaussian noise in trained neural networks) Thus, we construct the $X \in \mathbb{R}^{M \times N_X}$ neural representation matrix. The procedure for drawing neurons can be done either without replacement (subsampling) or with replacement (bootstrapping). In this paper we present the bootstrapped variance estimates, but the results do not qualitatively differ between the two procedures, provided that N is less than the total number of neurons in the layer; remarkably, even leaving just 1 out when sampling without replacement recapitulates these results qualitatively.

J Sampling Procedure for Section 4.3

In Section 4.3, at each value of M, we present M random images to the 10 regions of study. From each region, we randomly sample N random neurons 10 times to create 10 copies of each region, and compare all pairwise estimates of HSIC, CKD, and CKA between the copies. Hence, we construct a 100×100 matrix at each (M, N) pair, for $M, N \in \{10, 33, 100, 1000\}$. For the clusters, we apply MDS (10 dimensions) on the 100×100 matrix, then project onto the first two principal components.

Next, at each value of (M, N) we randomly sample 1 copy from each model to generate a 10×10 matrix which we can compute a Kendall's τ against a reference 10×10 matrix containing the true distances between each pair of models. We repeat this procedure 128 times and report the mean Kendall's τ in Figure 3. In this paper, we use the upper-triangular Kendall's τ, but using the row-wise Kendall's τ produces very similar results.