MUTUAL NEIGHBOR SIMILARITY: RELIABLE AND RICH REPRESENTATIONAL ALIGNMENT MEASUREMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Every representational similarity measure carries strengths and weaknesses. Mutual kNN (m-kNN) style approaches can report local alignment, but saturate at larger scales. In addition, these ordinal approaches, including kendall's tau and spearman rank correlation, discard all information-rich geometric details in favor of coarse rank-based descriptions, incurring various geometric failure modes. On the other end, many geometric approaches such as centered kernel alignment (CKA) provide global summary metrics but cannot describe alignment at local or intermediate scales. Furthermore, many of these methods can be easily confounded by factors such as embedding dimension and must be null-calibrated via expensive Monte-Carlo simulations to account for geometric priors and deliver robust estimates of alignment with meaningful baselines. To unify many of these strengths in a single measure while mitigating weaknesses, we introduce mutual neighbor similarity (MNS), a hybrid geometric and ordinal comparison of corresponding embedding space neighborhoods. With an analytically derived and simple to compute null expectation, MNS enables robust comparisons across all scales without saturation at either extreme, while accounting for the structure of point cloud geometries. We construct MNS, describe its invariances, derive its robust form, and run experiments on synthetic data demonstrating its ability to overcome various failure modes of existing approaches while introducing entirely new descriptive abilities. Finally, we present results on large scale real world data suggesting the superiority of our approach for transfer learning tasks.

1 INTRODUCTION AND RELATED WORK

Measuring the representational similarity of models is an increasingly relevant task within many different areas, ranging from representation learning (Peterson et al., 2018; Koh et al., 2020), model distillation and knowledge transfer (Hinton et al., 2015; Tian et al., 2022), interpretation (Huh et al., 2024), value alignment, (Wynn et al., 2024) computational neuroscience (Kriegeskorte et al., 2008; Jacoby et al., 2019; Yamins et al., 2014; Barrett et al., 2019), and neuroAI Muttenthaler et al. (2025); Sucholutsky & Griffiths (2023). However, existing measures are each limited in their ability to reliably describe the alignment of embeddings, requiring practitioners to attempt to glue together semantically distinct descriptions into a coherent picture Huh et al. (2024); Gröger et al. (2026). Existing approaches can be broadly classified into one of two categories: geometric and ordinal.

Geometric approaches act directly on kernel-Gram matrices of embeddings, and thus are aware of the full geometry of the point clouds they compare, accounting for complexities such as anisotropy - a defining feature of neural embeddings Liang et al. (2022); Godey et al. (2024). These approaches include CKA (Cortes et al., 2024; Kornblith et al., 2019) and its variants as well as Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008), a pioneering method in representational alignment. While these geometric methods deliver continuous global summary measures of alignment, they make the strong assumption that alignment is a linear property of space pairs (up to a choice of a predefined kernel) and they cannot be reliably modified to operate at different spatial scales. While centered kernel nearest neighbors alignment (CKNNA) makes an attempt at localizing CKA, it artificially inflates reported alignment by ignoring mismatches in nearest neighbor sets entirely Huh et al. (2024). Additionally, even the debiased form of CKA (Murphy et al., 2024) is hard to interpret, requiring simulation-derived null calibration on a per-comparison basis. Canonical correlation analysis (CCA) is the foundation for a subclass of geometric methods that measure

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.

alignment by finding correlation-maximizing linear transformations of each embedding space Raghu et al. (2017); Morcos et al. (2018). These methods suffer from many of the same deficiencies listed above, including expensive calibration, as well either invariance to arbitrary invertible linear transforms or sensitivity to orthogonal transforms depending on the formulation. Both of these were identified by Kornblith et al. (2019) as undesirable properties of a representational alignment measure. Finally, Procrustes analysis (Gower, 1975; Williams et al., 2022) acts on embeddings directly, by transforming each matrix into a shared space and measuring how closely the resulting shapes can be matched by some class of transformations - usually orthogonal transforms. In addition to the error incurred by the dimensionality matching transformations, this approach to alignment also carries rigid orthogonality assumptions.

Ordinal methods reduce similarities to their orderings and thus are free of linearity assumptions and are often definable at multiple spatial scales. For example, mutual k-NN (Huh et al., 2024) measures the size of the intersection of k-nearest neighbor sets in X and Y for arbitrary $1 \leq k \leq N-1$. However, as a binary set inclusion statistic, it is blind to the actual geometry of embeddings, inviting many failure modes. For example if a neighbor in X is not a neighbor in Y, the distance to the neighborhood is not considered. A near miss is penalized as much as a catastrophic one. Similarly, m-kNN cannot consider the internal structure of neighborhoods. While the k-nearest neighbors may match in X and Y, their ordering relative to an anchor can be perfectly reversed and the discrepancy will not be captured. Other ordinal approaches can remedy this specific issue, but they still do not account for geometry, and common approaches such as Spearman ρ or Kendall τ of similarities cannot be localized with respect to the number of neighborsKendall (1938); Spearman (1904). Additionally, m-kNN inherently saturates at the global extreme. Even with chance correction, m-kNN converges to 1 as k grows to $N-1$ rendering its measurements of global alignment unreliable and incomparable to measures of local alignment, an increasingly interesting axis to researchers Gröger et al. (2026). Since m-kNN is not reliable across spatial scales, the choice of k is forced to be made arbitrarily or heuristically, leaving measurements across many model pairings vulnerable to reordering depending on the choice of k.

We note that representations may also be compared by measuring the ability of a model (often a linear model/probe) to predict one representation from the other. In addition to being asymmetric, and carrying the assumption that embeddings should be related by affine transforms, this class of methods carries every inefficiency and inconsistency of model training.

2 Mutual Neighbor Similarity

Our goal is to define a measure that combines the advantages of existing approaches. Our measure should:

- Be definable and semantically consistent at all spatial scales.
- Not saturate at the local or global extremes.
- Give consistently interpretable measurements.
- Carry important invariances and sensitivities.
- Be geometric without linearity assumptions.
- Handle ties naturally.
- Scale to large data sets.

In order to proceed, we briefly define necessary preliminaries. Let $\mathcal{D}$ be a dataset of $N \in \mathbb{N}$ real events realized by a pair of models $f : \mathcal{D} \to \mathbb{R}^{d_X}$ and $g : \mathcal{D} \to \mathbb{R}^{d_Y}$ for embedding dimensions $d_X, d_Y \in \mathbb{N}$. The models generate embedding/representation matrices: $f(\mathcal{D}) = X \in \mathbb{R}^{N \times d_X}$ and $g(\mathcal{D}) = Y \in \mathbb{R}^{N \times d_Y}$. To achieve our goals, we take inspiration from the semantics of mutual k-NN and iteratively refine its coarse ability to discriminate between neighborhoods. Let $M = N - 1$ be the number of total neighbors of an anchor indexed by i and let $r_j^{X_i}, r_j^{Y_i}$ denote the neighbor rank to anchor $\mathcal{D}_i$ of the event $\mathcal{D}_j$ embedded in X and Y respectively. Notice that neighbor rankings to i are related by a permutation, or element of the symmetric group, $\pi_i \in S_M$ given by $\pi_i(r_j^{X_i}) = r_j^{Y_i}$.

m-kNN, a binary set inclusion statistic, can be expressed with respect to π_i:

$$\text{m-}k\text{NN}(i) = \frac{1}{2k}\sum_{r=1}^{k}[\pi_i(r) \leq k] + [\pi_i^{-1}(r) \leq k].$$

We can create a more information rich measure that considers the extent to which neighborhoods fail to correspond within a global context by computing the distance of π_i to the identity, $\text{id} \in S_M$. Specifically,

$$k \text{ scale misalignment at sample } i = \sum_{r=1}^{k} |r - \pi_i(r)| + |r - \pi_i^{-1}(r)|$$

While such an approach to comparing neighborhood construction would deliver some of our desired properties, such as being definable at all scales without saturation, it is still fundamentally non-geometric. One such geometric eventuality that the above statistic is blind to is high anisotropy, a key feature of many representational geometries in deep learning. In the highly dense regions that characterize these spaces, a large change in rank may be an effect of noise rather than carrying significant semantics, yet it can still be penalized drastically. We must further refine our approach to account for the global distribution of similarities to an anchor. A more information rich statistic would ask how close neighbors are to their original positions, when viewed in the opposite space, *in units of similarity*.

To measure this we define the "sorted similarity profiles" of i in each space. Let s^{X_i} be a vector of all similarities to the anchor X_i sorted in decreasing order so that the first entry is the similarity of the closest neighbor, and so on. We similarly define s^{Y_i}. The sorted profiles represent the neighbors in rank order in units of similarity to the anchor. We can then define the permuted profiles, $\pi_i^{-1}(s^{X_i})$ and $\pi_i(s^{Y_i})$ given by

$$\pi_i^{-1}(s^{X_i})_r = s^{X_i}_{\pi_i^{-1}(r)} \qquad \text{and} \qquad \pi_i(s^{Y_i})_r = s^{Y_i}_{\pi_i(r)}.$$

In other words, the r-th entry in the vector $\pi_i^{-1}(s^{X_i})$ is the similarity in X of the r ranked neighbor in Y. We are interested in using these objects to measure not just the rank changes between the two spaces, but the semantic consequences of those rank changes, encoded in geometry. We define two directional alignment measures called the "mutual neighbor similarity" at scale k and anchor i:

$$\text{MNS}_k^{X\to Y}(i) = \langle s^{Y_i}, \pi_i(s^{Y_i})\rangle_k \qquad \text{and} \qquad \text{MNS}_k^{Y\to X}(i) = \langle s^{X_i}, \pi_i^{-1}(s^{X_i})\rangle_k \tag{1}$$

where $\langle , \rangle_k$ is the truncated inner product, summing products of entries up to index k instead of M. $\text{MNS}_k^{X\to Y}(i)$ determines how aligned the k-neighborhood of X_i is to that of Y_i by asking how close the X closest neighbor is to the Y closest neighbor, measured in units of similarity to the anchor in Y, and so on. The alignment signal is carried by the change in neighbor rankings between the two spaces. This is the information encoded in π_i. However, the signal is measured in units of similarity, incorporating the geometries of X and Y.

MNS carries a consistent definition at all spatial scales and accounts for geometry without assuming a linear correspondence between similarities in X and Y. In fact similarities across spaces are never compared directly. MNS also handles ties naturally. It is immediate from the definition that given $r, r' \leq k$, with $r \neq r'$, if $s_r^{X_i} = s_{r'}^{X_i}$ and $s_r^{Y_i} = s_{r'}^{Y_i}$ then each directional MNS is invariant to a reordering of r and r' in π. When similarities are close in one space, the effect of a tie in the other is vanishing, rather than integer-valued. This is a structural advantage over ordinal methods which compute alignment in units of similarity ranking.

However, in its current form, the scale of MNS is subject to M and the choice of similarity kernel in X and Y. Furthermore, it has no statistically grounded baseline for what constitutes a strong measurement of alignment. In the next section, we fix these issues and define a symmetric aggregate measure across all anchors.

3 Adjusting MNS for robust measurements

We can analytically derive the maximum possible value of an MNS measurement as well as the mean under a carefully chosen null hypothesis and calibrate the raw statistic to remove the confounding

effects of point cloud cardinality and similarity kernel, and to establish a consistent and interpretable scale. We provide definitions and arguments in the direction $X \to Y$ but everything about our construction is symmetric. We define the adjusted mutual neighbor similarity as

$$\mathrm{AMNS}_k^{X\to Y}(i) = \frac{\mathrm{MNS}_k^{X\to Y}(i) - \mu_k^{X\to Y}(i)}{\max(\mathrm{MNS}_k^{X\to Y}(i)) - \mu_k^{X\to Y}(i)}. \tag{2}$$

Here,

$$\mu_k^{X\to Y}(i) = \mathbb{E}_{\pi_i \sim \mathrm{Unif}(S_M)}[\mathrm{MNS}_k^{X\to Y}(i) | s^{Y_i}],$$

or the expectation under the null hypothesis that the rank correspondence permutation π_i is a random variable: $\pi_i \sim \mathrm{Unif}(S_M)$, conditioned on the observed sorted profile s^{Y_i}. AMNS therefore naturally has a maximum value of 1, is positive for better-than-random alignment with respect to $\pi_i|_{[k]}$, and is exactly 0 at random alignment. Furthermore, AMNS is provably minimized when $\pi_i|_{[k]}$ is the anti-identity in S_M. In Appendix A we show this and derive the exact values of $\max(\mathrm{MNS}_k^{X\to Y}(i))$ and $\mu_k^{X\to Y}(i)$ solely as low compute functions of the already computed s^{Y_i}. A reliable closed form for $\max(\mathrm{MNS}_k^{X\to Y}(i))$ requires the assumption that all similarities are non-negative. As such, when we implement AMNS in practice, we use affine transformed cosine similarity: $\frac{\cos+1}{2}$ (havercosine kernel). Our derivations yield,

$$\max(\mathrm{MNS}_k^{X\to Y}(i)) = \sum_{r=1}^{k} \left(s_r^{Y_i}\right)^2 \quad \text{and} \quad \mu_k^{X\to Y}(i) = \overline{s^{Y_i}} \sum_{r=1}^{k} s_r^{Y_i}.$$

where $\overline{s^{Y_i}} = \frac{1}{M}\sum_{r=1}^{M} s_r^{Y_i}$ - the global mean of the similarity profile. Using these identities, we can express

$$\mathrm{AMNS}_k^{X\to Y}(i) = \frac{\langle s^{Y_i}, \pi(s^{Y_i}) - \overline{s^{Y_i}}\mathbf{1}\rangle_k}{\langle s^{Y_i}, s^{Y_i} - \overline{s^{Y_i}}\mathbf{1}\rangle_k} \quad (\text{resp. } Y \to X) \tag{3}$$

where $\mathbf{1}$ is the vector with 1 in every entry. While Equation 2 defines AMNS as a statistical normalization of MNS, the above reveals a geometric perspective as well. Specifically, AMNS is sensitive to the distribution of embedding similarities encoded by normalization with respect to the average similarity to the anchor, a reflection of both global anisotropy and local effects. AMNS therefore measures alignment by computing how close each neighbor is to its original ranking, in units of similarity, contextualized by the geometry of embeddings.

We can now define the symmetric form,

$$\mathrm{AMNS}_k^{XY}(i) = \frac{\mathrm{AMNS}_k^{X\to Y}(i) + \mathrm{AMNS}_k^{Y\to X}(i)}{2}$$

and the aggregate measure

$$\mathrm{AMNS}_k^{XY} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AMNS}_k^{XY}(i)$$

which is maximized when $\pi_{XY} = (\pi_1, \pi_2 \ldots, \pi_N) = \mathrm{id} \in S_M^N$. AMNS_k is a geometrically aware and naturally statistically calibrated measure of alignment with a highly interpretable and consistent range enabling alignment comparisons across model pairs and across spatial scales without saturation (Appendix B). Finally, we can also aggregate over k and form a scale free measure. In practice, we compute AMNS over a geometrically spaced k-grid, and so we use the $\log_{10}(k)$ normalized area to balance our observations across neighborhoods whose cardinalities vary over orders of magnitude:

$$\mathrm{AMNS}^{XY} = \frac{1}{\log(k_{\max}) - \log(k_{\min})} \int_{\log(k_{\min})}^{\log(k_{\max})} \mathrm{AMNS}_k^{XY}\, d\log(k).$$

Next, we demonstrate that AMNS carries all necessary invariances and sensitivities of reliable representational alignment measures Kornblith et al. (2019); Klabunde et al. (2025).

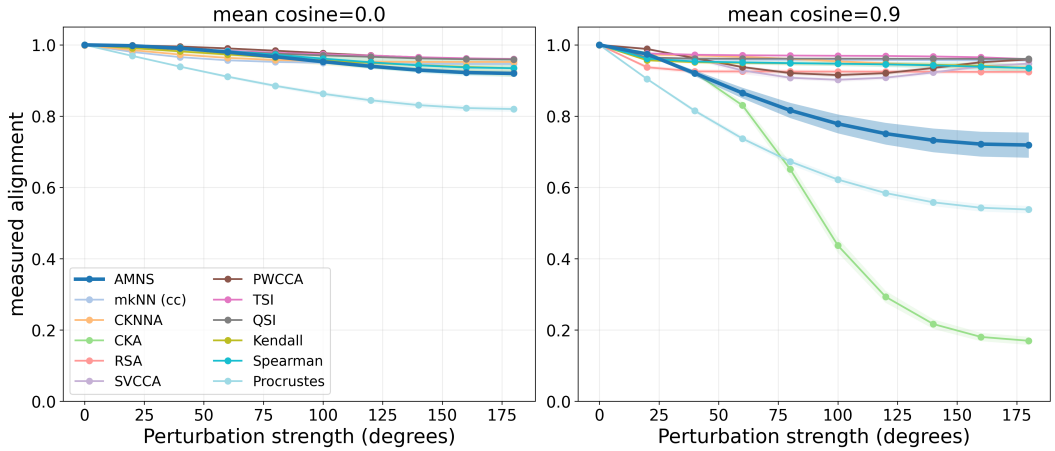

Figure 1: Degradation of different similarity measures when 2% of a synthetic dataset is perturbed with perturbation strengths ranging from negligible to antipodal. The left shows the response in a low anisotropy setting, measured by mean cosine similarity. The right is high anisotropy. In the low anisotropy setting, most degradation is limited, with only Procrustes alignment responding significantly to a small number of geometrically-consistent perturbations. Under high anisotropy, AMNS is unique in its ability to degrade gracefully in response to the perturbations without being overly sensitive to small perturbation strengths (Procrustes) or the small percentage of perturbed samples (CKA).

3.1 Invariances of AMNS

Invariance to orthogonal transformations, isotropic scalings, and neuron permutations AMNS is a composition of intra-space similarities defined by the havercosine kernel, which is preserved by orthogonal transformation, isotropic scaling, and neuron permutation. It is immediate that these invariances immediately carry over to AMNS.

Sensitivity to arbitrary invertible linear transformations We can disprove the statement that AMNS_k is invariant to arbitrary invertible linear transformations with a simple counter example. Notice that it is sufficient to find an example where $\mathrm{AMNS}_k^{XX'} < 1$ for $X' = LX$ with L an invertible linear transform. Consider an embedding space in $\mathbb{R}^2$ containing 3 points: $a = (1 \quad 1)$, $b = (1 \quad 0)$, and $c = (-1 \quad 0)$. We apply the invertible linear transform $L = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$ with $\det(L) = 1$. Then, $\pi_b = \pi_c = \mathrm{id} \in S_2$ but $\pi_a = (12)$. $s^0_{ac} = s^1_{ab} = \frac{2-\sqrt{2}}{4} = r_1$ and $s^0_{ab} = s^1_{ac} = \frac{2+\sqrt{2}}{4} = r_2$. Then $\mathrm{AMNS}_1(a) = \frac{1}{2}\left(\frac{r_1(r_2 - \frac{r_1+r_2}{2})}{r_1(r_1 - \frac{r_1+r_2}{2})} + \frac{r_1(r_2 - \frac{r_1+r_2}{2}}{r_1(r_1 - \frac{r_1+r_2}{2})}\right) = -1$. A slightly more tedious calculation shows $\mathrm{AMNS}_2(a) = -\frac{r_1^2+r_2^2}{(r_1-r_2)^2} < 0$ and so $\mathrm{AMNS}_k^{XX'}(a) < 1$ at both extremes $k = 1$ and $k = 2$.

4 Experiments

4.1 Controlled perturbation

Given a synthetic dataset with a geometry of interest, we can observe how similarity measures respond to controlled perturbations of the positions of a subset of its points. An optimal measure of alignment should respond coherently to the number of points being perturbed, perturbation strength, and features of the ambient point cloud geometry. For example, in high anisotropy settings, fully antipodal perturbations represent larger violations of the original geometry than they do in low anisotropy settings. As an extreme example, consider an anomaly detection model. In this setting, the small number of points considered anomalous (embedded far from a primary cluster) determines the model. If a perturbation of its representations changes the identities of the anomalous points, this should be catastrophic for alignment. As such, we analyze how a large library of canonical and

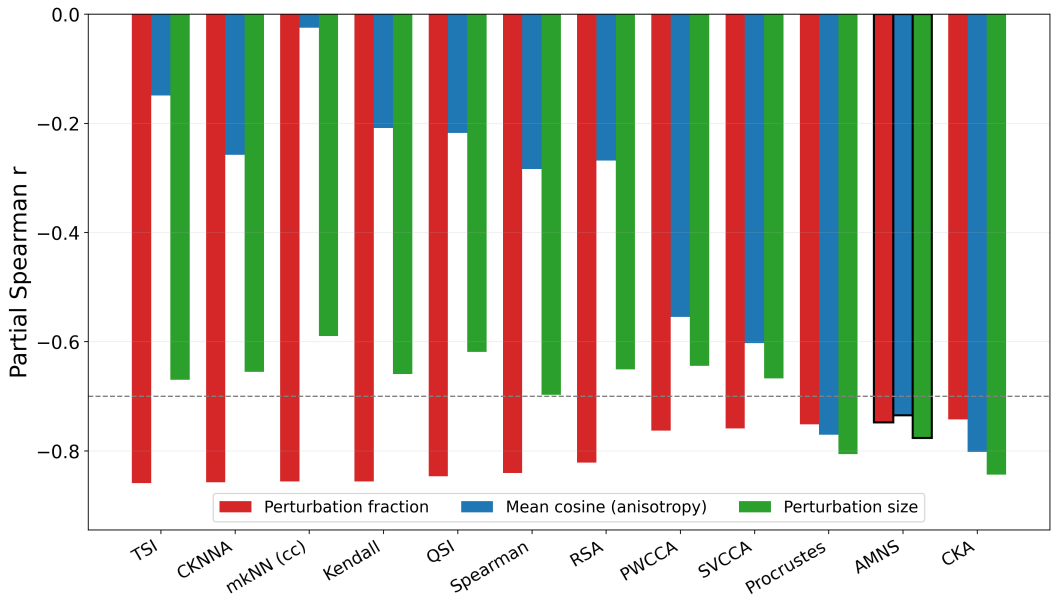

Figure 2: Partial Spearman correlation of each method with number of perturbed points (perturbation fraction), anisotropy, and perturbation size. Ordinal methods exhibit the failure mode of insensitivity to geometry. Among the sufficiently sensitive approaches (all partial correlations < -0.7), CKA and Procustes may exhibit outsized degradation responses when perturbation strength and/or perturbation fraction are low (Figure 1).

recently introduced measures (Appendix C) respond to perturbations of different sizes and strengths, and in different geometric settings. In these explorations, we use the scalar area under the curve of AMNS. The synthetic dataset cardinality is 10,000.

AMNS responds coherently to ambient geometry Figure 1 shows how different approaches respond to increasingly large perturbations of 2% of points in different geometric settings. In both settings, all alignment measures degrade as a function of perturbation strength. In the low anisotropy setting, degradation is limited except for Procrustes, which degrades significantly even when all perturbation factors are small. In the high anisotropy setting, CKA degrades catastrophically, Procrustes is again overly sensitive to even small perturbation strengths, and AMNS degrades gracefully while still registering strong alignment. All other methods fail to respond to the ambient geometry. We show in Appendix E that these findings carry over to perturbations of real data of varying geometries.

AMNS strongly correlates with all perturbation factors We generate synthetic data of different anisotropy levels, ranging from a mean cosine of 0 up to 0.99, perturb different sized subsets of its points, from 0.25% to 10%, with different angular perturbation strengths all the way up to antipodal perturbations. We compute the degradation response of different alignment methods over all triplets. Partial Spearman correlation then captures the ability of each measure to respond to each perturbation factor, while controlling for the other two. Figure 2 shows that only CKA, Procrustes, and AMNS exhibit strong negative correlation with each perturbation factor (< -0.7). However, we have seen that CKA and Procrustes are overly sensitive.

4.2 AMNS CURVES UNCOVER SHARED CLUSTERS IN EMBEDDING SPACES

A key feature of AMNS is its nearly unique ability to measure alignment at different scales. In fact, among common alignment measures, only CKNNA is also defined across scales without saturation effects. A question then, is whether we can go beyond discriminating between local and global alignment, and identify shared features in embedding spaces by scanning across neighborhood sizes. To test this, we generate two hierarchically clustered embedding spaces with corresponding clusters

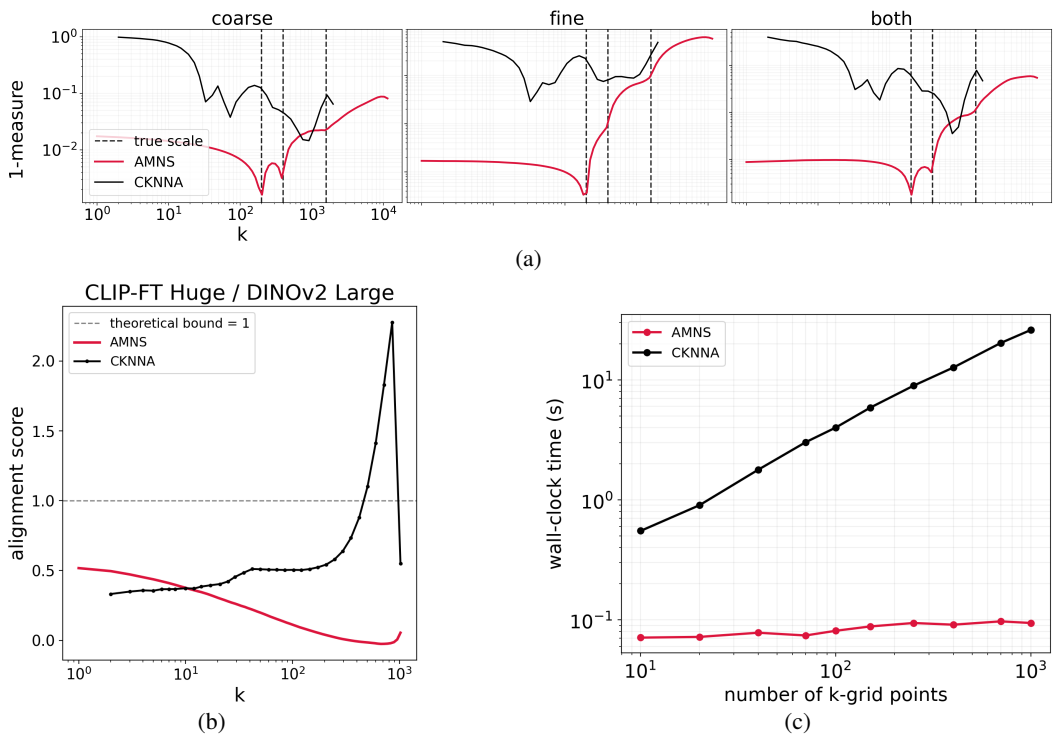

Figure 3: (a) AMNS and CKNNA curves over different configurations of hierarchically clustered data. Each vertical line marks the size of clusters that are shared between embedding spaces. The critical points of the AMNS curve capture the shared hierarchical clustering between the spaces. (b) AMNS and CKNNA curves for a pair of canonical vision models on the Wiki-based image-text dataset Srinivasan et al. (2021). (c) Log-log plot of runtime to compute alignment curves vs. number of neighborhood sizes tested.

and measure the ability of the AMNS and CKNNA curves to reveal the shared hierarchical clustering structure. We test 1, 2, and 3 levels of clustering and grow the cardinality of the embedding spaces accordingly. We also vary the level of agreement between the two geometries, defined by three regimes, depending on whether the internal distribution of points within clusters is shared ("fine" regime), the large scale distribution of clusters in space is shared ("coarse"), or if both properties hold simultaneously ("both").

In all configurations, we apply a random affine projection of each view into two spaces of different dimensions before adding Gaussian noise of modest variance. In every experimental setting, across the 9 total configurations and 20 seeds per configuration, the critical points of the AMNS curve perfectly recover the cluster sizes, reliably peaking whenever geometric structure is maximally shared (Figure 3a).

4.3 AMNS ENABLES RELIABLE COMPARISONS ACROSS SCALES

Spatial scale has emerged as an important axis for alignment comparisons. Recent findings suggest that frontier models of different modalities are more aligned locally than they are globally and that local spatial structure is a key generalization feature of the target representation in alignment-based distillation tasks Gröger et al. (2026); Singh et al. (2025). Currently, AMNS is uniquely suited to support these investigations, as no other alignment measure is equipped with a well-behaved scale parameter. To demonstrate this, we again compare to CKNNA. We perform a pairwise comparison of all distinct models in the experimental suite of the influential work, "The Platonic Representation Hypothesis" (Huh et al., 2024) up to 7 billion parameters. Figure 3b shows AMNS and CKNNA curves on this data. AMNS remains coherent in all cases, while CKNNA exhibits a tendency to inflate beyond its own theoretical bound as k grows before returning to measuring standard CKA

Rank	Method	P / baseline	Rank	Method	P / baseline
1	AMNS_{10}	161 ± 0.769	9	Spearman ρ	138 ± 0.561
2	AMNS_{M}	155 ± 0.958	10	Kendall τ	138 ± 0.529
3	cc m-kNN ($k = 10$)	152 ± 0.703	11	CKNNA (AUC)	137 ± 0.385
4	Procrustes	143 ± 0.752	12	TSI	137 ± 0.604
5	AMNS (AUC)	142 ± 0.534	13	CKNNA ($k = 10$)	134 ± 0.395
6	QSI	139 ± 0.885	14	SVCCA	116 ± 0.372
7	CKA	139 ± 0.567	15	PWCCA	106 ± 0.621
8	RSA	130 ± 0.903	Results scaled by 10^2 for readability		

Table 1: Median $\pm$ standard error of the average precision over baseline of classifiers fit on alignment measures to predict ImageNet-1k performance gaps between model pairs. Over 31,125 pairs, AMNS_{10} achieves an improvement over baseline $\sim 17\%$ higher than the best non-AMNS method.

at $k = N$. This was observed in $143/300$ cases. Additionally, the formulation of AMNS allows values at different neighborhood sizes to be computed from the same cumulative sums, enabling high resolution sampling while incurring negligible computational cost (Figure 3c). We expand on the computational complexity of AMNS in Section 5. CKNNA renormalizes masked Gram matrices with respect to each k, incurring a linear cost in the number of neighborhood sizes sampled.

4.4 AMNS PREDICTS FUNCTIONAL ALIGNMENT

A primary application of representational alignment is knowledge distillation and transfer learning Tian et al. (2022); Bhattarai et al. (2025); Park et al. (2019). To demonstrate the viability of AMNS in this domain, we seek to test how well it predicts functional similarity between models. We sample the final-layer semantic representations of ImageNet-1k from 250 models in the SEMASIA dataset (Pandolfo et al., 2026). For each of the resulting $\text{C}(250, 2) = 31,125$ unique model pairs, we compute the absolute gap in performance and the alignment reported by each measure. Unaligned models may have coincidentally small performance gaps, invalidating simple correlations between alignment and performance gap vectors. We are instead interested in precision: when a measure reports significantly high alignment, does that reliably predict a small performance gap? We threshold the performance gaps at different percentiles (5% closest pairs, 10%, 15%, 20%, 25%) and train one-feature decision tree classifiers to predict small performance gaps from alignment. We record the ratio of the precision of each decision tree over a random/label mean baseline and take the average across thresholds for each alignment measure to obtain a robust measurement. The procedure is then repeated over 40 splits of the data and medians are reported alongside standard errors. For every model in the dataset, all pairs containing it are limited to either the classifier training set or the holdout set for every split of the data. Table 1 shows that AMNS alignment is the strongest predictor of functional alignment of all measures tested. **In future work, at a time when we are less computationally constrained, we intend to build on these encouraging results and directly test the utility of AMNS in transfer learning.**

5 COMPUTATIONAL COMPLEXITY

Due to the formulation of AMNS as a mean over pointwise measurements, we are able to compute the aggregate measure in batches of size B, enabling large scale analyses. The time complexity of computing the raw $B \times N$ similarity matrices is $\mathcal{O}(BN(d_X + d_Y))$. Sorted similarity profiles and alignment permutations are concurrently generated by a full row-wise sort, incurring a cost of $\mathcal{O}(BN \log(N))$ with gather operations generating the two permuted profiles at a cost of $\mathcal{O}(BN)$. Finally the numerator and denominator of $\text{AMNS}_k^{X \to Y}$ are each indexed from cumulative sums: $\sum_r^k s_r^Y \pi(s^Y)_r, \sum_r^k s_r^Y, \sum_r^k (s_r^Y)^2$, (resp. $Y \to X$) which are each $\mathcal{O}(BN)$ incurring no extra cost per k sampled. The complexity at this step is determined by the largest neighborhood size considered only. The mean over the full sorted profiles used to normalize AMNS is a one time calculation that scales as $\mathcal{O}(BN)$ as well. The total time complexity over all batches is then $\mathcal{O}(N^2(d_X + d_Y + \log(N)))$, exactly on par with or below canonical methods (Table 2), while providing a uniquely information-rich and assumption-free measure. Additionally, we can estimate

Method	Time complexity
AMNS, cc m-kNN, Kendall τ, Spearman ρ	$\mathcal{O}\big(N^2(d_X + d_Y + \log N)\big)$
CKNNA	$\mathcal{O}\big(N^2(d_X + d_Y) + KN^3\big)$
CKA (debiased linear)	$\mathcal{O}\big(N(d_X^2 + d_Y^2 + d_X d_Y)\big)$
RSA	$\mathcal{O}\big(N^2(d_X + d_Y + \log N)\big)$
SVCCA	$\mathcal{O}\big(N(d_X^2 + d_Y^2) + \max(d_X, d_Y)^3\big)$
PWCCA	$\mathcal{O}\big(N(d_X^2 + d_Y^2 + d_X d_Y) + \max(d_X, d_Y)^3\big)$
Procrustes shape metric	$\mathcal{O}\big(N d_X d_Y + \min(d_X, d_Y)^2 \max(d_X, d_Y)\big)$
TSI, QSI	$\mathcal{O}\big(N^2 \log N\,(d_X + d_Y)\big)$

Table 2: Computational complexity of every measure tested against AMNS. N: full dataset size; d_X, d_Y: embedding dimensions; K: number of k-grid points evaluated; AMNS reads every k off one sort, so K does not appear in its complexity; CKNNA incurs its full masking and HSIC cost at every k. Full descriptions of each method appear in Appendix C.

AMNS by restricting to a subset of anchor points of size A, bringing the complexity down to $\mathcal{O}(AN(d_X + d_Y + \log(N)))$. We show empirically in Appendix D.3, the considerable reliability of estimating AMNS in different settings. However, in practice it is quite feasible to compute AMNS with all samples as anchors in a single batch.

6 Discussion

The diverse semantics of different representational similarity measures each invite strengths and weaknesses related to reliability and interpretation, information content of the alignment signal, usability at scale, and granularity of the analysis. We introduce mutual neighbor similarity as the basis for a hybrid approach between the geometric and ordinal families of measures that seeks to marry as many of the strengths of existing methods as possible in a single measure, while mitigating weaknesses. The final form, AMNS, offers multi-scale consistency without saturation, geometric sensitivity without linearity assumptions, and analytical calibration without Monte Carlo simulation. Of particular note, it is the only measure that enables reliable alignment comparisons across spatial scales. We demonstrate empirically that this property, in addition to supporting meaningful investigations of alignment (Gröger et al., 2026), enables the discovery of shared clusters between spaces (Section 4.2). This points to interesting directions for further interpreting AMNS measurements such as investigations into the meaning of global maxima of AMNS curves over real world data. Other dissections of AMNS, such as along the anchor axis, also represent interesting possibilities, such as identifying the semantics and statistics of divergent representations Hosseini et al. (2024). Finally, we demonstrate that the rich semantics of AMNS carry over to strong predictions of functional alignment between models (Section 4.4).

However, limitations remain. In particular, while AMNS can make measurements at different scales, it is unclear a priori which form of AMNS (local, global, aggregate) will be the most suitable for any task. In fact, our functional alignment experiment in Section 4.4 shows that the most information dense form, AMNS (AUC), actually underperforms relative to the single-k variants in that setting. Further investigation is required to understand the specific semantics of different forms of AMNS in order to properly match them to tasks. However, it is notable that AMNS can be computed over all possible scales at no extra cost over the computation at $k = M$, easing the burden of such explorations. The experimental evidence of the value of AMNS for important tasks such as transfer learning make this deeper exploration an exciting direction. As well, the discrepancy in the performance of different AMNS variants at a task may also illuminate which forms of alignment are inherently relevant to that task. For example, our experiments may be pointing to the primacy of local and global alignment over mesoscale alignment in predicting functional similarity. Lastly, large scale comparisons of many models with respect to representational alignment are of increasing experimental interest Williams et al. (2022). Clustering and other metric space-based analyses require similarity measures that give rise to proper mathematical metrics (or at least pseudometrics) (Yianilos, 1993; Dasgupta & Long, 2005). While it is not immediately obvious how to associate AMNS with a metric while retaining its full geometric descriptive power, this represents an exciting direction for immediate further investigation.

AI Use Statement In this work, we used Claude to aid in our literature review and certain synthetic data generation tasks, namely for our hierarchical clustering experiments. We have not used generative AI tools otherwise. We have reviewed all AI-assisted work. Every baseline used in this work was implemented in code by the human authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

David GT Barrett, Ari S Morcos, and Jakob H Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, 55: 55–64, 2019. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2019.01.007. URL https://www.sciencedirect.com/science/article/pii/S0959438818301569. Machine Learning, Big Data, and Neuroscience.

Prajjwal Bhattarai, Mohammad Amjad, Dmytro Zhylko, and Tuka Alhanai. Knowledge distillation through geometry-aware representational alignment, 2025. URL https://arxiv.org/abs/2509.25253.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment, 2024. URL https://arxiv.org/abs/1203.0550.

Sanjoy Dasgupta and Philip M. Long. Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, 70(4):555–569, 2005. doi: 10.1016/j.jcss.2004.10.006.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers, 2024. URL https://arxiv.org/abs/2401.12143.

John C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. doi: 10.1007/BF02291478.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view, 2026. URL https://arxiv.org/abs/2602.14486.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Eghbal Hosseini, Colton Casto, Noga Zaslavsky, Colin Conwell, Mark Richardson, and Evelina Fedorenko. Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024. doi: 10.1101/2024.12.26.629294. URL https://www.biorxiv.org/content/early/2024/12/26/2024.12.26.629294.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Nori Jacoby, Eduardo A Undurraga, Malinda J McPherson, Joaquín Valdés, Tomás Ossandón, and Josh H McDermott. Universal and non-universal features of musical pitch perception revealed by singing. *Current Biology*, 29(19):3229–3243, 2019.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Comput. Surv.*, 57(9), May 2025. ISSN 0360-0300. doi: 10.1145/3728458. URL https://doi.org/10.1145/3728458.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models, 2020. URL https://arxiv.org/abs/2007.04612.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019. URL https://arxiv.org/abs/1905.00414.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008. URL https://frontiersin.org.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, 2022. URL https://arxiv.org/abs/2203.02053.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation, 2018. URL https://arxiv.org/abs/1806.05759.

Alex Murphy, Joel Zylberberg, and Alona Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks, 2024. URL https://arxiv.org/abs/2405.01012.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations, 2025. URL https://arxiv.org/abs/2211.01201.

Mario Edoardo Pandolfo, Enrico Grimaldi, Lorenzo Marinucci, Leonardo Di Nino, Simone Fiorellino, Sergio Barbarossa, and Paolo Di Lorenzo. Semasia: A large-scale dataset of semantically structured latent representations, 2026. URL https://arxiv.org/abs/2605.09485.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation, 2019. URL https://arxiv.org/abs/1904.05068.

Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations, 2018. URL https://arxiv.org/abs/1706.02417.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, 2017. URL https://arxiv.org/abs/1706.05806.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure?, 2025. URL https://arxiv.org/abs/2512.10794.

Diogo Soares, Pankhil Gawade, Andrea Dittadi, and Ewa Szczurek. Scalable and interpretable representation alignment with ordinal similarity, 2026. URL https://arxiv.org/abs/2606.16379.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47):1393–1434, 2012. URL http://jmlr.org/papers/v13/song12a.html.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 2443–2449. ACM, july 2021. doi: 10.1145/3404835.3463257. URL http://dx.doi.org/10.1145/3404835.3463257.

Ilia Sucholutsky and Thomas L. Griffiths. Alignment with human representations supports robust few-shot learning, 2023. URL https://arxiv.org/abs/2301.11990.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation, 2022. URL https://arxiv.org/abs/1910.10699.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations, 2022. URL https://arxiv.org/abs/2110.14739.

Andrea Wynn, Ilia Sucholutsky, and Thomas L. Griffiths. Learning human-like representations to enable learning human values, 2024. URL https://arxiv.org/abs/2312.14106.

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

Peter N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '93)*, pp. 311–321, Philadelphia, PA, USA, 1993. Society for Industrial and Applied Mathematics.

A Appendix: Proof of the form of AMNS

Proposition. *If $s_M^{Y_i} \geq 0$ then for every $1 \leq k \leq M$, $\mathrm{MNS}_k^{X\to Y}(i) \leq \sum_{r=1}^k (s_r^{Y_i})^2$ with equality if and only if $\pi_i(r) = r$ for every $1 \leq r \leq k$ up to permutations within blocks of exactly tied similarities.*

Proof. $\mathrm{MNS}_k^{X\to Y}(i) = \sum_{r=1}^k s_r^{Y_i} s_{\pi_i(r)}^{Y_i}$. When $k = M$ the claim is exactly the rearrangement inequality, due to s^{Y_i} being nonincreasing. For $k < M$, the image of $\pi_i|_{[k]}$ is not restricted to $[k]$ so an additional step is necessary when that is not the case. If $\pi_i([k]) \neq [k]$ then there exists $r \leq k$, $p \in [k] \setminus \pi_i([k])$ and $q = \pi_i(r) \in \pi_i([k]) \setminus [k]$ such that $s_p^{Y_i} \geq s_q^{Y_i}$. Changing π_i so that $\pi_i(r) = p$ changes the sum by exactly $s_r^{Y_i}(s_p^{Y_i} - s_q^{Y_i}) \geq 0$ under our assumption that all similarities are non-negative. Iterating through $[k]$, we see that $\mathrm{MNS}_k^{X\to Y}(i)$ is maximized when $\pi_i([k]) = [k]$ which restores the case where the claim is identical to the rearrangement inequality. □

Notice that an inversion of the above argument shows that MNS is minimized when $\pi_i(r) = M - r + 1$ for $1 \leq r \leq k$.

Proposition. *$\mu_k^{X\to Y}(i) = \overline{s^{Y_i}} \sum_{r=1}^k s_r^{Y_i}$ where $\overline{s^{Y_i}} = \frac{1}{M}\sum_{r=1}^M s_r^{Y_i}$: the global mean similarity to anchor i in Y.*

Proof. Under the conditional null,

$$\mu_k^{X\to Y}(i) = \mathbb{E}_{\pi_i \sim \mathrm{Unif}(S_M)}[\mathrm{MNS}_k^{X\to Y}(i) | s^{Y_i}].$$

The symmetric group S_M has cardinality $M!$. For every pair $1 \leq r, q \leq M$ there are precisely $(M-1)!$ permutations in S_M where $\pi(r) = q$. Therefore,

$$\mu_k^{X\to Y}(i) = \frac{1}{M!}\sum_{\pi\in S_M}\sum_{r=1}^k s_r^{Y_i} s_{\pi(r)}^{Y_i} = \frac{(M-1)!}{M!}\sum_{q=1}^M\sum_{r=1}^k s_r^{Y_i} s_q^{Y_i} = \overline{s^{Y_i}}\sum_{r=1}^k s_r^{Y_i}$$

□

B Appendix: AMNS is well characterized at the extremes

We briefly demonstrate that AMNS does not degenerate at the extremes $k = 1$ or $k = M$. Let s be the similarity profile in Y of the event indexed by i with alignment permutation $\pi = \pi_i$. Notice,

$$\mathrm{AMNS}_1^{X\to Y}(i) = \frac{\pi(s)_1 - \overline{s}}{s_1 - \overline{s}}$$

which is maximized by any $\pi \in S_M$ that associates the closest neighbor in X with any similarity maximizing neighbor in Y, naturally accounting for ties. Minimization is analogous.

Proposition. $\mathrm{AMNS}_M^{X\to Y}(i) = \rho(s, \pi(s))$ *(resp. $Y \to X$).*

Proof. The statement is a direct result of the fact that at the global scale $k = M$, all truncated objects converge to global objects and $\overline{s} = \overline{\pi(s)}$. To proceed, we will also make repeated use of the general identity: $\langle v - \overline{v}\mathbf{1}, \overline{v}\mathbf{1}\rangle = 0$ which is true due to $\overline{v}\mathbf{1} = \mathrm{proj}_{\mathbf{1}}(v)$. Recall the geometric form of AMNS (Equation 3). When $k = M$, the numerator can be rewritten:

$$\begin{aligned}
\langle s, \pi(s) - \overline{s}\mathbf{1}\rangle &= \langle s, \pi(s) - \overline{\pi(s)}\mathbf{1}\rangle \\
&= \langle s + \overline{s}\mathbf{1} - \overline{s}\mathbf{1}, \pi(s) - \overline{\pi(s)}\mathbf{1}\rangle \\
&= \langle s - \overline{s}\mathbf{1}, \pi(s) - \overline{\pi(s)}\mathbf{1}\rangle \\
&= M\mathrm{Cov}(s, \pi(s))
\end{aligned}$$

Similarly, the denominator can be rewritten

$$\begin{aligned}
\langle s, s - \overline{s}\mathbf{1}\rangle &= \langle s - \overline{s}\mathbf{1}, s - \overline{s}\mathbf{1}\rangle \\
&= \langle s - \overline{s}\mathbf{1}, s - \overline{s}\mathbf{1}\rangle^{\frac{1}{2}} \langle \pi(s) - \overline{\pi(s)}\mathbf{1}, \pi(s) - \overline{\pi(s)}\mathbf{1}\rangle^{\frac{1}{2}} \\
&= M\sigma_s\sigma_{\pi(s)}
\end{aligned}$$

and so $\mathrm{AMNS}_M^{X\to Y} = \frac{M\mathrm{Cov}(s,\pi(s))}{M\sigma_s\sigma_{\pi(s)}} = \rho(s, \pi(s))$. □

C Appendix: Baselines

Here we describe the baseline methods compared to in this work, with many being definable with respect to the same central rank permutation object used by AMNS.

m-kNN / chance-corrected m-kNN (Huh et al., 2024) π_X, π_Y are the rank permutations of the M neighbors to an anchor with respect to similarities in X and Y:

$$J_k = \frac{1}{k}\big|\{j : \pi_X(j) \le k\} \cap \{j : \pi_Y(j) \le k\}\big|, \qquad \text{ccm-kNN}_k = \frac{J_k - k/M}{1 - k/M}.$$

CKNNA (Huh et al., 2024) With $K = XX^\top$, $L = YY^\top$ and mutual top-k mask $\mathrm{Mask}_k[\cdot, j] = \mathbb{1}[\pi_X(j) \le k] \cdot \mathbb{1}[\pi_Y(j) \le k]$ (rank permutations as above, applied per row):

$$\mathrm{CKNNA}_k = \frac{\mathrm{HSIC}(\mathrm{Mask}_k \odot K,\ \mathrm{Mask}_k \odot L)}{\sqrt{\mathrm{HSIC}(\mathrm{Mask}_k \odot K,\ \mathrm{Mask}_k \odot K)\,\mathrm{HSIC}(\mathrm{Mask}_k \odot L,\ \mathrm{Mask}_k \odot L)}},$$

$\mathrm{HSIC}(K, L) = \mathrm{tr}(KHLH)$ (biased) or the unbiased U-statistic estimator introduced by Song et al. (2012), $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$.

CKA (debiased linear) (Kornblith et al., 2019) Alignment on geometric similarity structure:

$$\mathrm{CKA} = \frac{\mathrm{HSIC}(K, L)}{\sqrt{\mathrm{HSIC}(K, K)\,\mathrm{HSIC}(L, L)}}, \qquad K = XX^\top,\ L = YY^\top,$$

with HSIC debiased via the unbiased dot-product estimator introduced by Song et al. (2012).

RSA (Kriegeskorte et al., 2008) Correlation of two representational dissimilarity matrices with $\mathrm{RDM}^X_{ij} = 1 - \cos(x_i, x_j)$ (resp. Y):

$$\mathrm{RSA} = \rho_{\mathrm{Spearman}}\big(\mathrm{vec}_{i<j}(\mathrm{RDM}^X),\ \mathrm{vec}_{i<j}(\mathrm{RDM}^Y)\big),$$

where $\rho_{\mathrm{Spearman}}(u, v) = \mathrm{corr}(\mathrm{rank}(u), \mathrm{rank}(v))$ is the Pearson correlation of the rank-transformed vectors (Spearman, 1904).

SVCCA (Raghu et al., 2017) Subspace/correlation alignment, where ρ_i are the canonical correlations of the 99%-variance SVD-truncated representations:

$$\mathrm{SVCCA} = \frac{1}{m}\sum_{i=1}^{m} \rho_i.$$

PWCCA (Morcos et al., 2018) ρ_i are the canonical correlations of representations, h_i the i-th canonical variable of X, and weights $\tilde{\alpha}_i = \sum_j |\langle h_i, x_j \rangle|$:

$$\mathrm{PWCCA} = \sum_i \alpha_i \rho_i, \qquad \alpha_i = \tilde{\alpha}_i \big/ \textstyle\sum_j \tilde{\alpha}_j.$$

Procrustes shape metric (Gower, 1975; Williams et al., 2022) Optimal orthogonal alignment distance. With X_c, Y_c the mean-centered, Frobenius-normalized representations and ρ_ℓ the singular values of $Y_c^\top X_c$ (the alignment achieved by the optimal orthogonal map, Williams et al. Prop. 2):

$$\theta(X, Y) = \arccos\Big(\textstyle\sum_\ell \rho_\ell\Big).$$

Kendall τ (Kendall, 1938) π is the rank permutation mapping an anchor's X-rank order to its Y-rank order over the M neighbors. $\mathrm{inv}(\pi) = |\{(j, k) : j < k,\ \pi(j) > \pi(k)\}|$:

$$\tau = 1 - \frac{4\,\mathrm{inv}(\pi)}{M(M-1)}.$$

Spearman ρ (Spearman, 1904) Same π as above:

$$\rho = 1 - \frac{6\sum_{j=1}^{M}(\pi(j) - j)^2}{M(M^2 - 1)}.$$

TSI (Soares et al., 2026) Anchor-relative ordinal agreement. With $\pi_{X,i}, \pi_{Y,i}$ the rank permutations of distances from anchor i:

$$\mathrm{TSI} = \Pr_{(i,j,k)\sim T}\Big[\operatorname{sign}\big(\pi_{X,i}(j) - \pi_{X,i}(k)\big) = \operatorname{sign}\big(\pi_{Y,i}(j) - \pi_{Y,i}(k)\big)\Big].$$

QSI (Soares et al., 2026) Not anchor-relative – compares two independent pairwise distances:

$$\mathrm{QSI} = \Pr_{(i,j,k,l)\sim Q}\Big[\operatorname{sign}\big(d_X(i,j) - d_X(k,l)\big) = \operatorname{sign}\big(d_Y(i,j) - d_Y(k,l)\big)\Big].$$

D Appendix: Robustness and efficiency checks

D.1 AMNS is not confounded by embedding dimension

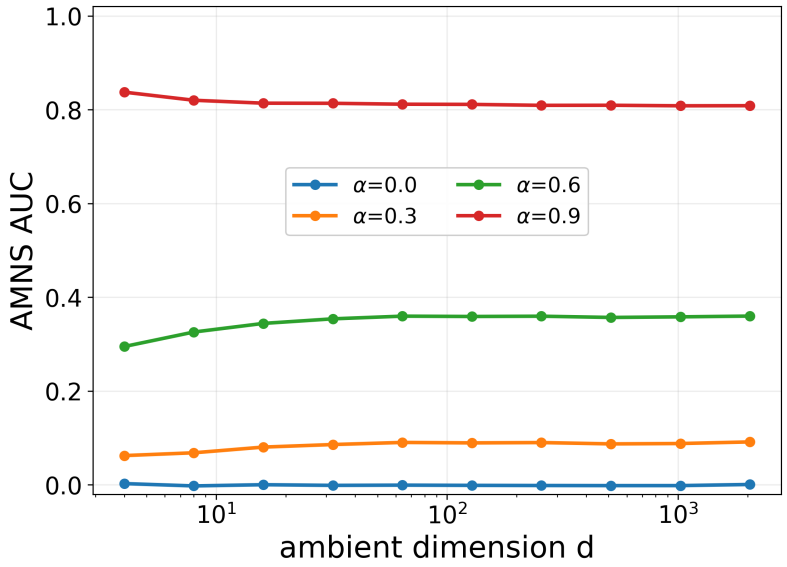

Figure 4: AMNS AUC vs. ambient embedding dimension d (4 to 2048), at four fixed levels of true alignment. True alignment is determined from a synthetic construction: X is a matrix ℓ_2-normalized i.i.d. Gaussian rows and $Y = \mathrm{rownorm}(\alpha X + \sqrt{1-\alpha^2}\, E)$ for independent Gaussian noise E. Since the expected dot product of two independent gaussian vectors is 0, $\mathbb{E}[\cos(x_i, y_i)] \approx \alpha$ regardless of d. The within-α range across the full d sweep is 0.005–0.065; between-α range at fixed d is 0.81–0.83 – a worst-case ratio of 0.081. In other words, d explains at most $\sim$8% as much AUC variation as true alignment does. Even this small residual movement is confined to the rare regime $d \lesssim 30$.

D.2 AMNS IS NOT CONFOUNDED BY NEIGHBORHOOD SIZE K

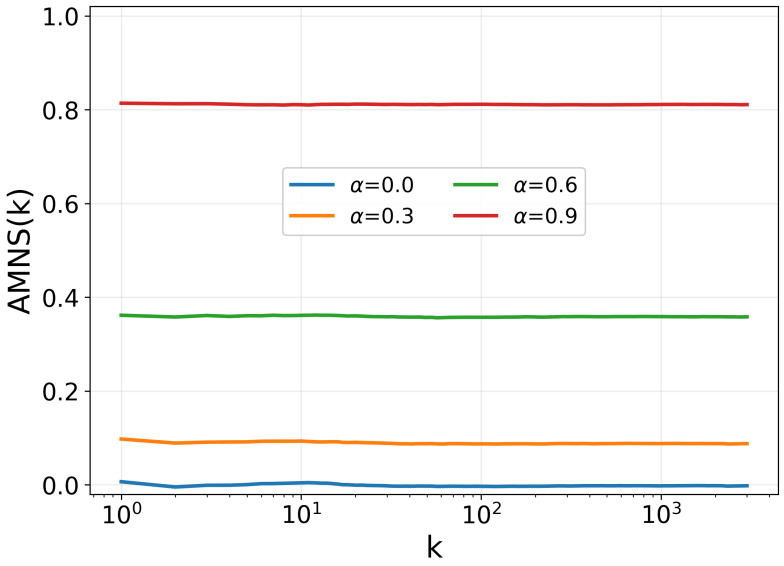

Figure 5: AMNS_k vs. neighborhood size k at the same four fixed levels of true alignment α described in Figure 4 for $N = 3000$, $d = 128$. Within-α range across the entire k-grid is 0.004–0.011; between-α range at matched k is 0.806–0.817 – a worst-case ratio of 0.013.

D.3 AMNS IS WELL-APPROXIMATED BY A SMALL NUMBER OF ANCHORS

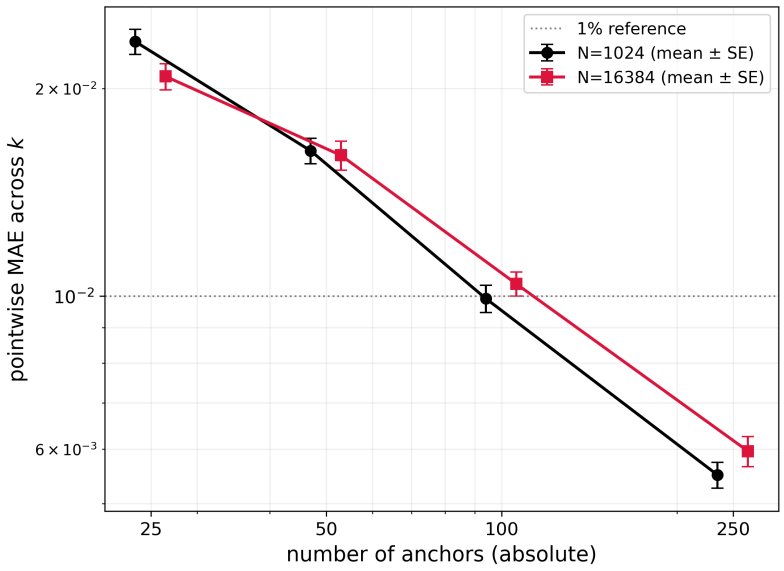

Figure 6: Mean MAE $\pm$ standard errors across all 32 pairs $\times$ 5 seeds vs. absolute anchor count, at $N = 1024$ and $N = 16384$. The two curves are nearly indistinguishable: at 250 anchors, median MAE is 0.0050 for $N = 1024$ and 0.0056 for $N = 16384$ despite a 16$\times$ difference in N. As N grows, the number of anchors needed to reasonably approximate AMNS for a given approximation quality remains stable.

We sample 32 model pairs from SEMASIA (Pandolfo et al., 2026) with diverse AMNS AUC values ranging from -0.10 to 0.98 (computed over $N = 1024$ anchors). We then subsample anchors and

record the mean absolute error across k between the full reference AMNS curve and the scores due to subsampling. We repeat this process over 5 subsampling seeds totaling $32 \times 5 = 160$ diverse observations. The reported mean MAE and standard error bars show within-1% convergence by approximately 100 anchors (Figure 6). We repeat this entire process for reference AMNS curves computed with $N = 16,384$ samples and observe the same absolute trend. This is evidence that AMNS is well approximated by a modest number of anchors regardless of the underlying sample size.

E Appendix: AMNS responds coherently to ambient geometry of real world data

We repeat the perturbation experiments of Section 4.1 on real world data, using the models and underlying dataset of Huh et al. (2024). Figure 7 shows that our observations about the unique coherent response of AMNS to ambient geometry carries over to perturbations of real data.

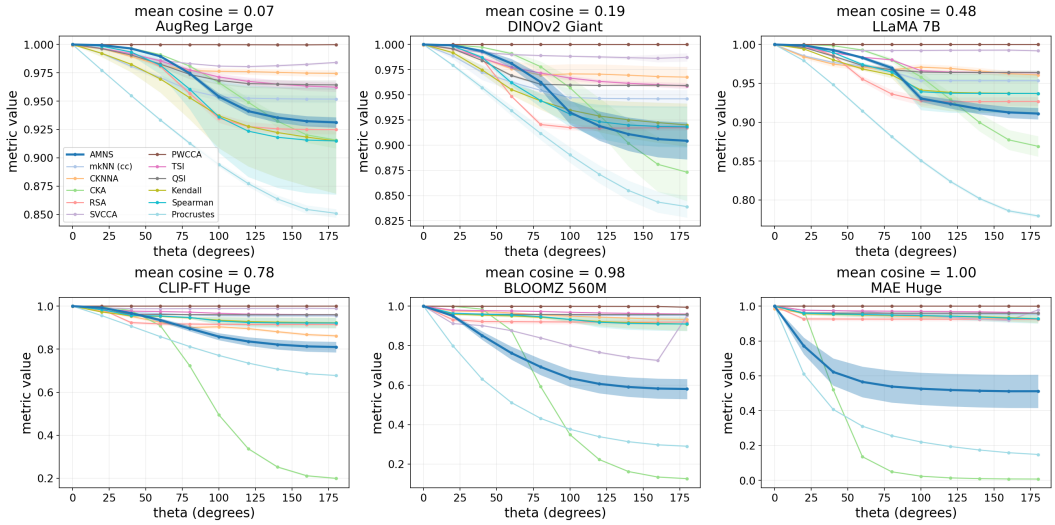

Figure 7: Degradation of different similarity measures when 2% of a real dataset is perturbed with perturbation strengths ranging from negligible to antipodal. The left shows the response in a low anisotropy setting. The right is high anisotropy.