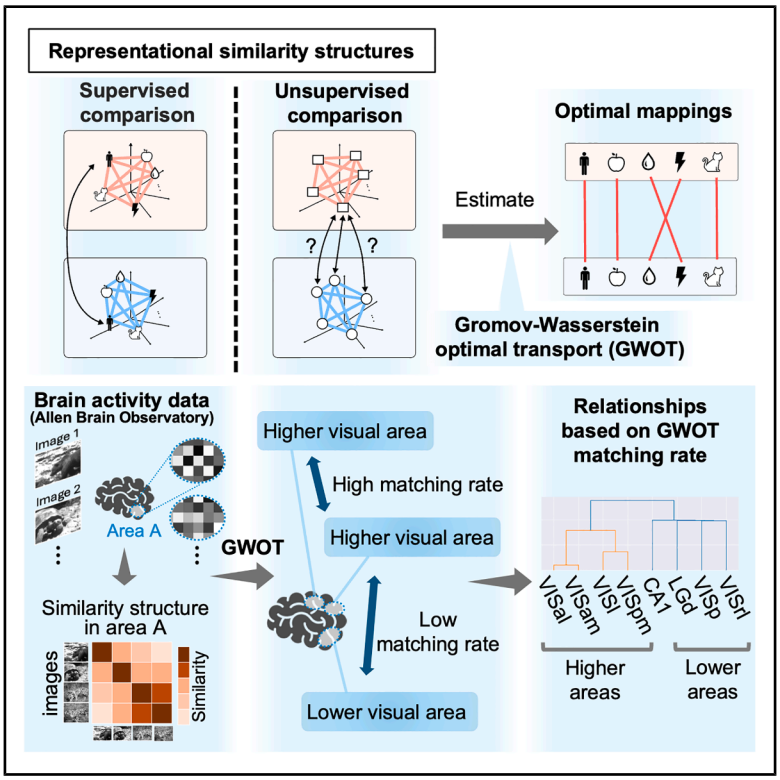
iScience

Article

Unsupervised alignment reveals structural commonalities and differences in neural representations of natural scenes across individuals and brain areas

Graphical abstract



Highlights

- Unsupervised alignment of representational similarity structures across visual areas
- Applied to large datasets of neural responses to visual stimuli in mice and humans
- Demonstrated the structural commonalities across individuals within same brain areas
- Alignable structures across different higher-order visual areas in both species

Authors

Ken Takeda, Kota Abe, Jun Kitazono, Masafumi Oizumi

Correspondence

c-oizumi@g.ecc.u-tokyo.ac.jp

In brief

Cognitive neuroscience; Psychology

Takeda et al., 2025, iScience *28*, 112427
May 16, 2025
https://doi.org/10.1016/j.isci.2025.112427

CellPress

iScience

CellPress
OPEN ACCESS

Article

Unsupervised alignment reveals structural commonalities and differences in neural representations of natural scenes across individuals and brain areas

Ken Takeda,[1,3] Kota Abe,[1,3] Jun Kitazono,[2] and Masafumi Oizumi[1,4,*]
[1]Graduate School of Arts and Science, The University of Tokyo, Tokyo, Japan
[2]Graduate School of Data Science, Yokohama City University, Kanagawa, Japan
[3]These authors contributed equally
[4]Lead contact
*Correspondence: c-oizumi@g.ecc.u-tokyo.ac.jp
https://doi.org/10.1016/j.isci.2025.112427

SUMMARY

Neuroscience research aims to identify universal neural mechanisms underlying sensory information encoding by comparing neural representations across individuals, typically using Representational Similarity Analysis. However, traditional methods assume direct stimulus correspondence across individuals, limiting the exploration of other possibilities. To address this, we propose an unsupervised alignment framework based on Gromov-Wasserstein Optimal Transport, which identifies correspondences between neural representations solely from internal similarity structures, without relying on stimulus labels. Applying this method to Neuropixels recordings in mice and fMRI data in humans viewing natural scenes, we found that the neural representations in the same visual cortical areas can be well aligned across individuals in an unsupervised manner. Furthermore, alignment across different brain areas is influenced by factors beyond the visual hierarchy, with higher-order visual areas aligning well with each other, but not with lower-order areas. This unsupervised approach reveals more nuanced structural commonalities and differences in neural representations than conventional methods.

INTRODUCTION

In the neuroscience field, researchers have long investigated the commonality of neural representations of sensory stimuli across individuals in an attempt to find universal neural mechanisms for the encoding of sensory information. It is typically assumed that sensory information is represented as the population activity of neurons or brain areas.[1,2] The difficulty in comparing neural representations across different brains is that there are no correspondences between neurons in different brains (Figure 1A).

One way to compare neural representations without correspondences between neurons is to focus on representational similarity structures, known as Representational Similarity Analysis (RSA).[4,5] RSA circumvents the need for direct neuron-to-neuron correspondences by focusing on similarity structures within the neural representations of different stimuli (Figure 1B). This method considers representational similarity structures—namely, the similarities or dissimilarities between neural responses to different stimuli within each brain. By evaluating the similarity of these structures, RSA allows for the comparison of neural representations across brains, even in the absence of direct neuronal correspondences. RSA extends its utility beyond neural responses, enabling comparison across different modalities, such as behavior or computational models, and thereby offers a versatile framework for understanding the commonality or differences in neural representations between different systems.[6–11] By comparing the similarity structure of neural representations, this approach has suggested the presence of structural commonalities in neural representations among individuals in both humans and animals.[12–16]

Despite its widespread application and usefulness, the conventional RSA framework is not without limitations. The framework typically assumes that there are direct correspondences between the neural representations of the same stimuli in different brains, and compares these neural representations based on the assumed correspondences, which we call "supervised" comparison. The supervised comparison framework includes other methods such as Hyperalignment[1] and encoding models.[17] Although this assumption might be valid or approximately valid in some limited situations, in general, there is no guarantee that the same stimuli are represented in the same relational way between different brains. For example, we previously discussed the possibility that experiences of colors might be relationally different between different individuals.[18] To address such a possibility, we proposed using "unsupervised" alignment to find the optimal correspondences based only on internal relationships, without relying on external correspondences, i.e., stimulus labels.[18] With the unsupervised alignment framework, we can explore the possibility that the same stimuli are not necessarily mapped to each other in different brains. For

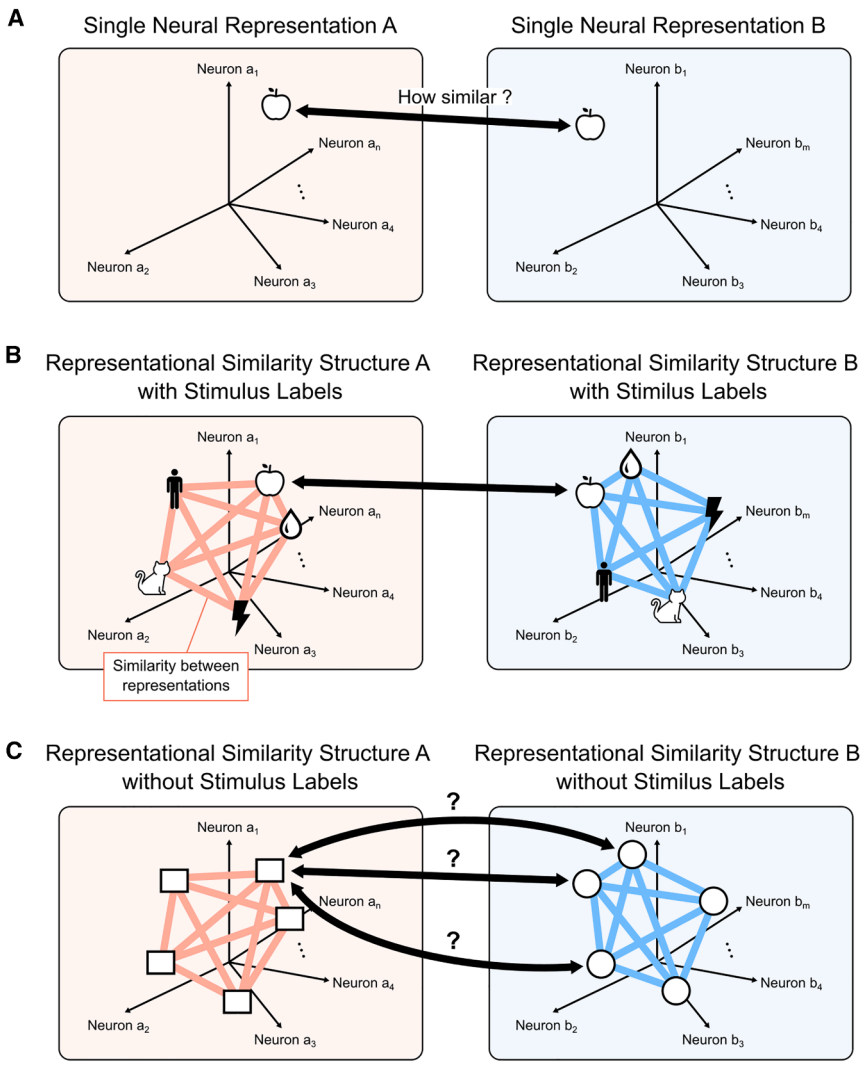


Figure 1. Comparison of neural representations across different individuals

(A) Comparison of single neural representations. Direct comparison of single neural representations across different individuals is challenging due to the lack of correspondence between neurons.

(B) Supervised alignment: By assuming a correspondence among stimulus labels and conducting comparisons of "representational similarity structures," we enable quantitative comparisons of neural representations across different individuals.

(C) Unsupervised alignment: In the situation where neural representations to different stimuli may correspond across individuals, the estimation of correspondence between neural representations must rely solely on each similarity structure, without the use of stimulus labels.

Adapted from [3].

example, it might happen that the neural representation of "red" in brain A does not correspond to "red" in brain B from a relational perspective but instead corresponds to "blue" in brain B. By its nature, the conventional RSA framework sidesteps this possibility by simply assuming the predefined correspondences.

Based on the unsupervised alignment framework, there are several possible scenarios of correspondences between neural representations (Figure 2). In principle, even if there are high correlations between the representational structures of different brains in terms of supervised comparison, the following scenarios are possible[3]: (1) Same one-to-one mapping (Figure 2A): The representations of the same stimuli correspond to each other one-to-one (e.g., the neural representation of an apple in brain A corresponds to that of an apple in brain B). This indicates that the representations are relationally "equivalent" across different brains. (2) Partially different one-to-one mapping (Figure 2B): The representations of some stimuli correspond to each other one-to-one, but not all. For example, the neural representation of an apple in brain A corresponds to the neural representation of a peach in brain B, while other representations, such as a cat in brain A, still correspond to a cat in brain B. This scenario indicates that while there is a generally consistent mapping, certain stimuli have non-identical correspondences between the two brains. (3) Coarse group-to-group mapping (Figure 2C): Neural representations correspond to each other at the group level, but not at the fine-item level. For example, the neural representations of fruits in brain A correspond to those of fruits in brain B, but the neural representation of an apple in brain A does not correspond specifically to an apple in brain B. Instead, it corresponds to any fruit in brain B, indicating a more generalized correspondence at the category level rather than at the individual item level. (4) Consistently shifted one-to-one mapping (Figure 2D): Most representations of different stimuli correspond to each other one-to-one, but are systematically shifted. For example, the neural representation of an apple in brain A consistently corresponds to the neural representation of a peach in brain B, and similarly, a cat in brain A corresponds to a dog in brain B. This scenario suggests a consistent but shifted mapping pattern, where each stimulus has a consistent but different counterpart in the other brain. Shifted mapping can occur when neural representations have a low-dimensional symmetric structure, such as a circle. Supervised comparison, such as RSA, cannot distinguish these possible cases. A key distinction is that unsupervised alignment optimizes the mapping between labels, while the conventional RSA (supervised comparison) uses a fixed mapping (Figure 2E). In unsupervised alignment, we can explore which of the possible mappings best aligns the two structures being compared. In contrast, RSA evaluates the similarity between the structures based on a single, pre-defined mapping, using a correlation metric.

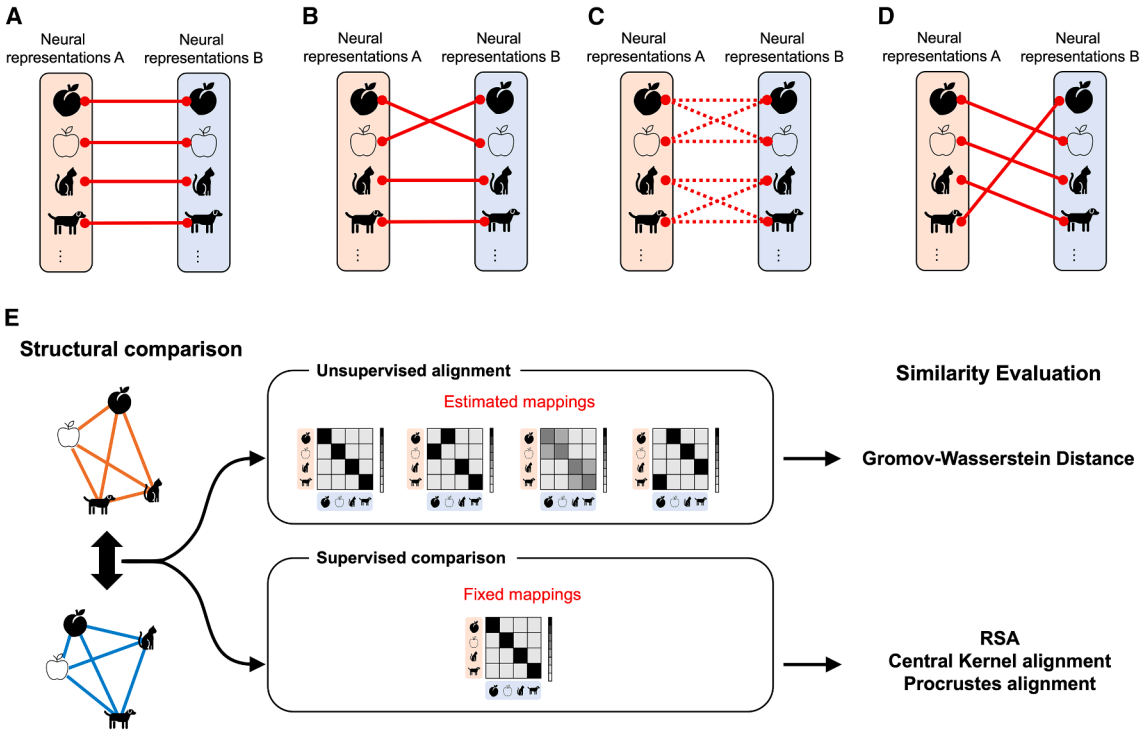


Figure 2. Possible consequences of unsupervised alignment

(A) Same one-to-one mapping.

(B) Partially different one-to-one mapping.

(C) Coarse group-to-group mapping.

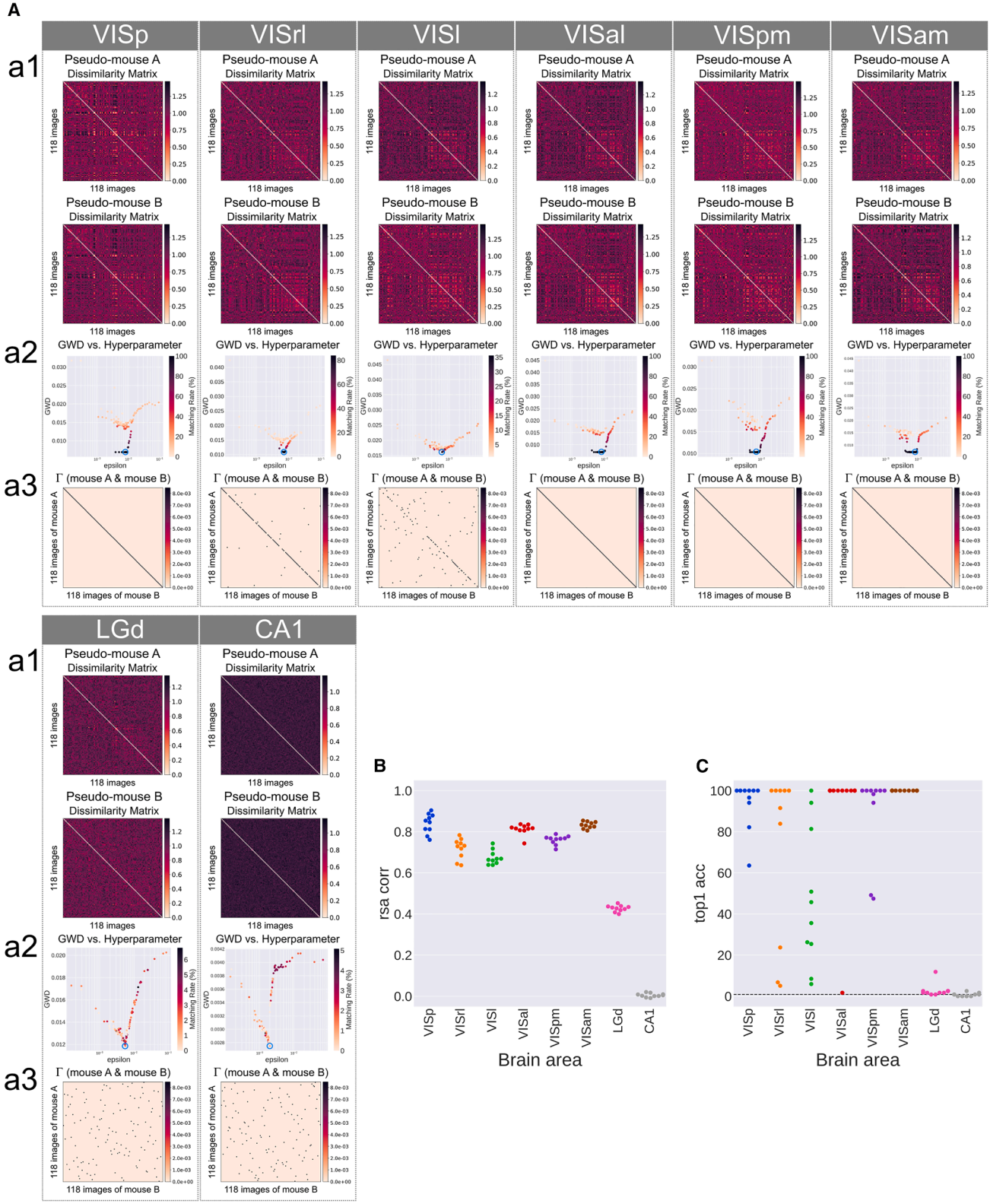
(D) Shifted one-to-one mapping.

(E) Difference between unsupervised alignment and supervised comparison. Unsupervised alignment uses estimated mappings, whereas supervised comparison uses fixed mappings for structural comparison.

Adapted from [3].

To further extend the framework of representational similarity analysis to evaluate optimal correspondences without presupposing specific stimulus correspondences, we have proposed an unsupervised alignment framework for comparing the similarity structures of neural representations. Our primary aim with this framework is to propose a novel similarity metric that diverges from RSA's approach by leveraging unsupervised alignment to analyze the similarity structures of neural representations. It is important to note that, such as RSA, GWOT operates within the same general framework for comparing representational similarity structures, and we do not intend to position GWOT as superior to the conventional RSA based on supervised comparison. This unsupervised alignment framework enables us to identify the optimal correspondence based solely on the internal relationship of neural representations, without presupposing any specific correspondence relationship (Figure 1C). This framework allows us to examine the assumption of correspondence itself. In our previous study, we examined whether the structure of subjective similarity of color is relationally equivalent across individuals, and used this unsupervised alignment technique to address this question.[18] Moreover, by identifying the optimal correspondence among many possible correspondences, the unsupervised alignment method can, in principle, distinguish the above mentioned possible scenarios (Figure 2), which are indistinguishable by supervised comparison such as RSA.[3] For the unsupervised alignment method, we use Gromov-Wasserstein Optimal Transport (GWOT),[19] a method that has been successfully applied in various fields,[20,21] including neuroscience.[18,22,23]

Our research leverages this unsupervised alignment framework based on GWOT to explore the commonality of representational similarity structures of natural scenes across different individuals. Specifically, we investigated this using large-scale open datasets of electrophysiological recordings in mice and fMRI recordings in humans. We obtained mouse electrophysiological data from the Allen Brain Observatory,[24] collected using Neuropixels probes, and human fMRI data from the Natural Scenes Dataset (NSD).[25] These datasets are suitable for our research purposes because they contain a large number of natural scene stimuli, allowing us to investigate the rich and complex similarity structures of neural representations of natural scenes. Furthermore, these datasets provide high-quality neural activity measurements across many individuals, enabling statistically reliable analysis. Using these ideal resources, we first investigated whether the similarity structures of natural scenes from the same brain regions can be aligned in an unsupervised



(legend on next page)

manner across different individuals for both mice and humans. In addition, we also investigated whether the similarity of representations can be aligned across different brain regions. This approach opens new avenues for revealing more nuanced structural similarities or differences in neural representations.

RESULTS

In the following analysis, we performed an unsupervised alignment of neural representations. See STAR Methods for details on the method.

Unsupervised alignment of mouse neural representations

We first performed unsupervised alignment of neural representations in mice using the Neuropixels dataset[24] to investigate whether neural representations of natural scenes can be aligned across different mice (see STAR Methods for the data description). We analyzed the neural activity in 6 areas of the visual cortex (VISp, VISrl, VISl, VISal, VISpm, and VISam), 1 area of the thalamus (LGd), and 1 area of the hippocampus (CA1). In the following, we first performed an unsupervised alignment between the same areas of different individual mice to investigate whether the representational similarity structures of these anatomically identical areas could be aligned across different mice. Then, we performed an unsupervised alignment between the different areas to investigate commonalities among the similarity structures between the different areas.

In this study, we performed an unsupervised alignment between two pseudo-mice (see STAR Methods) to obtain a statistically reliable estimate of representational similarity structures. We repeated the analysis 10 times, each time changing the division of the mice to construct different pairs of pseudo-mice. This approach was employed to examine the influence of mouse selection on the group-averaged representational similarity structures.

Representational similarity structures in each brain area

To perform unsupervised alignment, we estimated the representational similarity structures of 118 natural scene stimuli for each brain area (see STAR Methods). In Figure 3A1, we show a specific example of the dissimilarity matrices of a pair of pseudo-mice for the 118 stimuli in each area. We quantified the dissimilarity by the cosine distance between the trial-averaged normalized spike counts of two stimuli (see STAR Methods). As can be seen in Figure 3A1, the dissimilarity matrices of two pseudo-mice are highly similar for the visual cortical areas (VISp, VISrl, VISl, VISal, VISpm, VISam). In these areas, the correlations between the dissimilarity matrices are about 0.6–0.9 (Figure 3B). Compared to the visual cortical areas, the correlations in the thalamus (LGd) are lower, about 0.4 (Figure 3B). In contrast to these areas in the visual system, the correlations in the hippocampus (CA1) are close to 0, indicating that there are no common structures in CA1.

Unsupervised alignment between the same brain areas

We investigated whether there are sufficient structural similarities to allow unsupervised alignment of the neural representations within the same brain area across animals. To this end, we performed unsupervised alignment of the dissimilarity matrices in the same area across different pseudo-mice (Figure 3A). For each brain area, we performed GWOT with 100 trials on different ϵ values. The points in Figure 3A2 correspond to the estimated GWD (Equation 1) in each optimization trial. We selected the result of the trial with the lowest estimated GWD as the optimal solution (shown in the blue circle in Figure 3A2).

The optimal transportation plan Γ^* obtained through GWOT exhibited distinct characteristics between the visual cortical areas and other areas. Examples of the optimal transportation plan Γ^* for specific pairs of pseudo-mice are shown in Figure 3A3. In most areas of the visual cortex, the optimal transportation plans were close to diagonal matrices, i.e., most of the diagonal elements tended to have higher values than the off-diagonal elements. This means that the same stimuli are matched with each other across the different pseudo-mice. In contrast, in LGd, the optimal transportation plan was not diagonal but rather randomly mapped different stimuli over the pseudo-mice. The failure of unsupervised alignment in LGd was not fully expected because LGd exhibited a moderate degree of structural similarity, as indicated in Figure 3B, though the correlations are lower than the visual cortical areas. In comparison, in CA1, the optimal transportation plan also showed random matching, but this result was expected because there were no common structures observed at the level of correlation (Figure 3B). Taken together, these results suggest that the representational similarity structures within the same visual cortical area are highly consistent across animals, while they are notably less consistent in the thalamus, and there is no common structure at all in CA1.

We repeated the unsupervised alignment analysis 10 times, changing the division of the mice to construct different pairs of pseudo-mice each time. The average top 1 matching rates indicated that visual cortical areas generally share unsupervisedly alignable representational structures, while other areas do not (Figure 3C). The visual cortical areas showed the following average top 1 matching rates across 10 trials: 93.6% for VISp, 71.1% for VISrl, 47.4% for VISl, 90.2% for VISal, 88.9% for VISpm, and 100.0% for VISam. All of these significantly exceeded the chance level of 0.85%. On the other hand, subcortical areas showed average top 1 matching rates of 2.8% for

Figure 3. Unsupervised alignment between the same areas in different pseudo-mice

(A) The results of GWOT between the dissimilarity matrices of the same areas in different pseudo-mice. (A1) Dissimilarity matrices of a pair of pseudo-mice for the 118 stimuli in each area. (A2) Relationship between GWD (objective of GWOT) and the hyperparameter ϵ. (A3) Optimal transportation plan Γ^* between the dissimilarity matrices of a pair of pseudo-mice. This matrix corresponds to the point encircled in blue in (A2).

(B) The correlation coefficient of RSA between the dissimilarity matrices of a pair of pseudo-mice across 10 trials.

(C) The top 1 matching rate of the unsupervised alignment between the dissimilarity matrices of a pair of pseudo-mice across 10 trials.

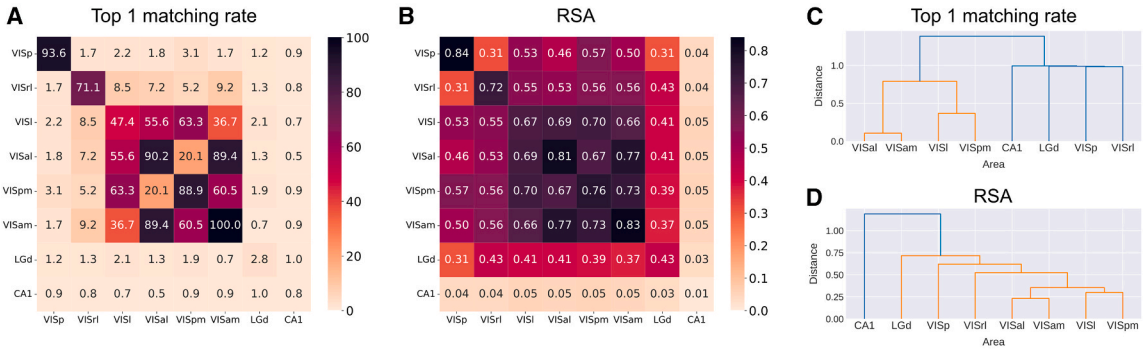


Figure 4. Unsupervised alignment between the different areas in different pseudo-mice

(A) The average top 1 matching rate of the unsupervised alignment for each pair of brain areas.

(B) The average correlation coefficient of the RSA for each pair of brain areas.

(C) Hierarchical clustering of brain areas based on the average top 1 matching rate. Here, the distance between areas is defined as $(100 - \text{top 1 matching rate})/100$ and Ward's method is employed as the clustering criterion.

(D) Hierarchical clustering of brain areas based on the average correlation coefficient. The distance between areas is defined as $(1 - \text{correlation})$.

LGd and 0.8% for CA1, which were nearly equivalent to the chance level.

Finally, we compared the results obtained by the unsupervised alignment based on GWOT with traditional supervised alignment (RSA) to assess the consistency or difference between the results. Upon calculating the average correlations between dissimilarity matrices using the conventional RSA framework, we observed consistent characteristics overall between GWOT (unsupervised, Figure 3C) and RSA (supervised, Figure 3B), but also subtle differences. For instance, while the dissimilarity matrices of LGd exhibited moderate correlations of 0.4, their top 1 matching rates were almost 0%. Additionally, the matching rates for VISl showed high variability, while the correlation values were much less variable and concentrated around 0.7. The important observation here is that high correlations do not necessarily mean high matching rates in unsupervised alignment.

We also analyzed the data for natural movie stimuli (natural movie 1 and natural movie 3) and obtained qualitatively similar results (see Figures S1 and S2).

Unsupervised alignment between the different brain areas

To investigate the extent to which representational similarity structures can be aligned across different brain areas, we also performed an unsupervised alignment of dissimilarity matrices between different areas of two pseudo-mice. Similarly to the same area experiment, we varied the random grouping of the animals 10 times and calculated the average top 1 match rate. It is worth mentioning that a single grouping of mice can yield two pairs of different areas, such as VISp of pseudo-mouse A and VISrl of pseudo-mouse B, as well as VISrl of pseudo-mouse A and VISp of pseudo-mouse B. Therefore, the resulting value represents the average of the top 1 match rates for 20 pairs.

Analysis of the average top 1 matching rate provided insights into the commonality of representational similarity structures across different areas (Figure 4A), with the primary findings being that: 1) the neural representations of VISp and VISrl exhibit a unique similarity structure distinct from those of other areas, and 2) the representational similarity structures of the higher-order visual cortical areas are similar to each other. First, the average top 1 matching rates between VISp and other areas were approximately 0–3%, which is remarkably low compared to the value of 93.6% observed within VISp itself, and close to the chance level of 0.85%. This suggests that the neural representations of VISp possess a unique similarity structure, differing not only from those in the thalamus and hippocampus but also from those in other visual cortical areas. A similar tendency was observed for VISrl, although this was not as pronounced as for VISp. Second, within VISl, VISal, VISpm, and VISam, there are pairs of areas that exhibit high top 1 matching rates exceeding 50%. This result reveals that there can be commonalities in representational similarity structures in the higher-order visual cortical areas, even among distinct areas. Moreover, similar insights were inferred from the results of the conventional RSA framework (Figure 4B), although the relationships between areas were observed more clearly through unsupervised alignment.

To further interpret the relationships between different areas as quantified by unsupervised alignment, we conducted the hierarchical clustering of the areas based on the average top 1 matching rate (Figure 4C). This hierarchical structure closely matched our earlier findings. Notably, VISp and VISrl formed independent clusters, distinct from other visual cortical areas. Additionally, VISl, VISal, VISpm, and VISam formed clusters, which is also consistent with our earlier findings.

Finally, we compared the hierarchical clustering results obtained by unsupervised alignment based on GWOT with those obtained by supervised alignment (RSA) to assess the consistency or difference between the results. Using average correlation coefficients from the conventional RSA framework for hierarchical clustering, we observed a different cluster structure compared to the top 1 matching rate result (Figure 4D). In the higher-order visual cortex, the formation of 2 sub-clusters obtained from the RSA was consistent with the results observed in the top 1 matching rate of unsupervised alignment. However, in contrast to the results of unsupervised alignment, where other areas formed independent single clusters, we observed a

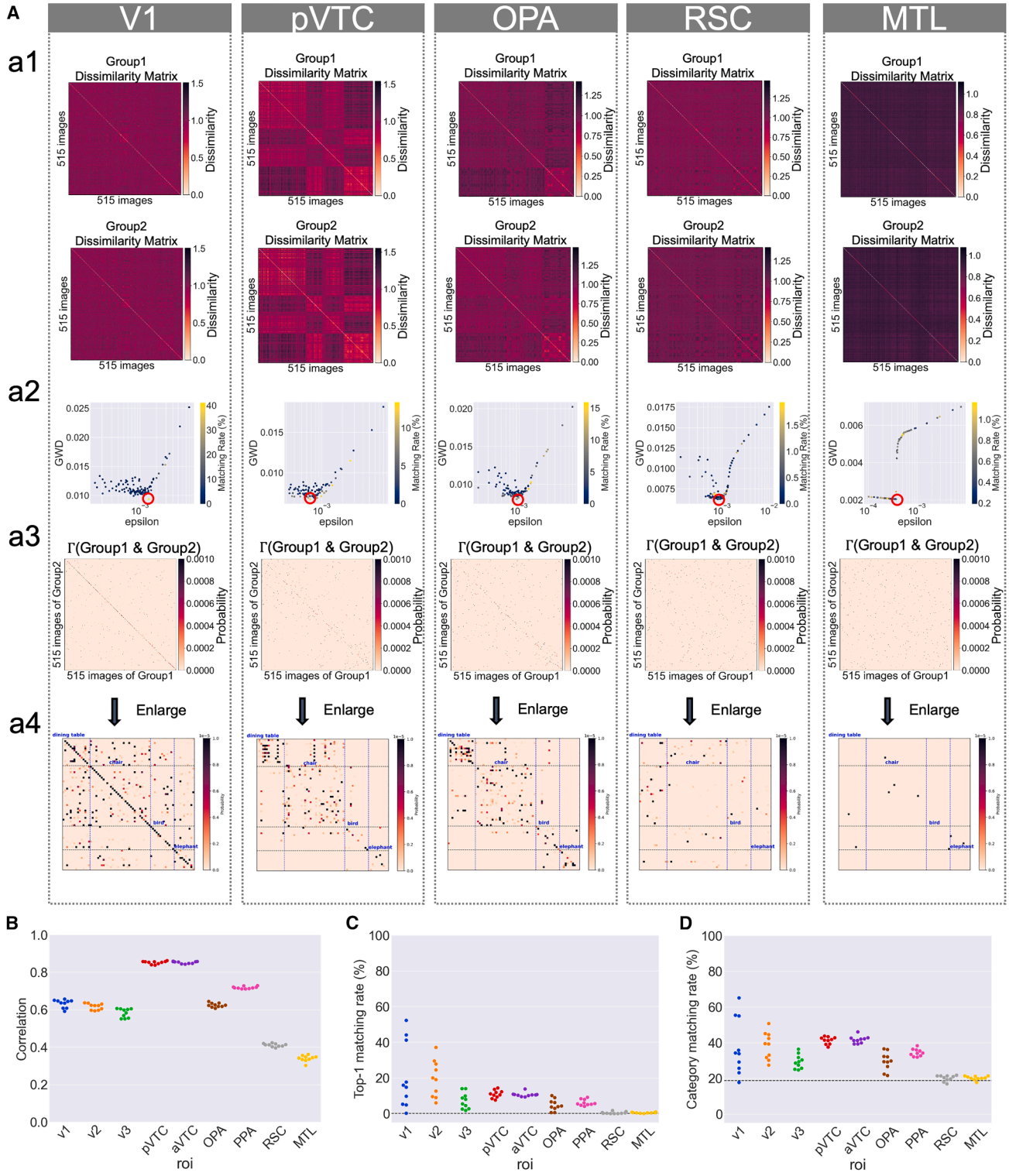


Figure 5. Unsupervised alignment between the same brain areas in different participants groups

(A) Results of GWOT between dissimilarity matrices of the same brain areas in different participants groups. (A1) Dissimilarity matrices of 515 stimuli of Group 1 and Group 2. (A2) Relationship between GWD and matching rate. Color represents top 1 matching rates. (A3) Optimal transportation plan Γ^* between the

(legend continued on next page)

gradual increase in distance from the higher-order visual cortex in the order of VISrl, VISp, and LGd. This order strongly reflects the functional hierarchy of the mouse visual system.[24] Thus, this comparison suggests that the inter-regional relationships based on unsupervised alignment via GWOT are not simply predictable from RSA correlations or hierarchical proximity.

We also analyzed the data for natural movie stimuli (natural movie 1 and natural movie 3) and obtained qualitatively similar results (see Figures S3 and S4).

Unsupervised alignment of human neural representations

Next, we investigated whether the neural representations of natural scenes in the human brain could be aligned across participants using the Neural Scenes Dataset (NSD)[25] (see STAR Methods for the data description). As in the previous section, we first conducted comparisons within the same areas to investigate whether similar structures in the same brain areas can be aligned across individuals. Next, we performed cross-area comparisons to investigate the extent to which representational similarity structures can be aligned across different brain areas.

As we did in the analysis of mouse neural representations, we also performed an unsupervised alignment between two participant groups (see STAR Methods). We repeated the analysis 10 times, changing the division of the participants to construct different pairs of groups each time. This approach was employed to examine the influence of participant selection on the group-averaged representational similarity structures.

Representational similarity structures in each brain area

First, we estimated the representational similarity structures of visual stimuli from 515 natural scenes for several visual areas (V1, V2, V3, pVTC, aVTC, OPA, PPA, and RSC) and a non-visual area (MTL) in each participant's group. In Figure 5A1, we show a specific example of pairs of dissimilarity matrices of the 515 stimuli in each area in two participant groups. See Figure S5 for examples from the other areas. We computed the mean dissimilarity matrix of the 515 stimuli for each area in each participant group, where the dissimilarity between stimuli is quantified by the correlation distance $(1-\rho)$ between the vectors of neural responses to the stimuli following convention (see STAR Methods for details). As seen in Figure 5A1, the appearance of the similarity structure in each region appears similar between the two groups. To quantitatively demonstrate the degree of similarity, we calculated the correlations between the two matrices. Figure 5B shows the correlations for ten randomly divided group pairs, showing relatively high correlations in all regions.

Unsupervised alignment between the same brain areas

We next investigated whether the neural representations of the same brain regions could be aligned between participant groups. To this end, we performed unsupervised alignment between the mean dissimilarity matrices of the same brain regions in different participant groups. We show the results of one of the 10 groupings in Figure 5A, and all of the results of the 10 different samples in Figures 5B–5D. We first performed GWOT between the two dissimilarity matrices for each brain region (Figure 5A1). We performed a total of 100 optimization iterations on different ϵ values. The points in Figure 5A2 correspond to the estimated GWD (Equation 1) in each optimization trial. We selected the result of the trial with the lowest estimated GWD as the optimal solution (shown in the blue circle in Figure 5A2). After the optimization, we obtained the optimal transportation plan Γ^* between the two dissimilarity matrices (Figure 5A3).

The optimal transportation plan Γ^* obtained through GWOT, indicates the structural commonality between the representational similarity structures of two different participant groups. As shown in Figure 5A3, in many cases, the diagonal elements in Γ^* have high values, indicating that many of the stimuli in one group correspond to the same stimuli in the other group with high probability. Figure 5A4 is an enlarged view of the optimal transportation plan Γ^* shown in Figure 5A3. As particularly indicated by pVTC and OPA in Figure 5A4, even in the case of mismatches, matching errors tended to occur within the same coarse category, such as dining table, chair, bird, and elephant. Mismatches also tended to occur within semantically similar categories, such as between a dining table and a chair. In contrast, in RSC, the optimal transportation plan was not diagonal but rather randomly mapped different stimuli over the participant groups. The failure of unsupervised alignment in RSC was not fully expected because RSC exhibited a moderate degree of structural similarity, as indicated in Figure 5B, though the correlations are lower than the other visual areas. In comparison, in MTL, the optimal transportation plan also showed random matching but this result was expected because there were no common structures observed at the level of correlation (Figure 5B).

By performing the same analysis with 10 different random groupings of the participants, we obtained the top 1 matching rate of the 10 random pairs (Figure 5C). The average of the top 1 matching rate over 10 random groupings was 21% for V1, 20% for V2, 7.2% for V3, 11% for pVTC, 11% for aVTC, 4.8% for OPA, 6.2% for PPA, 0.5% for RSC, and 0.5% for MTL. We also calculated the top 1 category matching rate of the 10 random samples (Figure 5D). The average of the top 1 category matching rate over 10 random samples was 37% for V1, 38% for V2, 29% for V3, 41% for pVTC, 42% for aVTC, 30% for OPA, 34% for PPA, 20% for RSC, and 20% for MTL. These values are significantly higher than the chance level (0.2% for the top 1 matching rate and 18.7% for the top 1 category matching rate) with the exception of RSC and MTL. Taken together, these results suggest that the representational similarity structures within the same visual cortical area are highly consistent across participants, while there is no common structure at all in RSC and MTL.

dissimilarity matrices of Group 1 and Group 2. (A4) Enlarged view of the some coarse categories of optimal transportation plan Γ^*. Blue dotted lines indicate the boundaries of the coarse categories.

(B) The RSA correlation coefficient between the dissimilarity matrices of the same brain areas in different participant groups.

(C) Top 1 matching rate of the unsupervised alignment for 10 random pairs of participant groups.

(D) Top 1 category matching rate of the unsupervised alignment for 10 random pairs of participant groups.

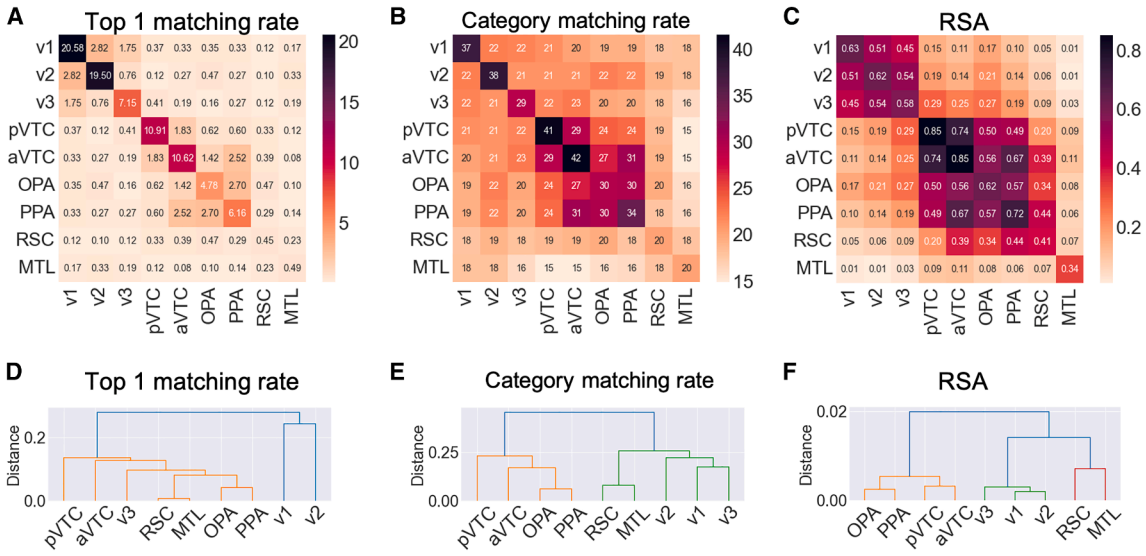


Figure 6. Unsupervised alignment between different brain areas

(A) Average top 1 matching rate of unsupervised alignment for each pair of brain areas.

(B) Average category matching rate of unsupervised alignment for each pair of brain areas.

(C) Average RSA correlation coefficient between the dissimilarity matrices of different brain areas.

(D) Hierarchical clustering of brain areas based on the average top 1 matching rate. Here, the distance between areas is defined as $(100 - \text{top 1 matching rate})/100$ and Ward's method is employed as the clustering criterion.

(E) Hierarchical clustering of brain areas based on the average category matching rate. The distance between areas is defined as $(100 - \text{category matching rate})/100$.

(F) Hierarchical clustering of brain areas based on the average correlation coefficient. Distance between areas is defined as $(1 - \text{correlation})$.

Unsupervised alignment between different brain areas

Next, we investigated whether the similarity structures of neural representations of natural scenes in different brain regions could be aligned across participants. To this end, we performed unsupervised alignment between the mean dissimilarity matrices of different brain regions in different participant groups.

For each case, we performed GWOT between the two dissimilarity matrices. We then calculated the matching rate of the unsupervised alignment and averaged the matching rate over 10 random groupings of participants. In Figure 6A, we show the average top 1 matching rate of unsupervised alignment between the different brain regions in different participant groups. The matching rates between the different brain areas were slightly higher than the chance level, but overall, much lower than the matching rate between the same brain regions. At this level of matching rate, we could not observe a clear relationship between the different regions.

In contrast, the category-level matching of the unsupervised alignment revealed a slightly clearer relationship between regions. First, within pVTC, aVTC, OPA, and PPA, there are pairs of areas that exhibit slightly higher matching rates, exceeding the chance level (18.7%). In contrast, RSC shows a unique representational structure that sets it apart from these areas. This result reveals that there can be commonalities in representational similarity structures in the higher-order visual areas, even among distinct areas.

A similar but much clearer tendency was observed in the framework of RSA. In Figure 6C, we show the RSA correlation coefficient between the dissimilarity matrices of different brain regions. The correlation coefficients showed high values in comparisons within V1, V2, V3, and within pVTC, aVTC, OPA, and PPA, whereas they were low in comparisons across these divisions. This trend was consistent with what we observed in unsupervised alignment, but was clearer and more discernible. On the other hand, RSC exhibited relatively high values in these higher-order visual areas, hinting at some degree of commonality in their representations. This tendency deviated from the patterns observed in unsupervised alignment, highlighting differences between the two methods.

To visually elucidate the relationships between the different brain regions, we applied hierarchical clustering to both the matching rate and the RSA correlation (Figures 6D–6F). Specifically, clustering based on the category matching rate and RSA revealed that higher-order areas such as pVTC, aVTC, OPA, and PPA formed a distinct cluster, separate from early visual areas such as V1, V2, and V3.

DISCUSSION

In this study, we propose a novel framework to compare neural representational similarity structures without using stimulus identities, but purely based on the internal relations of the representational structures, using an unsupervised alignment method based on GWOT. A fundamental difference between this framework and the conventional supervised framework is that this framework attempts to find the optimal correspondences

between neural representations of stimuli and to assess whether neural representations of the same stimuli are actually mapped across different individuals, challenging the implicit assumption behind the conventional supervised comparison. We consider that if two similarity structures are aligned at the one-to-one item level with the unsupervised alignment, this provides evidence for a stronger and finer level of structural correspondence beyond simple correlation.

In our analysis of the mouse Neuropixels dataset, we found that the similarity structures of neural representations could be well aligned in the same visual cortical areas (VISp, VISrl, VISl, VISal, VISpm, and VISam) but not in the thalamus (LGd), suggesting regional differences in the degree of commonality of representational structures across individuals. Given that LGd is located at the bottom of the mouse visual system hierarchy after the retina, and that LGd has strong feedforward projections into VISp,[24,26] it is not clear why LGd lacks alignable structures, whereas VISp (or higher-order visual areas) have consistently alignable structures across animals. Future work combining experimental and computational neuroscience would be needed to elucidate what kind of representational transformation from the thalamus to the visual cortex makes representations more consistent.

In addition, using unsupervised alignment, we obtained inter-regional structural relationships in the mouse visual system that are consistent with conventional knowledge, but still show some distinctive features that are not fully expected by conventional analysis. In particular, we found that neural representations in VISp and VISrl could not be aligned with other visual cortical areas closely related in the hierarchy, suggesting that these areas have idiosyncratic structures. In contrast, neural representations were well aligned within higher-order visual cortical areas (VISl, VISal, VISpm, VISam), and the dorsal-like regions VISl and VISpm or the ventral-like regions VISal and VISam[27] are more aligned with each other, as shown in the hierarchical clustering analysis (Figure 4C). Although these are reasonable results given the known functional similarities of these brain areas, we consider it still meaningful to show that these higher-order visual areas share enough common structures to allow unsupervised alignment.

In analyzing the human fMRI data, we also found that neural representations of the same brain regions can be unsupervisedly aligned across individuals to some extent, although the matching rate is lower than in the mouse data. This is probably due to the difference between the types of recordings, i.e., invasive recordings from Neuropixels and non-invasive recordings from fMRI, and the number of subjects used for averaging the similarity matrices, 16 for the mouse data and 4 for the fMRI data. Given that fMRI typically contains a lot of noise in the data, it may be significant that unsupervised alignment was successful even with an average of only 4 participants, indicating the high quality of this dataset and the strong commonalities between individuals, with the exception of RSC and MTL. The lower alignment in these regions may stem from their roles in integrating diverse types of information and their sensitivity to cognitive and contextual factors, which could introduce greater individual variability in representational structures.[28,29] Although less clear than in the case of the mouse data, we were also able to assess the proximity of some regions using unsupervised alignment between different regions. In particular, in the higher-order areas (pVTC, aVTC, OPA, and PPA), it was possible to assess the proximity of similarity structures by considering category-level matching. Future research is expected to expand the scope of analysis of the brain regions and comprehensively investigate their relationships.

In our analysis, we observed distinct differences between GWOT and RSA in how representational similarity structures are evaluated. One key distinction is that GWOT allows for the estimation of different types of mappings depending on the relationship between the brain areas being compared, whereas RSA relies on a fixed, predefined mapping. For instance, in the GWOT between VISam and VISal, mostly the same one-to-one mapping (Figure 2A) was estimated. In this case, the evaluations from both RSA and GWOT were consistent, with the RSA correlation being 0.77 and the GWOT matching rate being 89.4%, both indicating high similarity. However, GWOT further reveals that other possible mappings, such as incorrect one-to-one mappings or the group-to-group mapping (Figure 2C), were not optimal, suggesting that the correct one-to-one mapping was the most suitable and reflecting a strong level of similarity between these two brain areas. On the other hand, in the case of VISp and VISl, for example, nearly random mappings were found to be optimal, resulting in a very low matching rate of 2.2%. This result was not easily predicted from the moderate degree of similarity indicated by the RSA correlation (0.53). In this case, GWOT's optimization process revealed that the representational similarity structures between the two brain areas were, in fact, so dissimilar that unsupervised alignment is impossible, an insight that would not have been apparent through RSA alone.

The matching rate computed from the optimal transport plan obtained by GWOT tends to exhibit greater discontinuity compared to RSA. This is because, at the chance level, the matching rate is inherently the same by definition, whereas RSA values can vary even when all matching rates are at the chance level. One potential way to reduce this discontinuity is to use a top-k matching rate ($k > 1$) instead of the top-1 matching rate. However, if the transport matrix is sparse, the top-k matching rate remains nearly identical to the top-1 matching rate. In contrast, as the transport matrix becomes denser, the top-k matching rate becomes smoother because non-zero values are more likely to appear in diagonal elements. To utilize the top-k matching rate ($k > 1$) as a more continuous measure, it is crucial to control the sparsity of the transport matrix. The entropy regularization parameter (ϵ) plays a key role in this, as higher ϵ values lead to denser transport matrices. Therefore, deliberately using a higher ϵ could make the top-k matching rate of GWOT more continuous. Currently, ϵ is treated as an optimization parameter, and we evaluate the optimal transport solution for a single optimized ϵ value. Future work could explore refining the adjustment of ϵ to achieve a smoother transition between top-k and top-1 matching rates, further improving the continuity of this measure.

One important avenue for future research is the comparison of representations across different modalities. For instance, comparing neural representations with behavior is crucial for identifying the neural correlates of a given behavior or the neural

basis of subjective experiences. Understanding these connections can provide key insights into how neural activity underpins cognitive and behavioral processes.[4,30–33] Similarly, comparing brain representations with computational models is critical for investigating which learning methods and architectures lead to representations that resemble those found in humans or animals.[34–39] This comparison can help us understand how artificial systems can be designed to mimic human-like representational structures. GWOT can play an important role in investigating such comparisons. In particular, in our previous work,[22,40] we have used GWOT to compare behavioral representations in humans and computational models. Expanding such comparisons to include neural representations across modalities is a key direction for future research.

Limitations of the study

Although this study aims at alignment across different individuals, our analysis is limited to the group level. That is, we assess commonalities in neural representations by averaging similarity matrices across individuals, rather than directly aligning representations at the individual level. In both datasets we analyzed in this study, the results of unsupervised alignment at the individual level were statistically unreliable. Thus, based on this study alone, it is still unknown whether neural representations are common enough to be aligned in the unsupervised manner even at the individual level. Although the current datasets provide reliable answers, as datasets continue to increase in size, e.g., in terms of recorded neurons or recording durations, and include higher resolution and more information-rich recordings, it may become possible to extend our approach to individual-level alignment by using such datasets in the future. Achieving this could lead to a more precise understanding of inter-individual variability in neural representations and provide deeper insights into how representational structures differ across subjects.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Masafumi Oizumi (c-oizumi@g.ecc.u-tokyo.ac.jp).

Materials availability

This study did not generate new materials.

Data and code availability

- This article analyzes publicly available data from the Allen Brain Observatory (https://allensdk.readthedocs.io/en/latest/visual_coding_neuropixels.html) and the Natural Scenes Dataset (https://naturalscenesdataset.org/).
- Code used for the analyses in this article is publicly available at the following repository at GitHub: git@github.com:oizumi-lab/NeuRep_GWOT.git.
- Any additional information required to reanalyze the data reported in this article is available from the lead contact upon request.

ACKNOWLEDGMENTS

MO was supported by JST Moonshot R&D Grant Number JPMJMS2012 and Japan Promotion Science, Grant-in-Aid for Transformative Research Areas Grant Number 23H04834.

AUTHOR CONTRIBUTIONS

K.A. and M.O. conceived the initial analysis idea. K.A. conducted the analysis of the mouse data, and K.T. analyzed the human data. K.T., K.A., and M.O. wrote the initial draft of the article. J.K. extensively reviewed both the analysis and the draft. All authors reviewed, edited, and approved the final article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Data
  - Mouse neuropixels dataset: Allen Brain Observatory Visual Coding
  - Human fMRI dataset: Neural Scenes Dataset
  - Representational similarity structures
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Comparison of neural representations

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112427.

Received: September 2, 2024
Revised: February 10, 2025
Accepted: April 10, 2025
Published: April 15, 2025

REFERENCES

1. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science *293*, 2425–2430.
2. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. Science *303*, 1634–1640.
3. Takeda, K., Sasaki, M., Abe, K., and Oizumi, M. (2025). Unsupervised alignment in neuroscience: Introducing a toolbox for Gromov–Wasserstein optimal transport. J. Neurosci. Methods. *419*, 110443.
4. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.
5. Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. *17*, 401–412.
6. Cichy, R.M., Kriegeskorte, N., Jozwik, K.M., van den Bosch, J.J.F., and Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. Neuroimage *194*, 12–24.
7. Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. *10*, e1003915.
8. Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B.C., Grant, E., Groen, I., Achterberg, J., et al. (2023). Getting aligned on representational alignment. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.13018.
9. Mahner, F., Muttenthaler, L., Gucclu, U., and Hebart, M. (2024). Dimensions underlying the representational alignment of deep neural networks

with humans. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.19087.

10. Marjieh, R., Van Rijn, P., Sucholutsky, I., Sumers, T., Lee, H., Griffiths, T.L., and Jacoby, N. (2022). Words are all you need? language as an approximation for human similarity judgments. Preprint at arXiv. https://doi.org/10.48550/arXiv.2206.04105.
11. Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., and Griffiths, T.L. (2024). Large language models predict human sensory judgments across six modalities. Sci. Rep. *14*, 21445.
12. Deitch, D., Rubin, A., and Ziv, Y. (2021). Representational drift in the mouse visual cortex. Curr. Biol. *31*, 4327–4339.e6.
13. Charest, I., Kievit, R.A., Schmitz, T.W., Deca, D., and Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. Proc. Natl. Acad. Sci. USA *111*, 14565–14570.
14. Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron *60*, 1126–1141.
15. Raizada, R.D.S., and Connolly, A.C. (2012). What makes different people's representations alike: neural similarity space solves the problem of across-subject fMRI decoding. J. Cogn. Neurosci. *24*, 868–877.
16. Shinkareva, S.V., Malave, V.L., Just, M.A., and Mitchell, T.M. (2012). Exploring commonalities across participants in the neural representation of objects. Hum. Brain Mapp. *33*, 1375–1383.
17. Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. Neuroimage *56*, 400–410.
18. Kawakita, G., Zeleznikow-Johnston, A., Takeda, K., Tsuchiya, N., and Oizumi, M. (2025). Is my "red" your "red"?: Evaluating structural correspondences between color similarity judgments using unsupervised alignment. iScience *28*, 112029.
19. Mémoli, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching, Found. Comut. Math *11*, 417–487.
20. Alvarez-Melis, D., and Jaakkola, T. (2018). Gromov-Wasserstein alignment of word embedding spaces. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 1881–1890.
21. Demetci, P., Santorella, R., Sandstede, B., Noble, W.S., and Singh, R. (2022). SCOT: Single-Cell Multi-Omics alignment with optimal transport. J. Comput. Biol. *29*, 3–18.
22. Kawakita, G., Zeleznikow-Johnston, A., Tsuchiya, N., and Oizumi, M. (2024). Gromov-wasserstein unsupervised alignment reveals structural correspondences between the color similarity structures of humans and large language models. Sci. Rep. *14*, 15917.
23. Thual, A., Tran, Q.H., Zemskova, T., Courty, N., Flamary, R., Dehaene, S., and Thirion, B. (2022). Aligning individual brains with fused unbalanced gromov wasserstein. In Advances in Neural Information Processing Systems, *35*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds. (Curran Associates, Inc.), pp. 21792–21804.
24. Siegle, J.H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T.K., Choi, H., Luviano, J.A., et al. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. Nature *592*, 86–92.
25. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nat. Neurosci. *25*, 116–126.
26. Harris, J.A., Mihalas, S., Hirokawa, K.E., Whitesell, J.D., Choi, H., Bernard, A., Bohn, P., Caldejon, S., Casal, L., Cho, A., et al. (2019). Hierarchical organization of cortical and thalamic connectivity. Nature *575*, 195–202.
27. Bakhtiari, S., Mineault, P., Lillicrap, T., and Pack, C.C. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. Adv. Neural Info. Process. Syst. *34*, 25164–25178.
28. Alexander, A.S., Place, R., Starrett, M.J., Chrastil, E.R., and Nitz, D.A. (2023). Rethinking retrosplenial cortex: Perspectives and predictions. Neuron *111*, 150–175.
29. Squire, L.R., Wixted, J.T., and Clark, R.E. (2007). Recognition memory and the medial temporal lobe: a new perspective. Nat. Rev. Neurosci. *8*, 872–883.
30. Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P.A., and Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. Front. Psychol. *4*, 128.
31. Chikazoe, J., Lee, D.H., Kriegeskorte, N., and Anderson, A.K. (2014). Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci. *17*, 1114–1122.
32. Horikawa, T., Cowen, A.S., Keltner, D., and Kamitani, Y. (2020). The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. iScience *23*, 101060.
33. Koide-Majima, N., Nakai, T., and Nishimoto, S. (2020). Distinct dimensions of emotion in the human brain and their representation on the cortical surface. Neuroimage *222*, 117258.
34. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619–8624.
35. Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron *98*, 630–644.e16.
36. Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M.C., DiCarlo, J.J., and Yamins, D.L.K. (2021). Unsupervised neural network models of the ventral visual stream. Proc. Natl. Acad. Sci. USA *118*, e2014196118.
37. Rajalingham, R., Issa, E.B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J.J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. J. Neurosci. *38*, 7255–7269.
38. Konkle, T., and Alvarez, G.A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. Nat. Commun. *13*, 491.
39. Conwell, C., Prince, J.S., Kay, K.N., Alvarez, G.A., and Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. Nat. Commun. *15*, 9383.
40. Takahashi, S., Sasaki, M., Takeda, K., and Oizumi, M. (2024). Self-supervised learning facilitates neural representation structures that can be unsupervisedly aligned to human behaviors. In ICLR 2024 Workshop on Representational Alignment.
41. Alaux, J., Grave, E., Cuturi, M., and Joulin, A. (2018). Unsupervised hyper-alignment for multilingual word embeddings. Preprint at arXiv. https://doi.org/10.48550/arXiv.1811.01124.
42. Peyré, G., Cuturi, M., and Solomon, J. (2016). Gromov-Wasserstein averaging of kernel and distance matrices. In Proceedings of The 33rd International Conference on Machine Learning, *48*, pp. 2664–2672.
43. Peyré, G., and Cuturi, M. (2019). Computational optimal transport: With applications to data science, Found. Trends Mach. Learn. *11*, 355–607.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Mouse electrophysiological data	The Allen Brain Observatory (Siegle et al.[24])	https://allensdk.readthedocs.io/en/latest/visual_coding_neuropixels.html
Human fMRI data	The Natural Scenes Dataset (Allen et al.[25])	https://naturalscenesdataset.org/
Software and algorithms		
Data analysis code	This paper	GitHub: https://github.com/oizumi-lab/NeuRep_GWOT

METHOD DETAILS

Data

To investigate whether the neural representations of natural scenes in different individuals can be aligned, we used two datasets: the mouse Neuropixels dataset from Allen Brain Observatory Visual Coding[24] and the human fMRI dataset from the Neural Scenes Dataset (NSD).[25] All data are available from http://brain-map.org/explore/circuits for the mice Neuropixels and from http://naturalscenesdataset.org for the human fMRI.

Mouse neuropixels dataset: Allen Brain Observatory Visual Coding

We utilized the Allen Brain Observatory Visual Coding - Neuropixels dataset.[24] This dataset contains extracellular electrophysiology recordings of neural activities in various brain regions (visual cortex, hippocampus, thalamus, and midbrain) of mice, obtained by Neuropixels probes during the presentation of various types of visual stimuli. Among them, we focused on natural scenes stimuli in this study. Although we also analyzed natural movie 1 and natural movie 3 and obtained qualitatively similar results, we present only the analysis of natural scenes in the main text. See Supplementary Information for the methods and the results of the analysis of natural movie 1 and natural movie 3. The natural scenes stimuli consist of 118 black-and-white images of natural scenes, each presented 50 times in random order. There were a total of 32 mice for the experiment of natural scenes stimuli, and all of them were used in the analysis.

Human fMRI dataset: Neural Scenes Dataset

The Natural Scenes Dataset (NSD)[25] consists of high-resolution (1.8-mm) whole-brain 7T functional magnetic resonance imaging (fMRI) of 8 human participants who each viewed 9,000–10,000 color natural scenes, which differed among the participants. The set of natural scenes includes a special subset of 515 images that were commonly presented across participants. We used the recordings corresponding to these 515 shared images for the analysis. In the experiment, each image was presented three times to a given participant. To control the cognitive state and encourage the deep processing of the images, participants were instructed to perform a continuous recognition task in which they reported whether the current image had been presented at any previous point in the experiment.

Representational similarity structures

Mouse neuropixels dataset

Extracting neural representations. In the analysis of the mouse dataset, trial-averaged vectors of normalized spike counts were used as neural representations for the natural scenes stimuli. Initially, for each neuron in each mouse, we counted the number of spikes during the image presentation (250 ms), followed by standard normalization. Given that each image was presented 50 times, we averaged the normalized spike counts over these 50 trials to obtain the neural representations for each image for every mouse.
Brain regions. Among the recorded brain areas in the dataset, we analyzed the following 8 areas in the visual cortex and subcortical regions: 6 areas of the visual cortex (VISp: primary visual area, VISrl: rostrolateral visual area, VISl: lateromedial visual area, VISal: anterolateral visual area, VISpm: posteromedial visual area: VISam: anteromedial visual area), 1 area of the thalamus (LGd: dorsal part of the lateral geniculate nucleus), and 1 area of the hippocampus (CA1: cornu ammonis 1). The 6 areas in the visual cortex and LGd are components of the visual system, while CA1 is a part of the memory system. Not all brain areas were measured in each mouse, and the number of neurons used for our analysis ranged from 20 to 200 per mouse per region, with an average of approximately 60.

Group-averaged representational similarity structures. To obtain a statistically reliable estimate of representational similarity structures, we group-averaged the neural activity of multiple mice, considered as a ''pseudo-mouse''. To construct a pair of pseudo-mice, we randomly divided the total of 32 mice provided in the dataset into two non-overlapping groups of 16 mice each. We then concatenated the neural representations within each group along the neurons. Therefore, the neural representations for each pseudo-mouse are represented by vectors of 400 to 1,200 neurons, varying by brain area. Using these aggregated neural representations, we estimated the group-averaged representational similarity structures for the 118 natural scene stimuli in each area for each pseudo-mouse. The dissimilarity between two stimuli was quantified using cosine distance, based on the trial-averaged normalized spike counts.

Human fMRI dataset

We followed the data preprocessing procedure of the original study.[25]

Extracting neural representations. We extracted the single neural response vector for each of the 515 images across 8 participants. By applying the Generalized Linear Model (GLM), we obtained single-trial betas, which are the estimates of the fMRI response amplitude of each voxel to each trial conducted. The single-trial betas were averaged across the three repetitions of each image. Betas for each surface vertex were z-scored within each scan session, concatenated across sessions and averaged across repeated trials for each distinct image. The resulting single-trial betas were used as the neural representation of each of the 515 images.

Brain regions. Following the previous study,[25] we defined a set of vision-related brain regions (V1, V2, V3, pVTC, aVTC, ADDOPA, PPA and RSC) on the fsaverage surface and a non-visual brain region (MTL) on the volume space. This was done by mapping the manually defined V1, V2 and V3 from each participant to fsaverage, averaging across participants and using the result to guide the definition of group-level brain regions. We then defined a posterior and anterior division of the ventral temporal cortex (pVTC and aVTC, respectively) based on anatomical criteria. Also, we defined place-selective regions (OPA, PPA and RSC) based on results of the floc experiment offered from the original dataset. For each participant, we extracted betas for vertices within each brain region (concatenating across hemispheres). The extracted betas were used as the neural representations of the 515 images for each brain region.

Group-averaged representational similarity structure. We randomly divided 8 participants into two non-overlapping groups which consisted of 4 participants each, and estimated the group-averaged representational similarity structure for each group. We repeated this random division procedure 10 times. To estimate the group-averaged representational similarity structure, we first concatenated the neural representations within each group along the voxels, and then computed representational similarity structure. The procedure was as follows. For each brain region in each participant, we first extracted a single vector of neural responses to each of the 515 stimuli. We next concatenated the vector within each group along the voxels. Finally, we computed the dissimilarity matrix of the 515 stimuli for each brain region in each participant group, where the dissimilarity between stimuli was quantified by the correlation distance $(1 - \rho)$ between the vectors of neural responses to the stimuli.

QUANTIFICATION AND STATISTICAL ANALYSIS

Comparison of neural representations

Representational similarity structures

To compare neural representations across individuals, we use similarity structures of neural representations, which we call representational similarity structures. The neural representation of stimuli is characterized by the activity patterns of individual neurons, with each stimulus represented as a point in a space of neuron dimensions. Suppose we want to compare the neural representations between individuals A and B (Figure 1). Naively, one might think to directly compare the vectors in two domains, but this approach is not feasible because the correspondence of neuron groups characterizing stimuli cannot usually be estimated between individuals (Figure 1A). However, by focusing on the similarity structures, it becomes possible to compare neural representations beyond individuals. By measuring the distance between all pairs of representations within the same space, we obtain representational dissimilarity matrices (RDM). The dissimilarity matrix represents the distances between representations of each stimulus in a brain, forming a matrix of the size of the stimulus set by the stimulus set. In this study, we use the dissimilarity matrix to compare similarity structures across individuals. There are two methods for comparing the two similarity structures: supervised alignment and unsupervised alignment.

Supervised alignment

A simple approach is to compare structures assuming a correspondence between stimulus labels, namely comparing them in a supervised manner (Figure 1B). The representative method is Representational Similarity Analysis (RSA).[4,5] RSA is a method to calculate how similar two dissimilarity matrices are. Specifically, it is computed as follows: given two dissimilarity matrices, D and D', the lower triangular part of each matrix is first extracted, forming vectors d and d' respectively. Then, the rank correlation coefficient between d and d' is calculated. This correlation coefficient serves as an indicator of the similarity between the two dissimilarity matrices.

Unsupervised alignment

A second approach is to compare structures without assuming a correspondence between stimulus labels, namely by comparing them in an unsupervised manner (Figure 1C). In general terms, unsupervised alignment is a methodology for finding the optimal

mappings between items in different domains when the correspondences between the items are completely unknown or not entirely given. As a promising approach to unsupervised alignment, the Gromov-Wasserstein optimal transport (GWOT) method[19] has been applied with great success in various fields: for example, matching of 3D objects,[19] translation of vocabularies in different languages,[20,41] and matching of single cells in single-cell multi-omics data.[21] In neuroscience, it has been successfully applied to alignment of different individual brains in fMRI data[23] and comparison of color similarity structures between different individuals.[18,22] Here, we used the GWOT method to compare the neural representations of natural scenes in different individuals.

Gromov-Wasserstein optimal transport. Gromov-Wasserstein optimal transport[19] is an unsupervised alignment technique that finds the correspondence between two point clouds in different domains based only on internal distances within each domain. The key feature of this approach is that it does not rely on any explicit pairing between points in the two domains. Instead, GWOT works by leveraging the internal distances between points within each domain. In simple terms, it tries to match points from one domain with points from another domain by comparing how close each point is to others within its own domain, rather than requiring direct, pre-defined correspondences. For instance, in our case, each point in a point cloud represents a neural representation of natural scenes, and the distances between points capture how similar or dissimilar different representations are within each domain. The goal is to find a "transportation plan" that optimally matches points from one domain to points in another, such that the structure of each domain is preserved. This alignment allows us to compare neural representations of natural scenes across domains, even when the points in each domain may not have direct counterparts. Mathematically, the goal of the Gromov-Wasserstein optimal transport problem is to find the optimal transportation plan Γ between the two point clouds in different domains, given the internal dissimilarity matrices D and D' within each domain (Figure S6). The transportation cost, i.e., the objective function, considered in GWOT is given by,

$$\min_{\Gamma} \sum_{i,j,k,l} \left(D_{ij} - D'_{kl}\right)^2 \Gamma_{ik}\Gamma_{jl}. \qquad \text{(Equation 1)}$$

Note that a transportation plan Γ must satisfy the following constraints: $\sum_j \Gamma_{ij} = p_i$, $\sum_i \Gamma_{ij} = q_j$, $\sum_{ij} \Gamma_{ij} = 1$ and $\Gamma_{ij} \geq 0$, where $\mathbf{p}$ and $\mathbf{q}$ are the source and target distributions of resources for the transportation problem, respectively, whose sum is 1. Under this constraint, the matrix Γ is considered as a joint probability distribution with the marginal distributions being $\mathbf{p}$ and $\mathbf{q}$. For the distributions $\mathbf{p}$ and $\mathbf{q}$, we set $\mathbf{p}$ and $\mathbf{q}$ to be uniform distributions. Each entry Γ_{ij} describes how much of the resources on the i-th point in the source domain should be transported onto the j-th point in the target domain. The entries of the normalized row $\frac{1}{p_i}\Gamma_{ij}$ can be interpreted as the probabilities that the i-th point in the source domain corresponds to the j-th point in the target domain.

Entropy regularization ϵ. Previously, it has been demonstrated that adding an entropy-regularization term can improve computational efficiency and help identify good local optimums of the Gromov-Wasserstein optimal transport problem.[42,43]

$$\min_{\Gamma} \sum_{i,j,k,l} \left(D_{ij} - D'_{kl}\right)^2 \Gamma_{ik}\Gamma_{jl} - \epsilon H(\Gamma), \qquad \text{(Equation 2)}$$

where $H(\Gamma)$ is the entropy of a transportation plan. To find good local optimums, we need to conduct hyperparameter tuning on ϵ in Equation 2. We select the optimized transportation plan that minimize the Gromov-Wasserstein distance without the entropy-regularization term (Equation 1) following the procedure proposed in a previous study.[21]

Evaluation of unsupervised alignment. To evaluate the extent to which the points (stimulus images) were matched with their correct pairs by the obtained optimal transportation plan, we used a measure termed "correct matching rate", defined as follows. For an element i, the following function checks if the transportation amount Γ_{ii} to its counterpart in the other domain is the maximum among the amounts to any other elements:

$$\text{Match(i)} = \begin{cases} 1, & \text{if } \operatorname{argmax}_j\left(\Gamma_{ij}\right) = i \\ 0, & \text{otherwise.} \end{cases} \qquad \text{(Equation 3)}$$

Using this value, the matching rate is then defined as the percentage of points that are matched with their correct pairs:

$$\text{Matching rate} = \frac{\sum_{i=1}^{n} \text{Match(i)}}{n}. \qquad \text{(Equation 4)}$$

In the human fMRI dataset, the stimulus images are classified into 80 categories. To evaluate the extent to which the points were matched with the points in the same category, which are not necessarily the exact correct pairs, we used the measure "category matching rate", defined as follows. When i belongs to a category C, the following function checks whether the element j that receives the maximum amount of transportation from the element i is in the same category C as i:

$$\text{Category match(i)} = \begin{cases} 1, & \text{if } \operatorname{argmax}_j\left(\Gamma_{ij}\right) \in C \\ 0, & \text{otherwise.} \end{cases} \qquad \text{(Equation 5)}$$

The category matching rate is then defined as the percentage of indices that are matched with any indices within the same category:

$$\text{Category matching rate} = \frac{\sum_{i=1}^{n} \text{Category match}(i)}{n}. \quad \text{(Equation 6)}$$

To simplify the computation in the evaluation process, we restrict each image in the COCO dataset to a single category label, despite many images being originally associated with multiple labels. Among the multiple labels assigned to an image, we retain the most prominent label, which we define as the label most frequently shared with other images in the dataset. For instance, if an image is labeled with 'dog,' 'car,' and 'tree,' we select the label that is most commonly associated with other images.