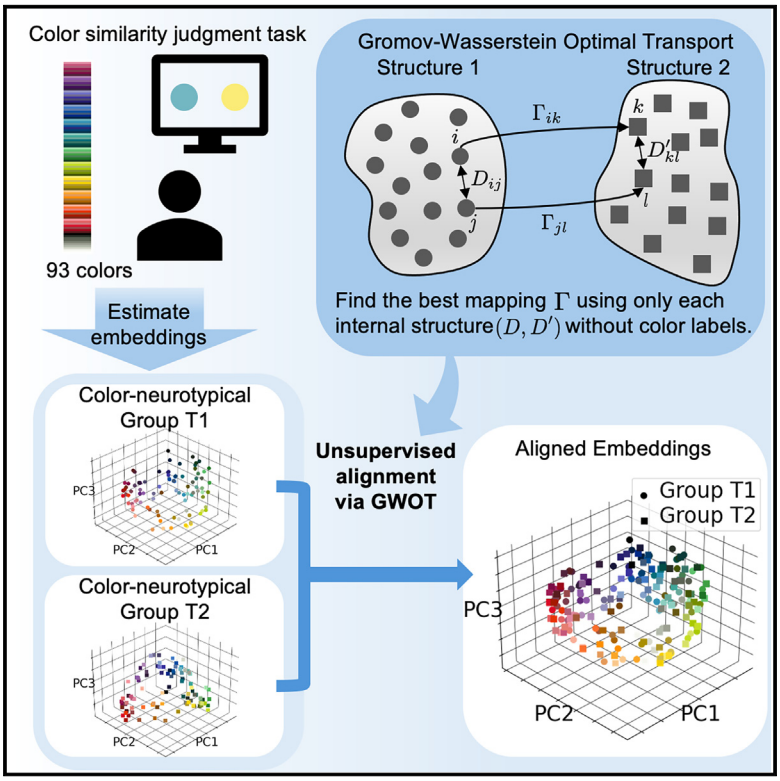
iScience

Article

Is my “red” your “red”?: Evaluating structural correspondences between color similarity judgments using unsupervised alignment

Graphical abstract



Highlights

- Structural approach to comparing different individuals’ experience of “red”
- Introduces unsupervised alignment of subjective similarity structures
- 93-color structures can be aligned between color-neurotypical people without labels
- Suggests color-typical and color-blind people differ in color experience structure

Authors

Genji Kawakita, Ariel Zeleznikow-Johnston, Ken Takeda, Naotsugu Tsuchiya, Masafumi Oizumi

Correspondence

naotsugu.tsuchiya@monash.edu (N.T.), c-oizumi@g.ecc.u-tokyo.ac.jp (M.O.)

In brief

Psychology

Kawakita et al., 2025, iScience 28, 112029
March 21, 2025
https://doi.org/10.1016/j.isci.2025.112029

CellPress

iScience

CellPress
OPEN ACCESS

Article

Is my "red" your "red"?: Evaluating structural correspondences between color similarity judgments using unsupervised alignment

Genji Kawakita,[1,2,7] Ariel Zeleznikow-Johnston,[3,4,7] Ken Takeda,[1,7] Naotsugu Tsuchiya,[3,4,5,6,8,*] and Masafumi Oizumi[1,8,9,*]

[1]Graduate School of Arts and Science, The University of Tokyo, Tokyo, Japan
[2]Department of Bioengineering, Imperial College London, London, UK
[3]School of Psychological Sciences, Monash University, Melbourne, VIC, Australia
[4]Turner Institute for Brain and Mental Health, Monash University, Melbourne, VIC, Australia
[5]Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Osaka, Japan
[6]Department of Qualia Structure, ATR Computational Neuroscience Laboratories, Kyoto, Japan
[7]These authors contributed equally
[8]These authors contributed equally
[9]Lead contact
*Correspondence: naotsugu.tsuchiya@monash.edu (N.T.), c-oizumi@g.ecc.u-tokyo.ac.jp (M.O.)
https://doi.org/10.1016/j.isci.2025.112029

SUMMARY

Whether one person's subjective experience of the "redness" of red is equivalent to another's is a fundamental question in consciousness studies. Intersubjective comparison of the relational structures of sensory experiences, termed "qualia structures", can constrain the question. We propose an unsupervised alignment method, based on optimal transport, to find the optimal mapping between the similarity structures of sensory experiences without presupposing correspondences (such as "red-to-red"). After collecting subjective similarity judgments for 93 colors, we showed that the similarity structures derived from color-neurotypical participants can be "correctly" aligned at the group level. In contrast, those of color-blind participants could not be aligned with color-neurotypical participants. Our results provide quantitative evidence for interindividual structural equivalence or difference of color qualia, implying that color-neurotypical people's "red" is relationally equivalent to other color-neurotypical's "red", but not to color-blind people's "red". This method is applicable across modalities, enabling general structural exploration of subjective experiences.

INTRODUCTION

The question of whether sensory experiences are intersubjectively equivalent is a central concern in the study of consciousness. Some researchers consider the question impossible to answer because of the intrinsic, ineffable, and private nature of subjective experience.[1] Although direct description of our experiences in a fashion that allows for intersubjective comparison may be impossible, indirect characterization of experience is empirically feasible and is considered a promising research program.[2–13] One notable approach is to analyze reports of subjective similarities between sensory experiences.[14–18] Relationships between sensory experiences, such as similarity, allow for the structural investigation of phenomenal consciousness.[19–21]

Based on this idea, we formally introduce a paradigm, which we call the "qualia structure paradigm". In our usage, qualia are simply the qualities of subjective experience, the "what it is like" character of conscious states such as "what it is like to see a green" or "what it is like to feel a toothache". This broad definition of qualia does not include whether qualia are fundamentally intrinsic, private, and incommunicable.[22] Although we do not argue here whether or not there might be a purely intrinsic and incommunicable aspect to qualia, in this paradigm we assume that at least some phenomenal properties of qualia can be empirically measured and communicated by focusing on the relationships between qualia, which we call "qualia structures". We use the term "structure" because "structure" is a more general concept than "space" in mathematics (see Kleiner et al.[13] for details and other similar terminology in the literature). Furthermore, "structure" is gaining more traction in consciousness research.[23]

The qualia structure paradigm (Figure 1A) consists of two main steps. The first step is to estimate relational structures of subjective experiences from psychophysics experiments, termed qualia structures, and the second step is to compare the estimated qualia structures between individuals. In this paper, we argue that it is essential to compare the estimated qualia structures without assuming the correspondence of qualia of the same sensory stimuli between individuals, and propose a framework for this comparison using unsupervised alignment. To the

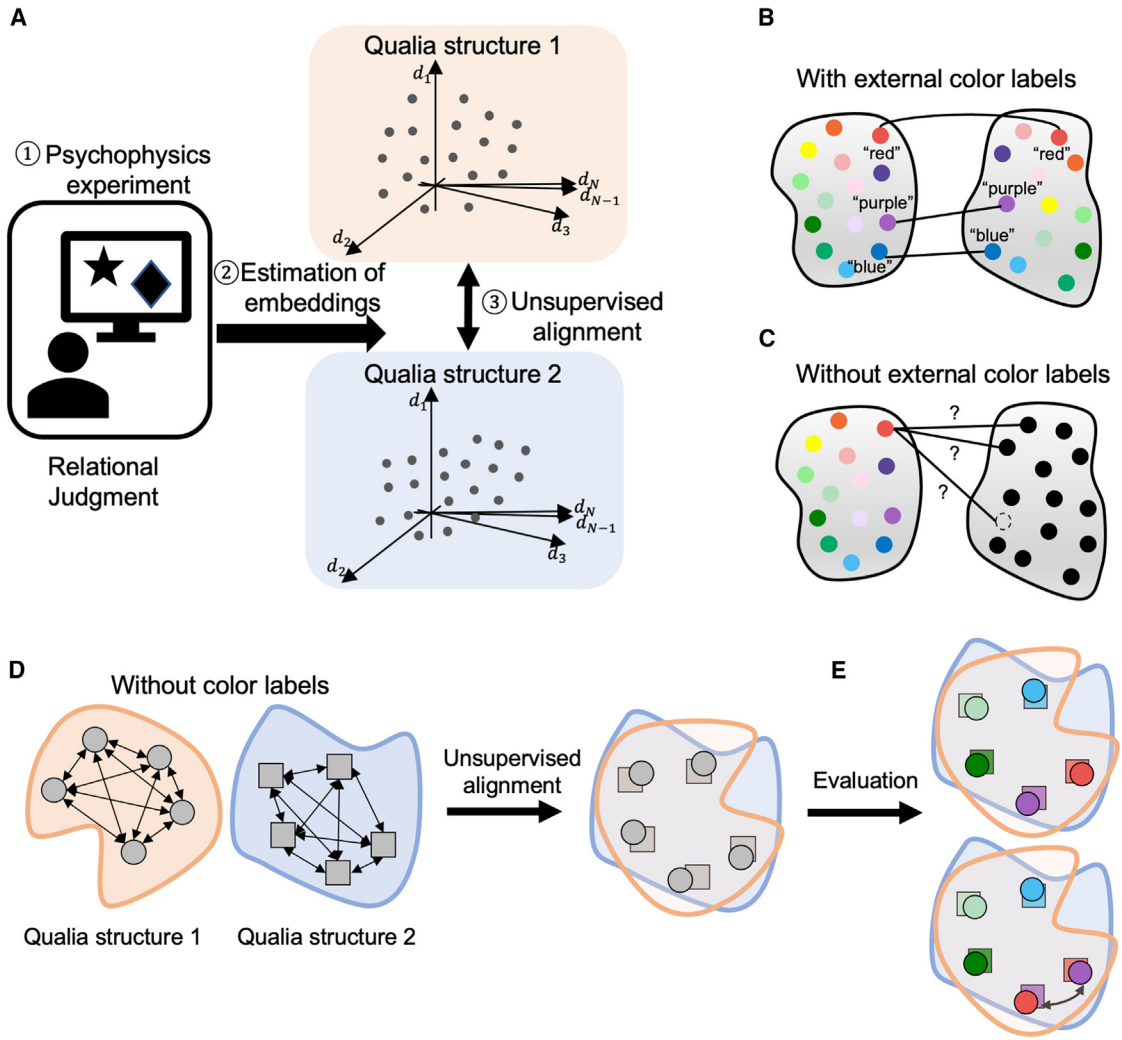


Figure 1. Schematics of concepts in the qualia structure paradigm

(A) Two steps in the qualia structure paradigm. The first step is to collect subjective reports through relational judgments between stimuli that enable estimation of the relational structure of sensory experiences, i.e., qualia structure. The second step is to align qualia structures from different individuals in an unsupervised manner to quantify the degree of similarity of their qualia structures.

(B) Supervised alignment of color qualia structures, which assumes correspondence between qualia evoked by the same external stimuli across different individuals.

(C) Unsupervised alignment of color qualia structures, which does not assume correspondence between qualia across different individuals. All possible correspondences are taken into consideration. A particular color quale for an individual may not have an exact correspondence to a particular quale of another individual, as indicated by the dotted circle.

(D) Aligning qualia structures of different individuals in an unsupervised manner without any external labels, solely based on the internal relationships of the embeddings.

(E) Evaluation of unsupervised alignment using external labels.

best of our knowledge, the importance of unsupervised comparison without relying on the labels of external sensory stimuli has not yet been emphasized, nor has a practical method for realizing such unsupervised comparison been proposed in the literature. The main contribution of the present work is to introduce a practical method for unsupervised alignment of estimated qualia structures and to demonstrate the utility of this approach on experimental data of color similarity structures. Both steps are described in detail below.

The first step is to collect detailed subjective reports about the relations between sensory experiences (qualia) through psychophysics experiments.[18] We then estimate the embeddings of qualia for different participants that best explain the participants' similarity judgments. The set of qualia embeddings is represented as points in space (Figure 1A) and is considered as a 'qualia structure'. Importantly, the relations of the qualia are represented as the "distances" of the embeddings, and we can estimate the dissimilarity matrices of the qualia based on the

distance (e.g., Euclidean distance) between the embeddings. Note that we do not consider the estimated embeddings based on psychophysical experiments to represent the entire quality of experience (e.g., color experience), but rather a particular aspect of experience. There is also the possibility that similarity judgments contain non-phenomenological information, including semantic concepts or other more cognitive information. Thus, an empirically estimated qualia structure should only be considered as an approximation of qualia structures or an empirically measurable aspect of qualia structures. We should always be cautious about the task requirements of psychological experiments, whether they can extract phenomenological aspects.

Having obtained two qualia structures from different participants (Figure 1A), the second step is to compare these structures and evaluate the extent to which they are similar, without assuming a correspondence between individual qualia from one structure to the other. This is in contrast to previous analyses of dissimilarity matrices, which typically assumes an "external" correspondence at the stimulus level: my experience of "red" evoked by a red stimulus corresponds to your experience of "red" (Figure 1B). This type of supervised comparison between dissimilarity matrices, known as representational similarity analysis (RSA), has been widely used in neuroscience to compare various similarity matrices obtained from behavioral and neural data.[24,25] However, there is no guarantee that the same stimulus will necessarily evoke the same corresponding subjective experience in different individuals. Accordingly, when considering which stimuli evoke which qualia for different individuals, we need to consider all possibilities of correspondence. For example, my "red" could correspond to your "red", "green", "purple", or it could lie somewhere between your "orange" and "pink" (Figure 1C).

If the unsupervised comparison of two different individuals' qualia structures (i.e., comparison without presupposing any correspondence between their qualia) results in an exact one-to-one mapping (e.g., red-to-red), what can we infer about their subjective experience? It should be emphasized that we do not take this as sufficient evidence that two people have exactly the same qualia (e.g., my "red" is your "red"). Rather, this should serve as one of the necessary conditions to be satisfied for two participants to possess the same experiences, which were previously called structural constraints.[7] We also conjecture that the contraposition is true, i.e., if two structures are not exactly mapped, two people would necessarily have different experiences. Our approach would provide a quantitative measure of the extent to which the structural conditions (or constraints) are met across individuals.

To account for all possible correspondences, we propose to use an unsupervised alignment method for quantifying the degree of similarity between qualia structures. As shown in Figure 1D, in unsupervised alignment, we do not attach any external (stimuli) labels to the embeddings. Instead, we try to find the best matching between qualia structures based only on their internal relationships (see STAR Methods). After finding the optimal alignment, we can use external labels, such as the identity of a color stimulus (Figure 1E), to evaluate how the embeddings of different individuals relate to each other. This allows us to determine which color embeddings correspond to the same color embeddings across individuals, and which do not. Checking the assumption that these external labels are consistent across individuals allows us to assess the degree of inter-individual correspondences between qualia structures obtained from different participants.

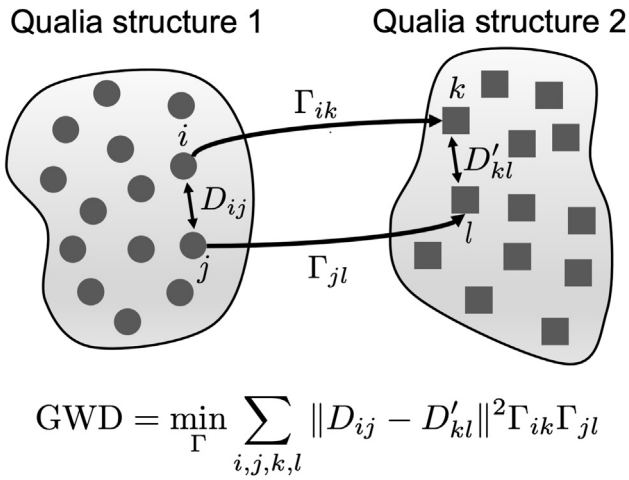


Figure 2. Schematic of Gromov-Wassersetin optimal transport
The elements of matrices D and D' are the distances between the embeddings. Γ is the transportation matrix indicating the probability of an embedding in one qualia structure corresponding to an embedding in the other qualia structure.

To this end, we used the Gromov-Wasserstein optimal transport (GWOT) method,[26] which has been applied with great success in various fields.[27–32] GWOT aims to find the optimal transportation plan Γ between two point clouds in different domains based on the distance D (or D') between points within each domain (Figure 2). Importantly, the distances (or correspondences) between points "across" different domains are not given while those "within" the same domain are given. GWOT aligns the point clouds according to the principle that a point in one domain should correspond to a point in the other domain that has a similar relationship to other points within its respective domain (see STAR Methods for the details). The optimal transportation plan Γ can be interpreted as the probability of an embedding in one qualia structure corresponding to an embedding in the other qualia structure. By using the optimal transportation plan Γ, we can evaluate the degree to which two qualia structures match correctly.

RESULTS

To assess the validity and utility of the qualia structure paradigm, we apply this framework to a similarity judgment dataset involving 93 colors as a representative and tractable case study. The relatively large number of colors enables us to investigate complex and nuanced qualia structures of colors, which is less feasible with previous datasets examining a smaller number of colors 14, 16–18. In addition, the large number of colors necessitates the computational efficiency of our method, as a brute-force search method considering all possible correspondences used in previous studies[33] would not be practical.

In this study, using data from both color-neurotypical and color-atypical participants, we address two questions: (1) whether color qualia structures can be aligned within color-neurotypical and color-atypical participant groups, separately, and (2) whether color qualia structures can be aligned across color-neurotypical and color-atypical participant groups. The first analysis is necessary to determine whether there are structures that are sufficiently common across participants to be alignable within color-neurotypical participants or within color-atypical participants. These cases also serve as a positive control, where we should be able to align two qualia structures, using our unsupervised alignment method, which relies only on the internal relationships. After establishing the validity of our methods, we quantified the degree to which the qualia structures of these two populations can be aligned.

Massive online experiment of color similarity judgment

We collected similarity judgments between 93 colors from 426 color-neurotypical and 257 color-atypical participants using an online cloud sourcing service. 257 color-atypical participants self-reported as color blind. Their reports were verified by a modified online Ishihara test (see STAR Methods). Each participant provided pairwise dissimilarity judgments for a randomly assigned subset of the 4,371 possible color pairs (including the same color pairs).

In this study, we considered the alignment between the color similarity structures on a participant group basis by aggregating the similarity judgments of many participants to estimate a group-level color similarity structure. This is because the number of color pairs reported by each participant was only 162 at most, which is too small to reliably estimate the entire color similarity structure of 93 colors. Thus, we did not address the problem of one-to-one alignment at the individual level (see discussion). As described further, we first considered alignment within color-neurotypical groups, then within color-atypical groups, and then finally between these participant groups.

Unsupervised alignment of color similarity structures

Unsupervised alignment between color-neurotypical participants

First, we considered the alignment within color-neurotypical participants (between subgroups of color-neurotypical participants). We aggregated the similarity judgment responses of 128 randomly selected color-neurotypical participants out of the total 426 color-neurotypical participants and created a pair of non-overlapping participant groups, each consisting of 128 participants. We repeated this random sampling 20 times. We show the results of one of the 20 samples in Figures 3A–3E, and all the results of the 20 different samples in Figure 3F.

As a demonstration, we show the embeddings of 93 colors for a certain random pair of groups, denoted as group T1 and T2 in Figure 3A. For each group, we estimated the embeddings that best explained the experimentally obtained similarity responses, based on the procedure described in detail in STAR Methods. We then applied principal-component analysis to reduce the dimensions of the embeddings to 3 for visualization (Figure 3A). From the estimated embeddings, we obtained the dissimilarity matrices D by computing the Euclidian distances between the estimated embeddings (Figure 3B), where the entry, D_{ij}, represents the subjective dissimilarity between the two experiences of the i-th and j-th colors.

We then compared the two color similarity structures by performing an unsupervised alignment based on entropic GWOT (Equation 11) on the estimated dissimilarity matrices (Figure 3B). Since entropic GWOT is a non-convex optimization problem involving hyperparameter search of ϵ, which controls the degree of entropy reguralization, we performed a total of 200 optimization iterations with different ϵ values and initializations of transportation plans to search for a global optimum. The points in Figure 3C correspond to the local minimum found in each iteration of the optimization performed on different ϵ values. We selected the local minimum with the lowest Gromov-Wasserstein distance (GWD) as the optimal solution (shown in the red circle in Figure 3C).

From the optimization process, we finally obtained the optimal transportation plan Γ between group T1 and T2 (Figure 3D). As shown in Figure 3D, most of the diagonal elements in Γ have high values, indicating that most of the colors in one group correspond with a high probability to the same colors in the other group. To quantitatively assess the degree of correspondence, we computed the top-1 matching rate of the 93 colors (see STAR Methods for details), which was 38%. As can be seen in Figure 3C, the local minima with low GWD (in the y axis) tend to yield a high matching rate (points with yellowish color), which is necessary for unsupervised alignment to achieve a high matching rate.

After applying GWOT, we performed an alignment of the two sets of embeddings, which is visualized in Figure 3E. Although the optimized transportation plan Γ provides the rough correspondence between the embeddings of the color similarity structures, we can find a more detailed mapping in the original space of the embeddings. As described in STAR Methods, we aligned the embeddings of group T1 (denoted by X) with those of group T2 (denoted by Y) by finding the orthonormal rotation matrix Q using the optimized transportation plan Γ obtained by GWOT. In Figure 3E, we plotted the embeddings of group T1, X, and the aligned embeddings of group T2, QY. This visualization clearly demonstrates that the embeddings of similar colors from both groups are closely located to each other, indicating that similar colors are ''correctly'' aligned by the unsupervised alignment (See Video S1 for visualization of the embeddings from multiple viewing angles). Note that although the colors in Figure 3E are used for evaluation purposes only, the entire alignment process was performed in a purely unsupervised manner, without relying on the color labels.

By performing the same analysis for all 20 random pairs of participant groups, we obtained the top-k matching rate of the 20 random samples (Figure 3F). The results of the particular example shown in Figure 3D are highlighted by the red arrows. The averages of the top 1, 3, and 5 matching rate over 20 random samples are 51%, 83%, and 94%, respectively. Importantly, all values of the matching rates from 20 random samples are well above the respective chance level (1.1%, 3.2%, and 5.4%). This suggests that there are sufficiently common structures among color-neurotypical participants to enable their alignment in an unsupervised manner. This result validates the effectiveness of our unsupervised alignment method based on GWOT,

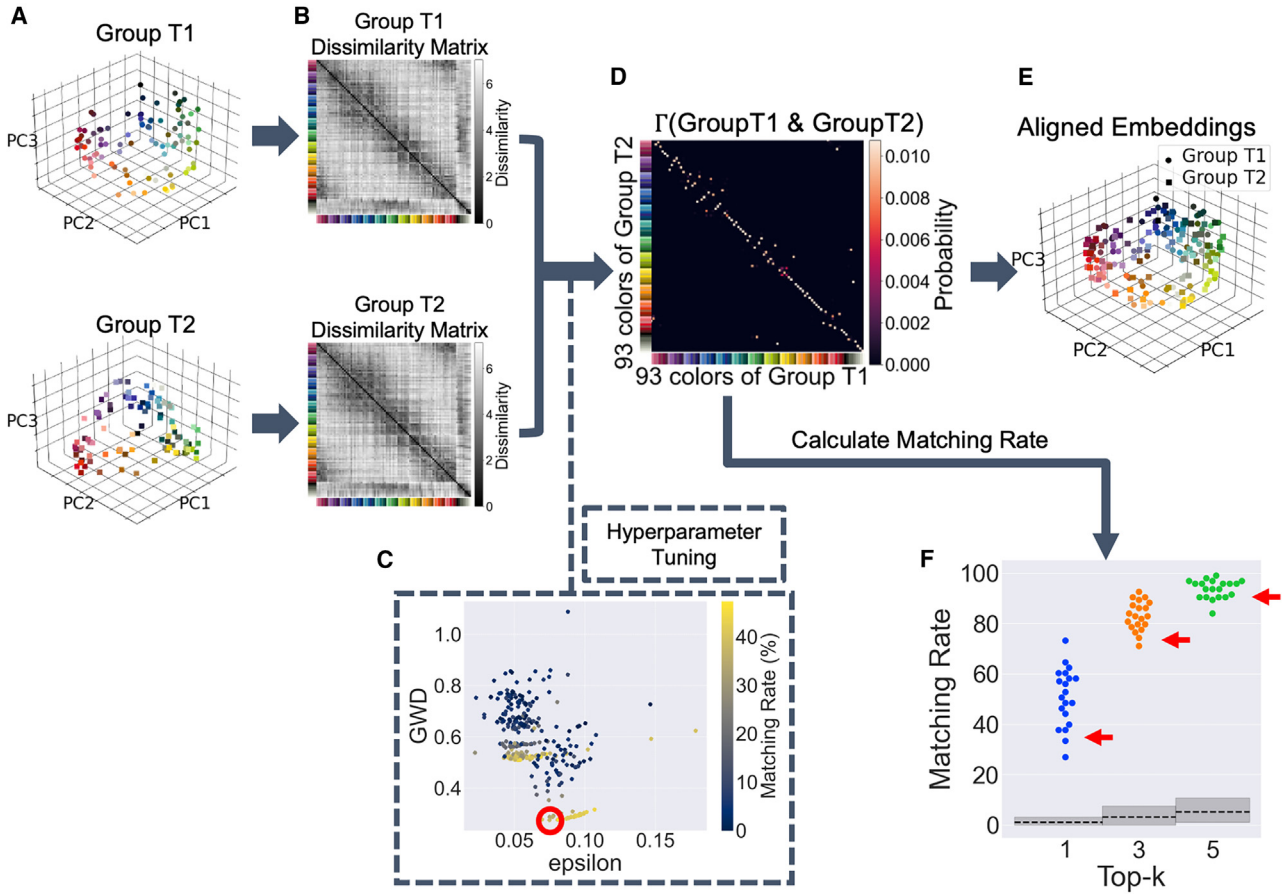


Figure 3. Unsupervised alignment between color similarity structures of color-neurotypical participant groups

(A) Three-dimensional embeddings of 93 colors for group T1 (top) and Group T2 (bottom). Embeddings were obtained by fitting a 20-dimensional model to each group's aggregated similarity judgments, then projecting onto the first three principal components (PC1, PC2, and PC3) for visualization. Each point corresponds to one color stimulus.

(B) Dissimilarity matrices for group T1 (top) and group T2 (bottom). Each matrix entry (i, j) indicates the estimated subjective dissimilarity between colors *i* and *j*, derived from the Euclidean distances of the embeddings in (A).

(C) Results of the entropic Gromov-Wasserstein optimization across multiple hyperparameter values (ϵ). Each point is a local optimum from a distinct initialization ($N = 200$ total initializations). The vertical axis shows the Gromov-Wasserstein distance (GWD), and the horizontal axis shows ϵ. Points are colored by the top-1 matching rate. The red circle denotes the selected optimal solution (lowest GWD).

(D) The optimal transportation plan Γ between group T1 and group T2. Rows correspond to the 93 colors of group T1, columns correspond to the 93 colors of group T2, and each cell shows the probability that color *i* in group T1 is matched to color *j* in group T2 under the optimal plan.

(E) Aligned embeddings for group T2 (squares) after rotation to match group T1 (circles). Points are plotted in group T1's embedded space. The axes (PC1, PC2, PC3) are the same principal components as in (A).

(F) Top-*k* matching rate (vertical axis) for 20 random splits of color-neurotypical participants into group T1 and group T2. Blue, orange, and green points show the matching rate for $k = 1, k = 3$, and $k = 5$, respectively, across different splits ($N = 20$). The dotted line represents the mean chance-level matching rate, and the gray shaded region represents the 95% percentile interval of the chance-level matching rate (computed via random permutations; see STAR Methods). Red arrows highlight the example results for the specific group T1 and group T2 shown in Panels A–E.

which relies solely on the internal relationships of color similarity structures, and demonstrates its effectiveness in scenarios where it is expected to work.

Unsupervised alignment between color-atypical participants

Next, we considered the alignment between different color-atypical participant groups (see STAR Methods and Figure S1 for screening details). Most of these participants are likely to have red-green color vision deficiencies, as further detailed in Figure S1.

We investigated whether the color similarity structures of color-atypical participant subgroups could be aligned similarly to those of color-neurotypical participants. To this end, we replicated the analysis performed with color-neurotypical participants, in which pairs of participant groups consisting of randomly selected 128 participants were formed 20 times. Figures 4A–4E show the results for a particular pair of color-atypical groups, group A1 and A2, all in the same format as Figure 3. Figure 4A, 4B, and 4C, show the embeddings, the dissimilarity matrices, and details of the GWOT

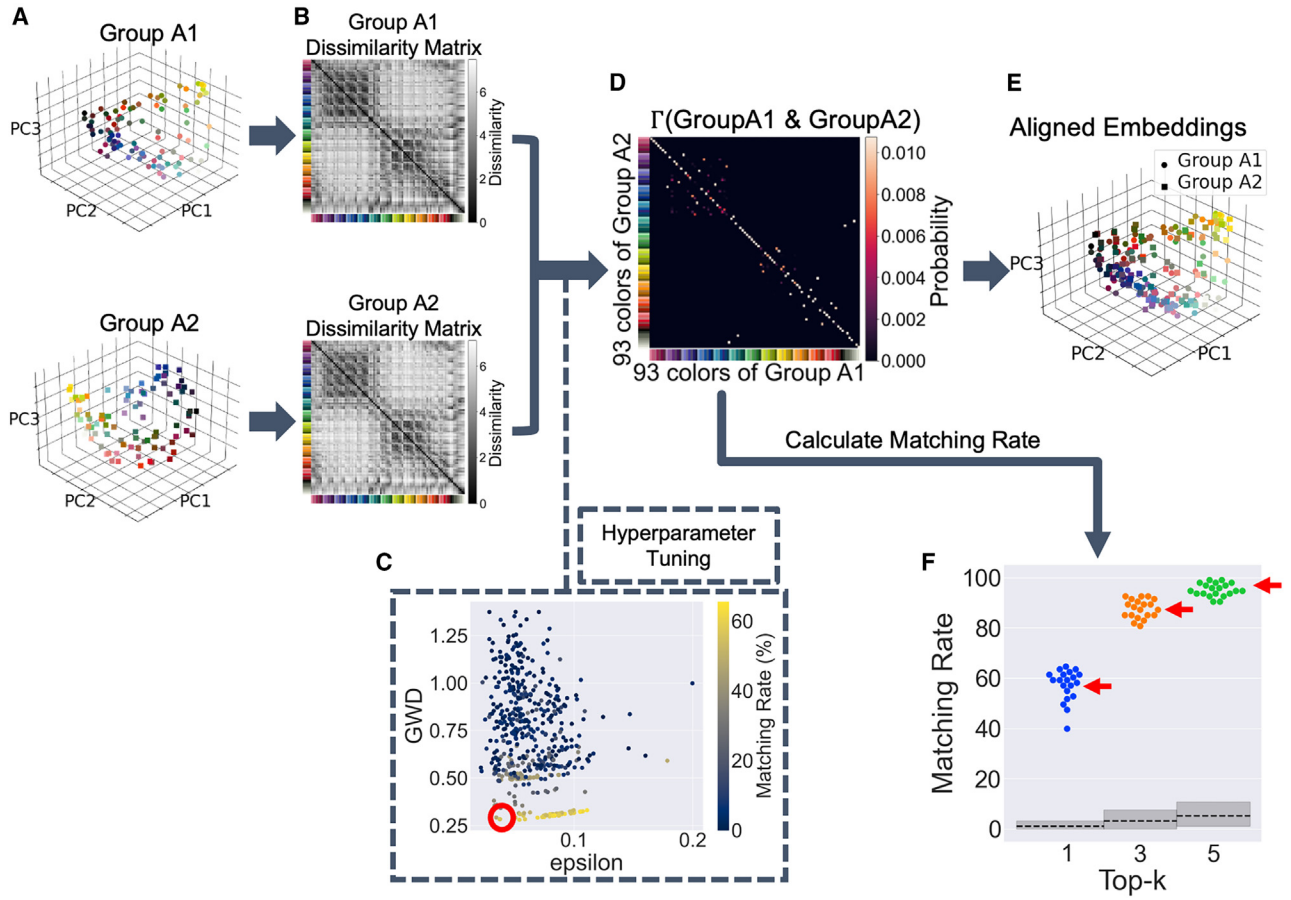


Figure 4. Unsupervised alignment between color similarity structures of color-atypical participant groups

(A) Three-dimensional embeddings of 93 colors for group A1 (top) and group A2 (bottom). Embeddings were obtained by fitting a 20-dimensional model to each group's aggregated similarity judgments, then projecting onto the first three principal components (PC1, PC2, and PC3) for visualization. Each point corresponds to one color stimulus.

(B) Dissimilarity matrices for group A1 (top) and Group A2 (bottom). Each matrix entry (i, j) indicates the estimated subjective dissimilarity between colors *i* and *j*, derived from the Euclidean distances of the embeddings in (A).

(C) Results of the entropic Gromov-Wasserstein optimization across multiple hyperparameter values (ϵ). Each point is a local optimum from a distinct initialization ($N = 200$ total initializations). The vertical axis shows the Gromov-Wasserstein distance (GWD), and the horizontal axis shows ϵ. Points are colored by the top-1 matching rate. The red circle denotes the selected optimal solution (lowest GWD).

(D) The optimal transportation plan Γ between group A1 and group A2. Rows correspond to the 93 colors of group A1, columns correspond to the 93 colors of group A2, and each cell shows the probability that color *i* in group A1 is matched to color *j* in group A2 under the optimal plan.

(E) Aligned embeddings for group A2 (squares) after rotation to match group A1 (circles). Points are plotted in group A1's embedded space. The axes (PC1, PC2, and PC3) are the same principal components as in (A).

(F) Top-*k* matching rate (vertical axis) for 20 random splits of color-atypical participants into group A1 and group A2. Blue, orange, and green points show the matching rate for $k = 1$, $k = 3$, and $k = 5$, respectively, across different splits ($N = 20$). The dotted line represents the mean chance-level matching rate, and the gray shaded region represents the 95% percentile interval of the chance-level matching rate (computed via random permutations; see STAR Methods). Red arrows highlight the example results for the specific group A1 and group A2 shown in A–E.

optimization results over 200 iterations, respectively. Figure 4D shows the optimal transportation plan Γ for group A1 and A2. We can see that most of the diagonal elements in Γ have high values, resulting in a top-1 matching rate of 59%. In Figure 4E, the embeddings of group A1 and the aligned embeddings of group A2 are plotted, demonstrating the effectiveness of the unsupervised alignment, as evidenced by the close placement of similar colors from both groups (See Video S2 for visualization of the embeddings from multiple viewing angles).

The top-k matching rate of the 20 random samples (Figure 4F) further confirms the validity of our methods. The average of the top 1, 3, and 5 matching rate over 20 random samples is 57%, 87%, and 95%, respectively. All values of the matching rates significantly exceed their corresponding chance levels, which are 1.1%, 3.2%, and 5.4%. This result suggests that at the group level, there are sufficient common structures among color-atypical participants that they can be aligned in an unsupervised manner, even though the degree of red-green color deficiency varies among individual participants.

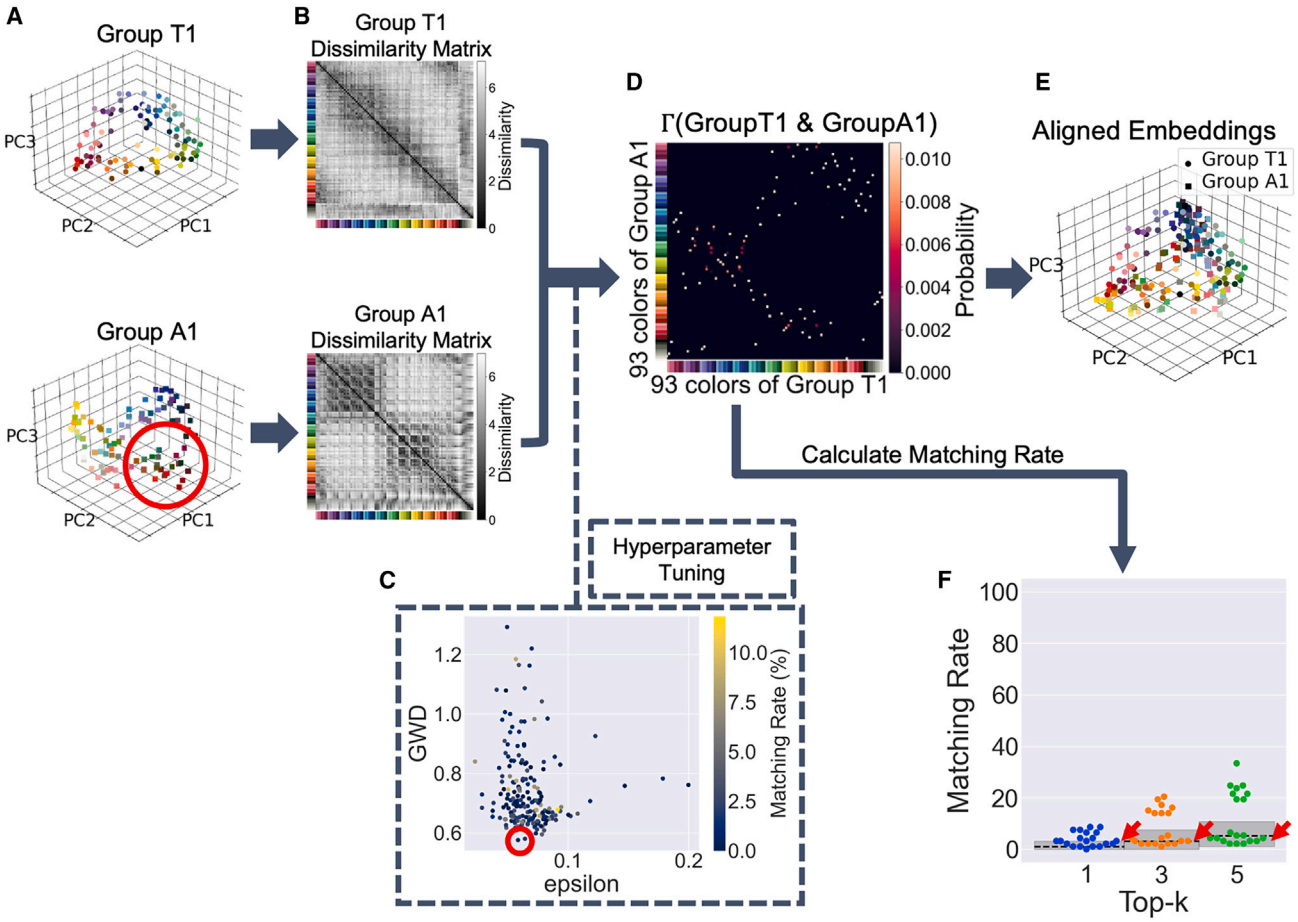


Figure 5. Unsupervised alignment between color similarity structures of color-neurotypical and atypical participant groups

(A) Three-dimensional embeddings of 93 colors for group T1 (top) and group A1 (bottom). Embeddings were obtained by fitting a 20-dimensional model to each group's aggregated similarity judgments, then projecting onto the first three principal components (PC1, PC2, and PC3) for visualization. Each point corresponds to one color stimulus.

(B) Dissimilarity matrices for group T1 (top) and group A1 (bottom). Each matrix entry (i, j) indicates the estimated subjective dissimilarity between colors *i* and *j*, derived from the Euclidean distances of the embeddings in (A).

(C) Results of the entropic Gromov-Wasserstein optimization across multiple hyperparameter values (ϵ). Each point is a local optimum from a distinct initialization ($N = 200$ total initializations). The vertical axis shows the Gromov-Wasserstein distance (GWD), and the horizontal axis shows ϵ. Points are colored by the top-1 matching rate. The red circle denotes the selected optimal solution (lowest GWD).

(D) The optimal transportation plan Γ between group T1 and group A1. Rows correspond to the 93 colors of group T1, columns correspond to the 93 colors of group A1, and each cell shows the probability that color *i* in group T1 is matched to color *j* in group A1 under the optimal plan.

(E) Aligned embeddings for group A1 (squares) after rotation to match group T1 (circles). Points are plotted in group T1's embedded space. The axes (PC1, PC2, and PC3) are the same principal components as in (A).

(F) Top-*k* matching rate (vertical axis) for 20 random splits, each pairing color-neurotypical participants (group T1) with color-atypical participants (group A1). Blue, orange, and green points show the matching rate for $k = 1$, $k = 3$, and $k = 5$, respectively, across different splits ($N = 20$). The dotted line represents the mean chance-level matching rate, and the gray shaded region represents the 95% percentile interval of the chance-level matching rate (computed via random permutations; see STAR Methods). Red arrows highlight the example results for the specific group T1 and group A1 shown in A–E.

Unsupervised alignment between color-neurotypical and color-atypical participants

Finally, we investigated to what extent the color similarity structures of color-neurotypical and color-atypical participants could be aligned. For this purpose, we separately sampled 128 participants from both color-neurotypical and color-atypical participants and paired a color-neurotypical participant group with a color-atypical participant group. This procedure was repeated 20 times, resulting in 20 pairs, each consisting of a group of color-neurotypical participants and a group of color-atypical participants.

As an illustrative case among these 20 random pairs, we show the estimated embeddings (Figure 5A) and the dissimilarity matrices of a color-neurotypical and a color-atypical participant group (Figure 5B), labeled as group T1 and A1, respectively. Upon visual inspection of the embeddings of the color-neurotypical and atypical group in Figure 5A, we can see that while the overall structures of the color-neurotypical and atypical group

are similar, there are distinct differences. In particular, greenish and reddish colors are close in the embedding space of the color-atypical participants, as highlighted by the red circle in Figure 5A, while they appear distant in the space of the color-neurotypical participants. Despite these differences, the two dissimilarity matrices in Figure 5B show a significant degree of similarity. This is quantitatively supported by the Pearson's correlation coefficient of 0.66 between the dissimilarity matrices of group T1 and group A1. While this coefficient is somewhat lower than the Pearson's correlation coefficient between the dissimilarity matrices of the color-neurotypical groups (group T1 and T2 in Figure 3B, with $\rho = 0.88$), and between the color-atypical groups (group A1 and A2 in Figure 4B, with $\rho = 0.91$), it still represents a substantial correlation.

Using the dissimilarity matrices, we performed the unsupervised alignment based on GWOT between the color-neurotypical participant group and the color-atypical participant group. We performed a total of 200 optimization iterations on different ϵ values and selected the local minimum with the lowest GWD as the optimal solution (highlighted by the red circle in Figure 5C). In Figure 5C, we observe that local minima with low values of GWD have low matching rates, leading to unsuccessful alignment in terms of matching rate. As can be seen in Figure 5D, the optimal transportation plan Γ is not lined up diagonally (the diagonal elements of Γ are very small), unlike the optimal transportation plan between the color-neurotypical participant groups shown in Figure 3D or that between the color-atypical participant groups shown in Figure 4D. The optimal solution with the lowest GWD has a top-1 matching rate of 1.1%, which is close to the chance level (1.1%). In Figure 5E, we plotted the embeddings of group T1 and the aligned embeddings of group A1. Unlike the results seen in Figure 3E and in Figure 4E, here the embeddings of similar colors from the two groups are not positioned closely, indicating that similar colors are not correctly aligned by the unsupervised alignment (See Video S3 for visualization of the embeddings from multiple viewing angles).

By performing the same analysis for all 20 random pairs of the participant group, we obtained the top-k matching rate of the 20 random samples (Figure 5F). The averages of the top 1, 3, and 5 matching rates over 20 random samples are 3.8%, 8.3%, and 11.7%, respectively, which are slightly higher than, but close, to their respective chance levels (1.1%, 3.2%, and 5.4%). Almost all values of the matching rates from 20 random samples are close to chance, with a few exceptions.

In addition to the matching rate, we also evaluated the GWD for all 20 random pairs of the participant group (Figure S2B). We found that the GWD was consistently higher than the GWD in the case of alignment within color-neurotypical participant groups or within color-atypical participant groups.

Taken together, these results showed that the structures between color-neurotypical participants and color-atypical participants are different, both in terms of matching rate and GWD.

DISCUSSION

For a long time, assessing the similarity of subjective experiences across participants has been primarily considered as a philosophical question, rather than an empirical problem to be tackled scientifically.[33–36] While many previous studies have employed multidimensional scaling (MDS) to analyze structural aspects of dissimilarity judgments between perceptual experiences, they have invariably refrained from investigating whether qualia can be uniquely identified by their relational properties alone.[19–21] To address this problem, we have proposed the "qualia structure" paradigm.

By using the proposed unsupervised alignment method, we were able to obtain qualitatively different results from those that would be obtained by the conventional supervised alignment method, such as RSA.[24] First, we showed that the color similarity structures within color-neurotypical or color-atypical participants can be aligned based only on similarity relationships of colors without using any external labels. One might think that these results are almost obvious, since the correlation between the dissimilarity matrices (Figures 3B and 4B) are very high ($\rho = 0.88$, $\rho = 0.91$, respectively). However, in simulations, it is easy to create examples where the two structures are not correctly aligned, even when the correlation coefficient between two structures is very high (e.g., $\rho = 0.9$) (see Figure 3 in the study by Sasaki et al.[32]). In general, similarity measures based on supervised comparison, such as the correlation coefficient alone, cannot tell us whether two structures are similar enough to align in an unsupervised manner.

In addition, we also showed that we could not unsupervisedly align the color similarity structures between color-neurotypical and color-atypical participants, even though the correlation coefficient between the dissimilarity matrices is reasonably high ($\rho = 0.66$). Given the high correlation coefficient, the failure of the unsupervised alignment is not entirely expected. As we show in the simulations shown in Figures S3, even though the correlation coefficients between the dissimilarity matrices are all equal at about 0.7, the matching rate of the unsupervised alignment can vary significantly, from near perfect alignment to chance level alignment. The unsupervised alignment probably failed because of local structural differences, i.e., greenish colors and reddish colors are close in the embedding space of color-atypical participants (Figure 5A), even though the overall structures are similar. Beyond traditional measures such as Pearson's correlation coefficient, our method provides a more essential structural characterization of how two structures are similar or different, which will be crucial for future investigations of qualia structures across psychological, neuroscientific, and computational fields.

As we emphasized in the introduction, the structural correspondences evaluated by our unsupervised approach are not direct evidence that different people have the same subjective experiences. Rather, it should be considered as one of the necessary conditions for different people to have the same subjective experiences. For sufficiency, other conditions may also need to be met. As a cautious example to illustrate that structural correspondence is not necessarily sufficient, consider a case of alignment between a qualia structure obtained from human participants and a similarity matrix generated by the internal representation of sensory stimuli in a neural network model. In another study using the same unsupervised approach,[37] we have indeed shown that recent large language models (LLM) can also generate a strikingly similar color

similarity structure that can be unsupervisedly aligned with human color-neurotypical participants. However, we do not claim that LLMs are conscious and possess equivalent qualia to humans, because structural correspondence alone is insufficient to make such a claim. Thus, the evidence for structural correspondence should be treated with caution and properly interpreted with full recognition of its limitations. Nevertheless, we believe that our unsupervised approach will be an important addition to the structural approach to consciousness, which has recently been actively explored with the aim of quantifying qualia from a structural perspective.[13]

Although in this paper, we only considered group-based alignment because the number of trials obtained from each participant was insufficient for reliable unsupervised alignment (see Figure S2), it is interesting to consider individual-based alignment and assess the degree of the individual difference even for color-neurotypical participants. To obtain statistically reliable results, we would roughly estimate from Figure S2 that at least more than 4,000 trials of similarity judgments are needed for each individual participant, which is almost same as the total number of pairs of 93 colors. Although we were not able to directly address the question of the correspondence of color similarity structures at the individual level, our goal in this paper is rather to propose a framework for exploring this question and to stimulate interest and further research in this area. To assess individual differences based on individual-level alignment, we plan to conduct experiments in which we collect full similarity judgments of pairs of 93 colors from individual participants as future research.

While we focused only on color similarity, our method has the potential to be applied to a wide range of subjective experiences and different modalities (e.g., natural objects,[38–40] emotions,[41–43] semantic concepts[44,45] etc.). Our unsupervised approach offers a powerful tool for assessing the intersubjective correspondence of various qualia structures and for deepening our understanding of qualia from a structural point of view.

Limitations of the study

While our current study focuses on unimodal color similarity judgments, we recognize the importance of considering cross-modal associations in the broader context of phenomenological experiences. The complex interplay between sensory modalities, such as the common description of certain musical experiences as having color-like qualities, highlights the need for a more comprehensive multimodal approach. Future research should extend this framework to include multimodal relational judgments about experiences based on similarity judgments, yielding large similarity matrices that encompass both within-modality and cross-modality comparisons. For example, while a particular musical tone might be judged to be similar to a particular color, it would generally be more similar to other auditory experiences than to visual ones if there is a categorical structure over similarity judgments within the same modalities. In this way, the particular auditory experience would not be confused with the particular visual experience in the embedding space. By accumulating a rich set of relational data across modalities, we can construct a more comprehensive understanding of the relational structure of subjective experience. The general principle is that the more relationships there are, the more uniquely the experience can be characterized.

It is important to recognize that the study of subjective experience involves a diverse range of experimental approaches, each with its own strengths and limitations. While our study employed pairwise similarity judgments to construct color similarity spaces, this is just one of several possible methods for investigating mental qualities. Alternative experimental techniques include odd-one-out tasks,[38] color naming experiments,[46] and just-noticeable-difference (JND) judgments. The JND method, which requires participants to discriminate between nearly identical stimuli, may provide a finer-grained representation of perceptual relationships compared to broader similarity judgments.

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

- Color similarity judgment data have been deposited at Open Science Framework and are publicly available at Open Science Framework: https://osf.io/9xwr2/.
- Code for the behavioral experiments has been deposited at Open Science Framework and is publicly available at Open Science Framework: https://osf.io/9xwr2/. Code used for the analyses in this article is publicly available at the following repository at GitHub: https://github.com/oizumi-lab/Color_Similarity_GWOT.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

G.K. and M.O. were supported by JST Moonshot R&D grant number JPMJMS2012. G.K. was supported by the Ezoe Memorial Recruit Foundation. N.T. and M.O. were supported by Japan Promotion Science, Grant-in-Aid for Transformative Research Areas grant numbers 20H05710 and 23H04830 (N.T.) and 20H05712 and 23H04834 (M.O.). N.T. was supported by Australian Research Council (DP180104128 and DP180100396). N.T. and A.Z.-J. were supported by National Health and Medical Research Council (APP1183280) and Foundational Question Institute (FQXi-RFP-CPW-2017) and Fetzer Franklin Fund, a donor advised fund of Silicon Valley Community Foundation. We thank Dominik Kirsten-Parsch and Lonni Gomes for their help in collecting the color dissimilarity data.

AUTHOR CONTRIBUTIONS

A.Z.-J. and N.T. designed and performed experiments to collect behavioral data. G.K. and M.O. conceived the initial idea of the data analysis. G.K., K.T., and M.O. analyzed the data. G.K., A.Z.-J., N.T., and M.O. wrote the initial draft of the manuscript. K.T. joined the project later, contributing extensively to the data analysis and revisions of the manuscript. All authors reviewed, edited, and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this work, the author(s) used GPT-4o and Claude 3.5 Sonnet in order to check English expressions. After using this tool or service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Ethics
  - Participants
  - Exclusion - color typical
  - Exclusion-color atypical
- METHOD DETAILS
  - Display apparatus
  - Procedure
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Estimation of embeddings and dissimilarity matrices at the group level
  - Unsupervised alignment using Gromov-Wasserstein distance
  - Supervised alignment
  - Unsupervised alignment
  - Gromov-Wasserstein Optimal Transport
  - Hyperparameter tuning
  - Initialization of transportation plan
  - Evaluation of unsupervised alignment

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112029.

Received: August 5, 2024
Revised: November 4, 2024
Accepted: February 11, 2025
Published: February 15, 2025

REFERENCES

1. Dennett, D.C. (1988). Quining qualia. In Consciousness in Modern Science, A.J. Marcel and E. Bisiach, eds. (Oxford University Press).
2. Nagel, T. (1974). What is it like to be a bat? Phil. Rev. *83*, 435–450.
3. Chalmers, D.J. (1996). The Conscious Mind: In Search of a Fundamental Theory (Oxford University Press).
4. Rosenthal, D. (2015). Quality spaces and sensory modalities. In Phenomenal Qualities: Sense, Perception, and Consciousness, P. Coates and S. Coleman, eds. (Oxford University Press).
5. Kleiner, J. (2020). Mathematical models of consciousness. Entropy *22*, 609.
6. Lee, A.Y. (2021). Modeling mental qualities. Philos. Rev. *130*, 263–298.
7. Fink, S.B., Kob, L., and Lyre, H. (2021). A structural constraint on neural correlates of consciousness. PhiMiSci *2*, 2–23.
8. Lau, H., Michel, M., LeDoux, J.E., and Fleming, S.M. (2022). The mnemonic basis of subjective experience. Nat. Rev. Psychol. *1*, 479–488.
9. Lyre, H. (2022). Neurophenomenal structuralism. a philosophical agenda for a structuralist neuroscience of consciousness. Neurosci. Conscious. *2022*, niac012.
10. Tallon-Baudry, C. (2022). The topological space of subjective experience. Trends Cogn. Sci. *26*, 1068–1069.
11. Malach, R. (2021). Local neuronal relational structures underlying the contents of human conscious experience. Neurosci. Conscious. *2021*, niab028.
12. Tsuchiya, N., and Saigo, H. (2021). A relational approach to consciousness: categories of level and contents of consciousness. Neurosci. Conscious. *2021*, niab034.
13. Kleiner, J. (2024). Towards a structural turn in consciousness science. Conscious. Cogn. *119*, 103653.
14. Helm, C.E. (1964). Multidimensional ratio scaling analysis of perceived color relations. J. Opt. Soc. Am. *54*, 256–262.
15. Tversky, A. (1977). Features of similarity. Psychol. Rev. *84*, 327–352.
16. Shepard, R.N., and Cooper, L.A. (1992). Representation of colors in the blind, color-blind, and normally sighted. Psychol. Sci. *3*, 97–104.
17. Epping, G.P., Fisher, E.L., Zeleznikow-Johnston, A.M., Pothos, E.M., and Tsuchiya, N. (2023). A quantum geometric framework for modeling color similarity judgments. Cogn. Sci. *47*, e13231.
18. Zeleznikow-Johnston, A., Aizawa, Y., Yamada, M., and Tsuchiya, N. (2023). Are color experiences the same across the visual field? J. Cogn. Neurosci. *35*, 509–542.
19. Indow, T. (1988). Multidimensional studies of munsell color solid. Psychol. Rev. *95*, 456–470.
20. Paramei, G.V., Bimler, D.L., and Cavonius, C.R. (2001). Color-vision variations represented in an individual-difference vector chart. Color Res. Appl. *26*, S230–S234.
21. Boehm, A.E., MacLeod, D.I.A., and Bosten, J.M. (2014). Compensation for red-green contrast loss in anomalous trichromats. J. Vis. *14*, 19.
22. Chalmers, D.J. (2002). Consciousness and its place in nature. In Philosophy of Mind: Classical and Contemporary Readings, D.J. Chalmers, ed. (Oxford University Press), pp. 102–142.
23. Lee, A.Y., and Fink, S.B. (2023). Structuralism in the Science of Consciousness – Editorial Introduction (Preprint at OSF). https://doi.org/10.31234/osf.io/dwvby.
24. Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. *17*, 401–412.
25. Rosenthal, I.A., Singh, S.R., Hermann, K.L., Pantazis, D., and Conway, B.R. (2021). Color space geometry uncovered with magnetoencephalography. Curr. Biol. *31*, 1127–1128.
26. Mémoli, F. (2011). Gromov-Wasserstein distances and the metric approach to object matching. Found. Comut. Math. *11*, 417–487.
27. 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, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, eds. (Association for Computational Linguistics), pp. 1881–1890.
28. 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.
29. Thual, A., Tran, H., Zemskova, T., Courty, N., Flamary, R., Dehaene, S., and Thirion, B. (2022). Aligning individual brains with fused unbalanced gromov-wasserstein. Preprint at arXiv. https://doi.org/10.48550/arXiv.2206.09398.
30. Peyré, G., and Cuturi, M. (2016). In Solomon J.Gromov-wasserstein averaging of kernel and distance matrices, M.F. Balcan and K.Q. Weinberger, eds. (Proceedings of ICML. JMLR.org), pp. 2664–2672.
31. Peyré, G., and Cuturi, M. (2019). Computational optimal transport: With applications to data science. FNT. in Machine Learning *11*, 355–607.
32. Sasaki, M., Takeda, K., Abe, K., and Oizumi, M. (2023). Toolbox for Gromov-Wasserstein Optimal Transport: Application to Unsupervised Alignment in Neuroscience (bioRxiv).
33. Palmer, S.E. (1999). Color, consciousness, and the isomorphism constraint. Behav. Brain Sci. *22*, 923–989.

34. Block, N. (2007). Wittgenstein and qualia. Philos. Perspect. *21*, 73–115.
35. Lee, G. (2006). The experience of left and right. In Perceptual Experience, T.S. Gendler and J. Hawthorne, eds. (Oxford University Press), pp. 291–315.
36. Griffin, L.D. (2001). Similarity of psychological and physical colour space shown by symmetry analysis. Color Res. Appl. *26*, 151–157.
37. 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.
38. Hebart, M.N., Zheng, C.Y., Pereira, F., and Baker, C.I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat. Hum. Behav. *4*, 1173–1185.
39. Hebart, M.N., Contier, O., Teichmann, L., Rockter, A.H., Zheng, C.Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., and Baker, C.I. (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. Elife *12*, e82580.
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. ICLR 2024 Workshop on Representational Alignment *1*, 1–12.
41. Cowen, A.S., and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc. Natl. Acad. Sci. USA *114*, e7900–e7909.
42. Nummenmaa, L., Hari, R., Hietanen, J.K., and Glerean, E. (2018). Maps of subjective feelings. Proc. Natl. Acad. Sci. USA *115*, 9198–9203.
43. 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.
44. Roads, B.D., and Love, B.C. (2020). Learning as the unsupervised alignment of conceptual systems. Nat. Mach. Intell. *2*, 76–82.
45. Gärdenfors, P. (2000). Conceptual Spaces: The Geometry of Thought (The MIT Press).
46. Saji, N., Imai, M., and Asano, M. (2020). Acquisition of the meaning of the word orange requires understanding of the meanings of red, pink, and purple: constructing a lexicon as a connected system. Cogn. Sci. *44*, 12813.
47. Birch, J. (1997). Efficiency of the ishihara test for identifying red-green colour deficiency. Ophthalmic Physiol. Opt. *17*, 403–408.
48. Pouw, A., Karanjia, R., and Sadun, A. (2017). A method for identifying color vision deficiency malingering. Graefe's Arch. Clin. Exp. Ophthalmol. *255*, 613–618.
49. Bosten, J. (2019). The known unknowns of anomalous trichromacy. Current Opinion in Behavioral Sciences *30*, 228–237.
50. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. *32*, 1–12.
51. Gower, J.C., and Dijksterhuis, G.B. (2004). Procrustes Problems (Oxford University Press).
52. Alaux, J., Grave, E., Cuturi, M., and Joulin, A. (2019). Unsupervised hyperalignment for multilingual word embeddings. Proceedings of ICLR *1*, 1–12.
53. Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Association for Computing Machinery), pp. 2623–2631.
54. Flamary, R., Courty, N., Gramfort, A., Alaya, M.Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., and Gautheron, L. (2021). Pot: Python optimal transport. J. Mach. Learn. Res. *22*, 1–8.
55. Bergstra, J., Yamins, D., and Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Proceedings of the 30th International Conference on Machine Learning, S. Dasgupta and D. McAllester, eds. (JMLR.org), pp. 115–123.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Behavioral experiment data	This paper	Open Science Framework: https://osf.io/9xwr2/
Software and algorithms		
Behavioral experiment code	This paper	Open Science Framework: https://osf.io/9xwr2/
Data analysis code	This paper	GitHub: https://github.com/oizumi-lab/Color_Similarity_GWOT

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethics

Experimental procedures were approved by the Monash University Human Research Ethics Committee (Project ID: 17674). Participants were provided electronically with written consent forms prior to the commencement of the experiment and provided electronic consent to participate. Participants were compensated for their time at a rate of £5.27 for an experimental duration of approximately 40 minutes.

Participants

Participants were recruited remotely through Prolific, an online participant recruitment platform. Participants accessed the experiment and provided data using their own personal computers. Only English native speakers were recruited. We recruited 488 general-population (color-neurotypical; (Group T)) and 548 self-identified color-atypical (Group A) participants prior to data cleaning.

Exclusion - color typical

Participants who failed to meet the inclusion criteria were excluded from the analysis. Firstly, we removed participants who failed to complete the experiment. Secondly, we excluded participants with a catch score of <77%. Catch scores from trials that were included to assess participant attention, which were scattered randomly among the main trials. Catch trials involved no presentation of colored stimulus patches. Instead, participants were shown a response screen where they were prompted to click a specific number. All other aspects of the response screen were the same. Lastly, the experiment was designed as a 'double-pass paradigm', meaning participants performed each sequence of main trials twice. Participants whose responses across the two passes were correlated <0.5 were excluded, as low 'double-pass' correlation is indicative of inattentive or neglectful responding.[17,18] 62 out of 488 color-neurotypical participants were excluded, leaving 426 (87%) for the main analysis.

Exclusion-color atypical

We collected a cohort of 548 participants who self-identified as color blind. In addition to the general exclusion criteria, these participants were also screened using a modified online Ishihara test. Participants viewed a set of 28 Ishihara color plates and were asked to report the number they observed. 16 of the plates were standard, so that identification of the number requires color-typical vision. We excluded participants if their identification performance was > 80%, as we suspected they falsely identified themselves as red-green color blind.[47] Additionaly, 12 plates were red- or blue- shifted, so that identification of the number did not require color-typical vision.[48] We excluded participants if their identification performance was 80%, as we suspected they intentionally tried to mimic behaviors of color-atypical people. After these additional exclusion criteria, 257 of 548 (47%) participants who self-identified as color blind were used for the main analysis (Figure S1). As the vast majority of individuals with color vision deficiencies possess some form of red-green color blindness (such as deuteranomaly or protanomaly), and because the Ishihara plates we selected did not screen for blue-yellow color blindness (such as tritanomaly or tritanopia), it is probable that the included participants overwhelmingly possess some degree of red-green colour blindness.[49]

METHOD DETAILS

Display apparatus

Due to the nature of online experimentation, participants used their own computer screen to perform the experiment. The stimuli for the current study were based on the color swatches used by.[46] This 93-color set was selected by[46] from the Practical color

Co-ordinate System (PCCS). All stimuli were presented as solid colored circles 120 pixels in diameter on a grey (#7F7F7F) background.

Procedure

After recruitment through Prolific, participants were directed to the experiment hosted on Pavlovia. The first page of the experiment was a consent form that they could electronically sign by pressing the spacebar. Participants were informed that the data collection process was anonymous and that they could quit the experiment at any time. Following consent, participants were provided written instructions on how to complete the experiment. This was followed by 9 practice trials, seven of which were color similarity judgments and the rest were catch trials.

Main trials for color-neurotypical participants proceeded as follows. First, a fixation cross was presented in the centre of the screen for 250 ms. Following this, the two stimuli were presented as solid-colored circles for 250 ms. A duration of 250 ms was used, rather than leaving colors on the screen for longer durations, because the duration of direct eye fixation of stimuli can vary from trial to trial and from participant to participant under unconstrained conditions. As we demonstrated in a previous study,[18] reliable color similarity judgments are easily feasible with these short durations. Considering the centre of the screen as the midpoint, each stimulus was presented 180° apart and at a radius of 8% of the width of their screen. The stimuli were randomly assigned to a position within ±30° of horizontal meridian in order to prevent retinal adaptation between trials. Lastly, the participants were presented with a response screen and were directed to select a specified value from 0 (most similar) to 7 (most dissimilar). After responding, participants were asked to click on the centre of the screen to initiate the next trial.

A typical participants were presented with a slightly updated version of the same task. Instead of stimuli being presented randomly within ±30° of horizontal meridian, they were presented randomly in two out of four possible locations equidistant from the centre of the screen and maximally spaced from each other. Additionally, participants reported using values from -4 to $+4$ (with zero excluded) instead of 0 to 7. All other parameters remained the same.

During practice trials, participants were provided feedback on what selection they made, consisting of both the value they selected and the text 'Very Similar', 'Similar', 'Different' or 'Very Different' for selections of 0/1, 2/3, 4/5, 6/7 respectively for the color-neurotypical participants, or $-4/-3$, $-2/-1$, 1/2, 3/4 for the color-atypical participants. At the cessation of these practice trials they were asked to press the SPACE button to proceed to the main trial set.

Following the practice trials, participants completed the main task. As with the practice trials, catch trials were randomly inserted among the main trials. Each participant was randomly allocated a set of color pairs out of the total 4371 unique pairs of 93 colors (including pairs of the same color), which were presented in a random sequence. Color-neurotypical participants were allocated 162 color pairs. After providing a response for each color pair once, color-neurotypical participants performed a repeat of the first 162 trials, identical in stimuli and sequence (double-pass). In total, this comprised of 324 main trials and 20 randomly interspersed catch trials. Color-atypical participants were allocated 81 color pairs, which were also presented in a double pass manner for a total of 162 main trials and 10 catch trials.

QUANTIFICATION AND STATISTICAL ANALYSIS

Estimation of embeddings and dissimilarity matrices at the group level

Aggregating similarity judgments

To estimate embeddings at the group level, we aggregated similarity ratings from multiple participants. We fixed the number of similarity ratings taken from each participant to 75, which corresponds to the minimum number of unique color pairs judged among all color-neurotypical and color-atypical participants. We randomly chose 75 similarity ratings without replacement from each participant. Then, we aggregated the similarity responses from the fixed number of participants Z and made a group of participants. The participants were chosen randomly from the entire participants (426 for color-typical participants, 257 for color-atypical participants).

To assess how many trials of similarity judgments are needed to reliably determine whether two similarity structures are aligned in an unsupervised manner, we varied the number of participants in a group, $Z = 16, 32, 64, 128$. As we can see in Figure S2, we found that $Z = 128$ (9600 trials) is necessary to obtain an accuracy of unsupervised alignment that is unquestionably higher than the chance level for any random samples. Based on this analysis, we only showed the results of the alignment when $Z = 128$ in the main text. See Figure S2 for the other cases.

Estimation of embeddings based on the similarity judgments

Based on the aggregated similarity judgment responses, we estimated the embeddings of 93 colors. The embeddings are estimated by training a one-layer linear neural network model with the similarity judgment data by using PyTorch.[50] The procedure is as follows.

First, the initial embedding of each color denoted by $\boldsymbol{e}_i$ is given by a one-hot vector of 93 dimensions. Second, the initial embeddings are linearly transformed into 20 dimensional embeddings as

$$\boldsymbol{x}_i = W\boldsymbol{e}_i, \qquad \text{(Equation 1)}$$

where $\boldsymbol{x}_i$ is the embedding of the i-th color, W is the weights of the neural network that need to be learned so that the loss function defined below is minimized. We set the embeddings of dimensions large enough to capture the similarity structure of 93 colors. Note that to avoid over-fitting, the number of dimensions set here is effectively reduced by the hyperparameter tuning of the L1 regularization term through the usual cross-validation procedure.

The similarity ratings between the pair of colors are given by the Euclidean distance

$$D_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2. \quad \text{(Equation 2)}$$

Then, by using the empirically obtained similarity rating S_{ij} for the color pair i and j, the loss function to minimize is defined as

$$L_{\text{train}} = \sum_{(i,j)}^{n_{\text{train}}} \| D_{ij} - S_{ij}\|^2 + \lambda \sum_{i=1}^{m} \| \boldsymbol{x}_i\|_1. \quad \text{(Equation 3)}$$

where the summation of the first term is taken over all the color pairs (i,j) in the training dataset, n_{train} is the total number of color combinations in the training dataset, m is the number of the colors, $\|\cdot\|_1$ denotes the L1 norm $\|z\|_1 = \sum_i |z_i|$, and λ is a hyperparameter that determines the strength of the $L1$ regularization. The loss function was optimized by the Adam algorithm with a fixed number of 100 epochs using PyTorch. The hyperparameter λ was optimized by 5-fold cross-validation to minimize the following validation error,

$$L_{\text{valid}} = \sum_{(i,j)}^{n_{\text{valid}}} \| D_{ij} - S_{ij}\|^2, \quad \text{(Equation 4)}$$

where n_{valid} is the total number of color combinations in the validation dataset.

Unsupervised alignment using Gromov-Wasserstein distance

In this section, we provide an overview of unsupervised alignment methods for aligning two qualia structures (two sets of embeddings) by using Gromov-Wasserstein optimal transport. With this method, we can quantify the degree of similarity between the qualia structures. Also, the results from this analysis can inform us in what way those are similar or different, which can be examined by detailed analysis of the correspondence between the embeddings of the two qualia structures.

General problem setting

We consider the problem of aligning two sets of embeddings X and Y, which in our case correspond to the embeddings of the color qualia structures. X and Y are $d \times n$ matrices where n is the number of embeddings and d is the dimension of embedding vectors.

$$X = (\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_n), Y = (\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_n) \quad \text{(Equation 5)}$$

Here, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are column vectors, which are the embeddings of the ith-color quale of X and Y, respectively.

The general problem setting in this study is to find the optimal alignment between X and Y without assuming any correspondence by solving the following problem:

$$\min_P \min_Q \| X - QYP \|_F^2, \quad \text{(Equation 6)}$$

where $\|\cdot\|_F$ is the Frobenius norm $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, P is the $n \times n$ assignment matrix that establishes correspondence between the column vectors of X and those of Y (i.e., $\boldsymbol{x}_j \leftarrow \sum_i P_{ij} \boldsymbol{y}_i$), and Q is the $d \times d$ orthogonal matrix that rotates Y to fit into X. If we only allow one element in each column of P to be 1 and set the other elements to 0, the problem becomes finding a one-to-one correspondence between the columns of X and Y, or equivalently, finding the optimal permutation of the column indexes of X. In this study, we examine a more general scenario where the elements of matrix P can take on a real number between 0 and 1. These values represent the degree of correspondence between the i-th column of matrix X and the j-th column of matrix Y. This more flexible approach allows us to model the correspondences between the columns of X and Y in a more comprehensive manner.

Supervised alignment

When the assignment matrix P is given, the optimization problem becomes the well-known Procrustes problem,[51] which has a closed form solution. For instance, if we simply assume that the column indexes of X match those of Y, and therefore P is the identity matrix, the optimization problem is given by

$$\min_Q \| X - QY \|_F^2. \quad \text{(Equation 7)}$$

Given the singular value decomposition $U\Sigma V^\top$ of $XY^\top$, the solution to the Procrustes problem is given by $Q^* = UV^\top$.

Unsupervised alignment

In this study, we consider the scenario where the assignment matrix P is not given. In this case, we need to jointly optimize P and Q in Equation 6, which is a non-convex optimization problem without a closed-form solution. To address this, we first find an optimal assignment matrix P using Gromov-Wasserstein optimal transport (GWOT) in an unsupervised manner. We then compute the Procrustes solution Q based on the assignment matrix obtained from the GWOT analysis. This approach has been effective in unsupervised language translation tasks.[27,52] Denoting the optimal transportation plan (the assignment matrix) by Γ^*, the problem to solve becomes

$$\min_{Q} \|X - QY\Gamma^*\|_F^2 . \quad \text{(Equation 8)}$$

The solution can be found by the singular value decomposition of $X(Y\Gamma^*)^\top$.

Gromov-Wasserstein Optimal Transport

To obtain the assignment matrix P, which establishes the correspondence between the embeddings (the column vectors) of X with the embeddings of Y, we use Gromov-Wasserstein optimal transport (GWOT).[26] GWOT is an unsupervised alignment technique that can find correspondence between two point clouds (embeddings) in different domains based on internal distances within each domain. Unlike classic optimal transport problems, the points in the two domains do not necessarily reside in the same metric space and any information about correspondences or distances between points "across" different domains is not given. In this study, the internal distances within the domains are represented by two different $n \times n$ dissimilarity matrices D_{ij} and D'_{ij} obtained from different participant groups, where n is the number of colors and D_{ij} denotes the subjective rating of dissimilarity between the i-th and j-th color.

The goal of Gromov-Wasserstein optimal transport problem is to find the optimal way to transport the distribution of resources (e.g., a pile of sand) from one domain to the other. There is a certain amount of the pile on each point in one domain. The distribution of the pile is given by $\boldsymbol{p}$ where p_i is the amount of the pile at the i-th point in the source domain. We wish to transport the piles onto the points in the other domain so that the distribution of the pile matches with the target distribution $\boldsymbol{q}$ where q_i is the amount of the pile at the i-th point in the target domain.

With this setting, we wish to find the optimal transport plan that minimizes a certain transportation cost. The transportation cost considered in GWOT is given by

$$\min_{\Gamma} \sum_{i,j,k,l} (D_{ij} - D'_{kl})^2 \Gamma_{ik} \Gamma_{jl}. \quad \text{(Equation 9)}$$

Note that a transportation plan Γ needs to satisfy the following constraints: $\sum_j \Gamma_{ij} = p_i$, $\sum_i \Gamma_{ij} = q_j$ and $\sum_{ij} \Gamma_{ij} = 1$. Under this constraint, the matrix Γ is considered as a joint probability distribution with the marginal distributions being $\boldsymbol{p}$ and $\boldsymbol{q}$. We set $\boldsymbol{p}$ and $\boldsymbol{q}$ to be the uniform distributions, i.e., $p_i = q_i = 1/n$. Each entry Γ_{ij} describes how much of the pile 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 embedding $\boldsymbol{x}_i$ corresponds to the embeddings $\boldsymbol{y}_j$.

With the transportation plan, the embeddings of Y are mapped to the embeddings of X as follows

$$\boldsymbol{x}_j \leftarrow \sum_{i=1}^{n} \Gamma_{ij} \boldsymbol{y}_i. \quad \text{(Equation 10)}$$

Then, this mapping is subsequently used for finding the rotation matrix Q in Equation 8.

We chose the Gromov-Wasserstein Optimal Transport (GWOT) method for its practical effectiveness in addressing the unsupervised alignment problem. GWOT stands out as one of the few viable approaches for aligning complex similarity structures without predefined correspondences. GWOT's strengths lie in its simplicity, interpretability, and computational efficiency. Unlike more complex deep neural network approaches, GWOT's objective function is relatively straightforward and is rooted in the extension of the well-established Wasserstein distance. This allows for the computation of both the Gromov-Wasserstein distance, which measures the similarity between internal relational structures, and the optimal transportation plan that maps items between domains. Moreover, GWOT maintains computational tractability, with reduced complexity making it feasible to handle large datasets, which is a critical consideration in applications involving large structural data. This combination of factors makes GWOT a practical and effective choice for our alignment needs, while still leaving room for potential improvements through future advancements in the field.

Hyperparameter tuning

Previously, it has been demonstrated that adding an entropy-regularization term can improve the computational efficiency and help to find good local optimums of the Gromov-Wasserstein optimal transport problem.[30,31]

$$\min_{\Gamma} \sum_{i,j,k,l} (D_{ij} - D'_{kl})^2 \Gamma_{ik} \Gamma_{jl} - \epsilon H(\Gamma), \quad \text{(Equation 11)}$$

where $H(\Gamma)$ is the entropy of a transportation plan Γ and ϵ is a hyperparameter that determines the strength of the entropy regularization.

To find good local optimums, we conducted hyperparameter tuning on ϵ in Equation 11 by using the GWTune toolbox that we developed.[32] This toolbox uses Optuna[53] for hyperparameter tuning and Python Optimal Transport (POT)[54] for GWOT optimization. We sampled 200 different values of ϵ ranging from 0.02 to 0.2 by a Bayesian sampler called TPE (Tree-structured Parzen Estimator) sampler.[55] We chose the value of ϵ, where the optimal transportation plan minimizes the Gromov-Wasserstein distance without the entropy-regularization term (Equation 9) following the procedure proposed by a previous study.[28]

Initialization of transportation plan

To avoid getting stuck in bad local minima, it is effective to randomly initialize the transportation plan and try many random initialization, as proposed in.[32] Each element in the initial transportation plan was sampled from the uniform distribution [0,1] and was normalized to satisfy the following conditions: $\sum_j \Gamma_{ij} = p_i$, $\sum_i \Gamma_{ij} = q_j$ and $\sum_{ij} \Gamma_{ij} = 1$. For each value of ϵ, the transportation plan was randomly initialized.

Evaluation of unsupervised alignment

To assess the degree of similarity between the two qualia structures in the unsupervised setting, the correct matching rate of color labels are computed between two groups based on the optimal transportation plan. Denote the color labels in group 1 and 2 as c_1 and c_2 respectively. The matching rate is calculated by comparing the transportation plan Γ with these color labels. For each color i in group 1, denoted by c_{1i}, the matching condition can be formalized as:

$$\text{Match}(i) = \begin{cases} 1, & \text{if } \Gamma_{ij} = \max_{j \in \{1,\dots,n\}}(\Gamma_{ij}) \text{ and } c_{1i} = c_{2j} \\ 0, & \text{otherwise} \end{cases} \quad \text{(Equation 12)}$$

This function indicates whether the i-th color in group 1, c_{1i}, matches with the same color in group 2, c_{2j}. The matching rate is then the percentage of colors in group 1 that match with the same colors in group 2, which can be calculated as

$$\text{Matching Rate} = \frac{\sum_{i=1}^{n} \text{Match}(i)}{n}, \quad \text{(Equation 13)}$$

where n is the total number of colors ($n = 93$). In this study, the row and column of Γ are sorted in the same order of colors and thus, the matching rate corresponds to the percentage of the diagonal elements Γ_{ii} that are the largest among Γ_{ij} for any j.

The matching rate defined above is top 1 matching rate. More generally, we also define top k matching rate. For each color i in group 1, we can define a function to determine if the probability of the i-th color corresponding to the same color in group 2 is within the top-k probabilities:

$$\text{Top}_k(i) = \begin{cases} 1, & \text{if } \Gamma_{ij} \text{ is in the top-}k \text{ for } j \in \{1,\dots,n\} \text{ and } c_{1i} = c_{2j} \\ 0, & \text{otherwise} \end{cases} \quad \text{(Equation 14)}$$

The top-k matching rate can then be calculated as

$$\text{Top-}k \text{ Matching Rate} = \frac{\sum_{i=1}^{n} \text{Top}_k(i)}{n}. \quad \text{(Equation 15)}$$

A high matching rate between two color similarity matrices suggests that two different groups have similar similarity structures of colors.

The chance-level matching rate was estimated by simulation as follows. First, 10,000 random matrices satisfying the constraints of a transportation plan were generated. For each matrix, the matching rate was computed, yielding a null distribution of matching rates. Finally, the 2.5th–97.5th percentiles of this distribution were taken as the 95% percentile interval for the chance-level matching rate.