Is my "red" your "red"?: Unsupervised alignment of qualia structures via optimal transport

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

[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, Australia.
[4]Turner Institute for Brain and Mental Health, Monash University, Melbourne, Australia.
[5]Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Osaka, Japan.
[6] Advanced Telecommunications Research Computational Neuroscience Laboratories, Kyoto, Japan.

*Corresponding author(s). E-mail(s): naotsugu.tsuchiya@monash.edu; c-oizumi@g.ecc.u-tokyo.ac.jp;
Contributing authors: g.kawakita22@imperial.ac.uk; ariel.zeleznikow-johnston@monash.edu; tkkentakeda1248@g.ecc.u-tokyo.ac.jp;
†These authors contributed equally to this work.
‡These authors contributed equally to this work.

Abstract

To what extent are sensory experiences equivalent between individuals? One promising approach to this fundamental question about consciousness is to intersubjectively compare the similarity relationships of sensory

experiences, named "qualia structures". Conventional methods for comparing the similarity relationships largely sidestep the issue by assuming experiences evoked by the same stimuli are matched across individuals, precluding the possibility that my "red" might be your "blue". Here, we present an unsupervised optimal transport method for assessing the similarity of qualia structures without assuming correspondence between individuals. As proof of concept, we analyzed two massive datasets: dissimilarity judgments for 93 colors and 1854 natural objects. In both cases, we found that qualia structures can be "correctly" aligned across participants based solely on similarity relationships, providing quantitative evidence for the structural equivalence of qualia of color and natural objects across individuals. This method is applicable to any modality of experience, enabling general structural exploration of subjective experiences.

Keywords: qualia, qualia structure, consciousness, unsupervised alignment, optimal transport, Gromov-Wasserstein distance

The question of inter-subjective equivalence of sensory experience is a central concern in the study of consciousness. Some researchers consider the question impossible to determine due to the intrinsic, ineffable and private nature of subjective experience [1]. Though direct description of our experiences for inter-subjective comparison may be impossible, indirect characterization of experience is empirically possible and considered as a promising research program [2–12]. One particular approach is to analyse reports of subjective similarities between sensory experiences [13–17]. Relationships between sensory experiences, such as similarity, allow structural investigation of phenomenal consciousness.

Based on this idea, we formally introduce a new paradigm, which we call "qualia structure" (Fig. 1a). This paradigm consists of two main steps. The first involves collecting detailed subjective reports about the relationships between sensory experiences (qualia) through psychophysics experiments that capture the relational structure of the qualia [17]. The second involves comparing qualia structures between participants without assuming correspondence between individual qualia, and evaluating the extent to which the qualia structures are similar.

To demonstrate the utility of this paradigm, we investigate two distinct datasets in this study. The first dataset consists of color similarity judgments, which is original data collected for this study. The second dataset, known as THINGS data, includes similarity judgments of naturalistic objects previously collected and made publicly available by other researchers [18–20]. We will initially illustrate the qualia structure paradigm mainly using the color similarity judgment data as an example and then, show the analysis of the color similarity judgments. Subsequently, we will apply the same paradigm to the THINGS data to further demonstrate our paradigm's utility and generalizability.

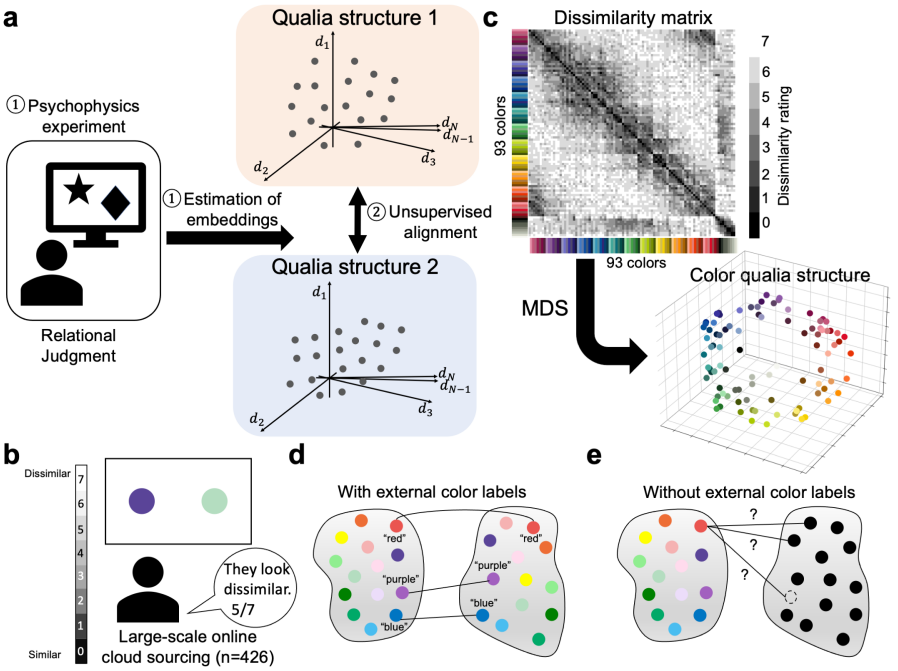

Fig. 1 Schematics of concepts in the qualia structure paradigm. (a) Two steps in the qualia structure paradigm. The first step is to collect subjective reports through relational judgments between stimuli that enable estimation of the relational structure of sensory experiences, i.e., qualia structure. The second step is to align qualia structures from different individuals in an unsupervised manner to quantify the degree of similarity of the qualia structures. (b) A color dissimilarity judgment task. A dissimilarity rating between a pair of colors is collected. (c) A color dissimilarity matrix and its embedding. An element in the dissimilarity matrix represents the average dissimilarity rating between a pair of colors reported by the participants. From a dissimilarity matrix, the embeddings of colors are estimated using multidimensional scaling (MDS). (d) Supervised alignment of color qualia structures, which assumes correspondence between qualia evoked by the same external stimuli across different individuals. (e) 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.

For the first step of our qualia structure paradigm, we need to estimate the similarity relationships between qualia. In the case of colors, we directly asked participants to report the (dis)similarity between color pairs from a set of 93 unique colors. These ratings defined the similarity relationships between the 93 colors for the subsequent analyses. In contrast, in the case of natural objects, participants provided odd-one-out judgment among triplet of objects without directly providing pair-wise similarity ratings between object pairs. Thus, we first need to estimate the embeddings of the natural objects from the odd-one-out judgments and then, estimate the pairwise similarity relationships between the objects based on these embeddings.

The similarity relationships were represented in a matrix D, where the rows and columns correspond to different experiences. For example, the entry, D_{ij}, shows the subjective dissimilarity between the two experiences of i-th and j-th colors (Fig. 1c). We applied multidimensional scaling (MDS) [21] to obtain the vector embeddings of the colors as an approximation of the qualia structure (Fig. 1c). In the case of natural objects, their vector embeddings are first estimated and then a dissimilarity matrix D is computed as the "distance" between the embeddings of the objects.

For the second step, we compare qualia structures in a purely unsupervised manner, without assuming any correspondence between individual qualia across participants. One might be tempted to compare two dissimilarity matrices assuming stimulus-level "external" correspondence: my "red" corresponds to your "red"(Fig. 1d). 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 [22]. However, there is no guarantee that the same stimulus will necessarily evoke the same subjective experience across different participants. Accordingly, when considering which stimuli evoke which qualia for different individuals, we need to consider all possibilities of correspondence: my "red" might correspond to your "red", "green", "purple", or might lie somewhere between your "orange" and "pink"(Fig. 1e).

To account for all possible correspondences, we use an unsupervised alignment method for quantifying the degree of similarity between qualia structures. As shown in Fig. 2a, in unsupervised alignment, we do not attach any external (stimuli) labels to the qualia embeddings. Instead, we try to find the best matching between qualia structures based only on their internal relationships (see Methods). After finding the optimal alignment, we can use external labels, such as the identity of a color stimulus (Fig. 2b), 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 or which do not. Checking the assumption that these external labels are consistent across individuals allows us to assess the plausibility of determining accurate inter-individual correspondences between qualia structures of different participants.

To this end, we used the Gromov-Wasserstein optimal transport (GWOT) method [23], which has been applied with great success in various fields [24–28]). GWOT aims to find the optimal mapping between two point clouds in different domains based on the distance between points within each domain. 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 another point in the other domain that has a similar relationship to other points.

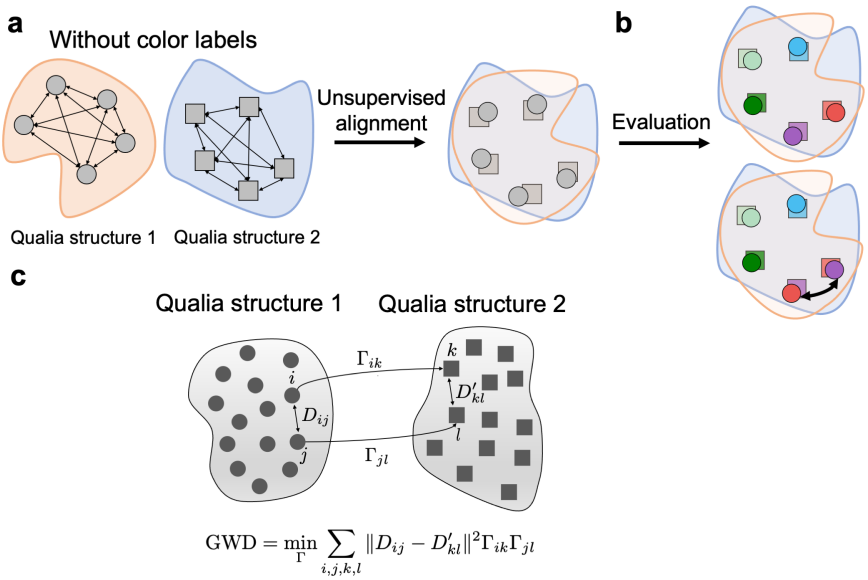

Fig. 2 Schematic of unsupervised alignment. (a) Aligning qualia structures of different individuals in an unsupervised manner without any external labels, solely based on the internal relationships of the embeddings. (b) Evaluation of unsupervised alignment using external labels. (c) Schematic of Gromov-Wasserstein 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.

The goal of GWOT is to find the optimal mapping Γ that minimizes the Gromov-Wasserstein distance (GWD), defined as

$$\mathrm{GWD} = \min_{\Gamma} \sum_{i,j,k,l} (D_{ij} - D'_{kl})^2 \Gamma_{ik}\Gamma_{jl}, \tag{1}$$

where D and D' are distance matrices in the two domains (Fig. 2c). The elements of the matrix Γ_{ik} can be interpreted as the "probability" of the i-th point in one domain corresponding to the k-th point in the other domain (see Methods). This cost function encourages the mapping of pairs of points i and j in one domain to pairs of points k and l in the other domain if the distances between the points, D_{ij} and D'_{kl}, are similar because the transportation cost represented by $(D_{ij} - D'_{kl})^2$ becomes small.

The optimal mapping, the matrix Γ^*, is obtained by solving this minimization problem (Eq. 1). Adding an entropy-regularization term, $H(\Gamma)$, as shown in Eq. 2, improves the efficiency of the optimization and the empirical results [24, 28]:

$$\mathrm{GWD}_{\epsilon} = \min_{\Gamma} \sum_{i,j,k,l} (D_{ij} - D'_{kl})^2 \Gamma_{ik}\Gamma_{jl} + \epsilon H(\Gamma), \tag{2}$$

where the parameter, ϵ, determines the strength of regularization. We optimized the parameter ϵ by minimizing GWD (Eq. 1) (see Methods).

As a proof of concept, we first applied this method to the similarity data involving 93 colors. The relatively large number of colors enables us to investigate more complex and nuanced qualia structures of colors, which is not possible with previous datasets examining a smaller number of colors [13, 15–17]. Additionally, the large number of colors allows us to more convincingly demonstrate the effectiveness of our method, as a brute-force search method considering all possible correspondences used in previous studies [29] would be infeasible.

To assess whether the color dissimilarity structures from different participants can be aligned in an unsupervised manner, we divided color pair similarity data from a large pool of 426 participants into five participant groups (85 or 86 participants per group) to obtain five independent and complete sets of pairwise dissimilarity ratings for 93 color stimuli (Fig. 3a). Each participant provided a pairwise dissimilarity judgment for a randomly allocated subset of the 4371 possible color pairs. We computed the mean of all judgments for each color pair in each group, generating five full dissimilarity matrices referred to as Group 1 to Group 5.

We first computed the GWD for all pairs of the dissimilarity matrices of the 5 groups (Group 1-5) using the optimized ϵ. In Fig. 3b, we show the optimized mapping Γ^* between Group 1 and Groups 2-5 (see Supplementary Figure S1 for the other pairs). As shown in Fig. 3b, most of the diagonal elements in Γ^* show high values, indicating that most colors in one group correspond to the same colors in the other groups with high probability.

We next performed unsupervised alignment of the vector embeddings of qualia structures. Although Γ^* provides the rough correspondence between the embeddings of qualia structures, we should find a more precise mathematical mapping between qualia structures in terms of their vector embeddings to more accurately assess the similarity between the qualia structures. Here, we consider aligning the embeddings of all the groups in a common space. We chose Group 1 as a reference group, which the other groups (Group 2 to 5) are aligned to.

By applying MDS, we obtained the 3-dimensional embeddings of Group 1 and Groups 2-5, referred to as X and Y_i, where $i = 2, ..., 5$ (Fig. 3c). We then aligned Y_i to X with the orthogonal rotation matrix Q_i, which was obtained by solving a Procrustes-type problem using the optimized transportation plan Γ^* obtained through GWOT (see Methods). Fig. 3d shows the aligned embeddings of Group 2-5 (Q_iY_i) and the embedding of Group 1 (X) plotted in the embedded space of X. Each color represents the label of a corresponding external color stimulus. Note that even though the color labels are shown in Fig. 3d, this is only for the visualization purpose and the whole alignment procedure is performed in a purely unsupervised manner without relying on the color labels. As can be seen in Fig. 3d, the embeddings of similar colors from

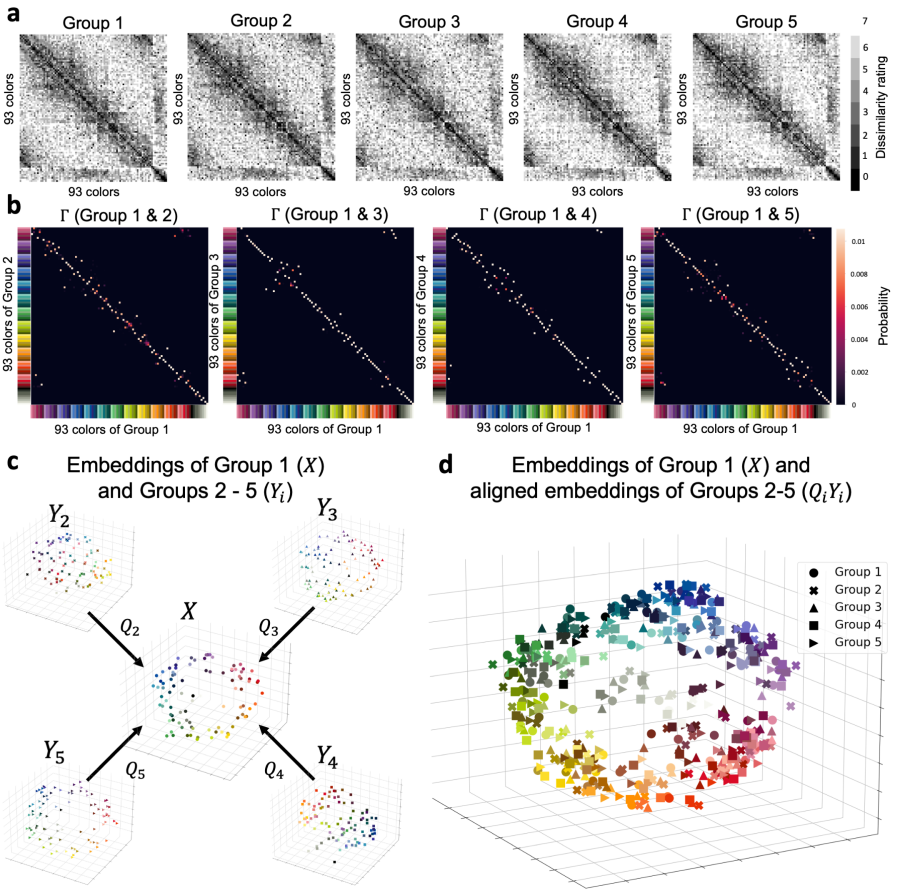

Fig. 3 Unsupervised alignment of qualia structures (a) Dissimilarity matrices of 93 colors in five subgroups (85 or 86 participants per group). (b) Optimized mappings Γ^* between the dissimilarity matrix of Group 1 and those of Groups 2 to 5. (c) The embeddings of Group 1 (X) at the center and those of Groups 2 to 5 (Y_2, Y_3, Y_4, Y_5) at the corners plotted in each group's embedded space. Y_i is aligned to X by the rotation matrix Q_i. Colors represent the labels of external color stimuli, which are not used for the alignment. (d) Aligned embeddings of Group 2 to 5 (Q_iY_i) and embeddings of group 1 (X) plotted in the Group 1's embedded space.

the five groups are located close to each other, indicating that similar colors are 'correctly' aligned by the unsupervised alignment method.

To evaluate the performance of the unsupervised alignment, we computed the k-nearest color matching rate in the aligned space. If the same colors from two groups are within the k-nearest colors in the aligned space, we consider that the colors are correctly matched. We evaluated the matching rates between all the pairs of Groups 1-5. The averaged matching rates are 51% when $k = 1$, 83% when $k = 3$, and 92% when $k = 5$, respectively. This demonstrates the

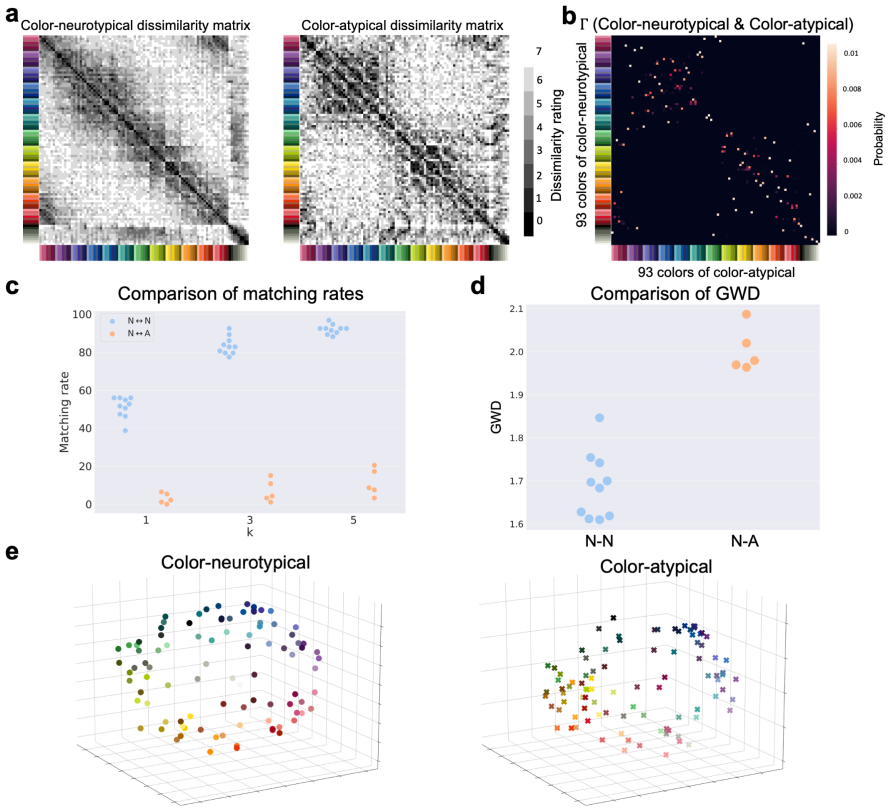

Fig. 4 Unsupervised alignment between qualia structures of color-neurotypical and color-atypical participants (a) Dissimilarity matrix of Group 1 (color-neurotypical participants) and Group 6 (color-atypical participants). (b) Optimized mappings Γ^* between the dissimilarity matrix of Group 1 (color-neurotypical participants) and that of Group 6 (color-atypical participants). (c) Matching rate of unsupervised alignment between the color-neurotypical participants groups (N vs N) and that between the color-neurotypical participants and color-atypical participants (N vs A). Each dot represents the k-nearest matching rate of the alignment between a particular group pair (${}_5C_2 = 10$ dots in total for N vs N and 5 dots in total for N vs A). (d) The values of GWD between the color-neurotypical participants groups (N vs N) and those between the color-neurotypical participants and color-atypical participants (N vs A). (e) Embeddings of Group 1 (X) and aligned embeddings of Group 6 ($Q_6 Y_6$) plotted in the Group 1's embedded space.

effectiveness of the GW alignment for correctly aligning the qualia structures of different participants in an unsupervised manner.

Although we confirmed that we were able to align color qualia structures of color-neurotypical participants fairly well, it is important to investigate whether we can align possibly different color qualia structures of color-atypical participants with those of color-neurotypical participants. For this purpose, we conducted color similarity judgment tasks on color-atypical participants. We obtained 207 participants who both self-reported as color blind and were verified by a modified online Ishihara test (see Methods).

By using the dissimilarity matrix of the color-atypical participants, we performed the GWOT unsupervised alignment with color-neurotypical participants. As shown in Fig. 4a, the dissimilarity matrix of color-neurotypical participants (Group 1) and that of atypical participants (denoted by Group 6) do not look much different. In fact, the Pearson's correlation coefficient between the dissimilarity matrices of Group 1-5 (color-neurotypical) and that of Group 6 (color-atypical) is $\rho = 0.51$ on average, which is only slightly lower than the Pearson's correlation coefficient between the dissimilarity matrices of Group 1-5 ($\rho = 0.60$ on average). However, as can be seen in Fig. 4b, the optimized mapping Γ^* is not lined up diagonally unlike the optimized mappings between color-neurotypical participants groups shown in Fig. 3b (see Supplementary Figure S1 for the other pairs). Accordingly, top k matching rate between Group 1-5 and Group 6 is 3.0% when $k = 1$ (Fig. 4c), which is only slightly above chance ($\sim$ 1%). The matching rate did not improve even when we relaxed the criterion (6.9% and 11% for $k = 3$ and $k = 5$, respectively). Moreover, all of the GWD values between Group 1-5 and Group 6 are larger than any of the GWD values between color-neurotypical participant groups (Fig. 4d). These results indicate that the difference between the qualia structures of neuro-typical and atypical participants is significantly larger than the difference between the qualia structures of neuro-typical participants. Finally, we visualized the embeddings of Group 1 (X) and the aligned embeddings of Group 6 ($Q_6 Y_6$) in Fig. 4e. A notable difference is that greenish colors and reddish colors are close in the embedding space of color atypical participants while they are distant in the embedding space of color neurotypical participants. This structural difference is likely to prevent the unsupervised alignment between the embeddings of color-neurotypical and atypical participants even though the correlation coefficient between the dissimilarity matrices of color neuro-typical and atypical participants is reasonably high.

To further demonstrate the effectiveness and generalizability of this paradigm, we next analyzed the THINGS dataset containing human similarity judgments for 1854 naturalistic objects [18–20]. In this dataset, participants performed an odd-one-out task, where they were presented with three naturalistic objects from the THINGS dataset and asked to report which item in the triplet was the most dissimilar to the other two objects. This dataset includes approximately 4.70 million similarity judgments from about 12,000 participants collected through online crowdsourcing.

Using this dataset, we assessed whether the dissimilarity structures of natural objects from different participants can be aligned in an unsupervised manner, as we did with the color similarity judgment data. We first divided the THINGS similarity judgments data into four non-overlapping participant groups. Each participant group contains $\sim$ 1.17 million similarity judgments, a sample size proven to be sufficient for estimating meaningful and consistent representations of natural objects [19, 20]. We next estimated the embeddings of the 1854 natural objects separately for each of the participant groups (see Methods for the details). We then obtained the dissimilarity matrix of the 1854

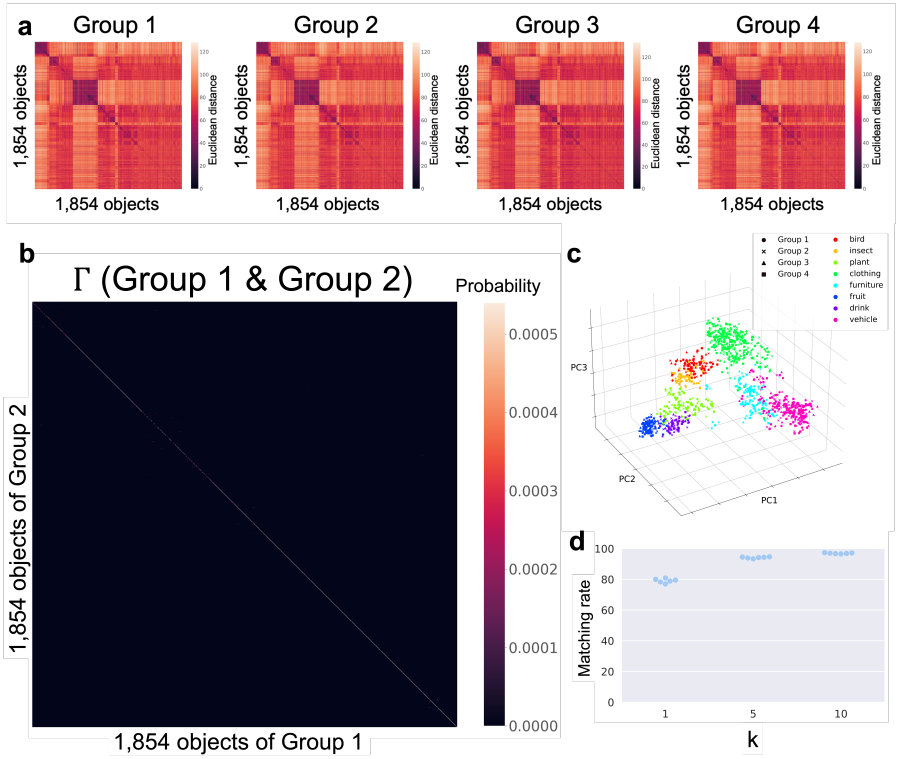

Fig. 5 Unsupervised alignment between qualia structures of naturalistic objects (a) Dissimilarity matrices of 1854 naturalistic objects in four subgroups. (b) Optimized mappings Γ^* between the dissimilarity matrices of Group 1 and Group 2. (c) Embeddings of Group 1 (X) and aligned embeddings of Groups 2-4 ($Q_i Y_i$) plotted in the Group 1's embedded space. Only objects with certain coarse category labels ("bird", "insect", "plant", etc.) are shown here. Each dot represents the embedding of an object. Colors represent the coarse categories of the objects. (d) Matching rates of unsupervised alignment between Groups 1-4. Each dot represents the k-nearest matching rate of the alignment between a pair of groups($_4C_2 = 6$ dots in total).

natural objects for each participant group (Fig. 5a), where the dissimilarity between objects is quantified by the Euclidean distance between the embeddings of the objects. Although these matrices are highly correlated (~ 0.97 on average), this does not trivially guarantee a one-to-one matching of each object between participant groups. This is because, in general, when similarity matrices have some structure induced by coarse categories (as in the THINGS data), the correlation can be high when the objects are matched only at the coarse categorical level (e.g., matching between different birds in the same bird category) but not matched at the one-to-one object level. See Supplementary Fig. S2 for an illustrative example. Thus, even with this level of high correlation, it is not obvious whether the qualia structures of naturalistic objects can be correctly aligned in an unsupervised manner at the one-to-one object level.

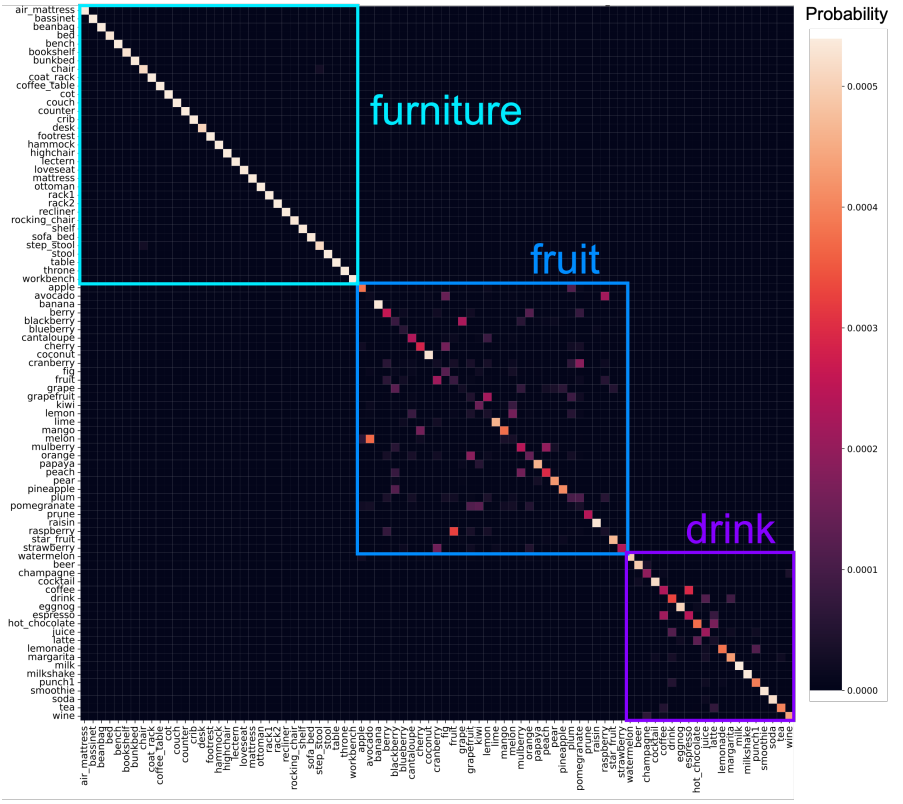

Fig. 6 Enlarged view of optimized mappings Γ^* between the dissimilarity matrices of Group 1 and Group 2. Only objects whose coarse category label is "furniture" or "fruit" or "drink" (85 objects in total out of 1854 objects) are depicted here. Rectangles represent the boundaries of the categories (furniture, fruit, and drink from top to bottom).

We first computed the GWD between all pairs of the dissimilarity matrices of the 4 groups (Group 1-4) using the optimized ϵ. In Fig. 5b, we show the optimized mapping Γ^* between Group 1 and Group 2 (see Supplementary Fig. S3 for the other pairs). As shown in Fig. 5b, most of the diagonal elements in Γ^* show high values, indicating that most objects in one group correspond to the same objects in the other groups with high probability.

Fig. 6 and Supplementary Fig. S4 provide a closer examination of the optimized mapping Γ^* depicted in Fig. 5b. As can been seen in Fig. 6 and Fig. S4, even in cases of mismatch, matching errors tend to occur within the same coarse object category, such as furniture, fruit, and drink. This tendency can be also confirmed in Fig. 5c, where the aligned embeddings of subsets of objects are shown for Groups 1-4 (note that only the first three principal components are shown here for visualization, although the original dimension of the embeddings is 90). The embeddings of objects in the same coarse category from the four groups are located close to each other, indicating that similar or identical objects are 'correctly' aligned by the unsupervised alignment method.

It is important to note that objects are not only matched at the coarse categorical level, but also at the one-to-one object level (Fig. 5b and Fig. 6). As mentioned above, a very high correlation between the dissimilarity matrices (Fig. 5a) does not guarantee a very high one-to-one matching rate in unsupervised alignment. In fact, many matching errors occur between objects in the fruit category, while objects in the furniture category are perfectly matched at the one-to-one object level (Fig. 6).

Finally, we calculated the top k nearest-neighbor matching rates between all the pairs of groups (Fig. 5d). The average matching rates are 79% when $k = 1$, 94% when $k = 5$, and 97% when $k = 10$. Considering that the chance level of the matching rate is only 0.05% when $k = 1$, the high matching rates demonstrate the effectiveness of the GW alignment for correctly aligning the qualia structures of 1854 natural objects in an unsupervised manner.

For a long time, assessing the similarity of subjective experiences across participants has been challenging [30–33]. To address this problem, we proposed the "qualia structure" paradigm, which focuses on quantitative structural comparisons of subjective experiences. Using an unsupervised alignment method, we were able to match the qualia structures of colors and natural objects of different groups of participants based only on the way the qualia relate to each other, without using any external labels.

Our results on color qualia structures are consistent with an idea that the relational properties of color qualia are universally shared by color-neurotypical individuals [33, 34]. Intriguingly, our results also suggest that individuals with color-atypical vision may have a different structure of their color experiences, rather than just failing to experience a certain subset of colors. Longstanding thought experiments that challenge the feasibility of inter-subjective color comparisons, such as individuals with color qualia inversion [30, 31, 33, 35, 36], should be resolvable with our relational unsupervised approach. Beyond traditional measures such as Pearson's correlation coefficient, our method provides a more fundamental 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.

While we focused only on similarities of colors and natural objects, our method has the potential to be applied to a wide range of subjective experiences and different modalities (e.g. emotions [37, 38], semantic concepts[29, 39], etc.). Our approach offers a novel and powerful tool for quantitatively exploring various aspects of subjective experiences and advancing our understanding of consciousness.

Methods

Color similarity judgment data

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.

Design

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 (neurotypical) and 360 self-identified color-atypical (atypical) participants prior to data cleaning.

Exclusion - General

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 (see below) of $< 77\%$. Catch trials were included to ensure participant attention and scattered randomly among the main trials. 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 [16, 17]. 62 out of 488 neurotypical participants were excluded, leaving 426 (87%) for the main analysis.

Exclusion - Color Atypical

We collected a cohort of 360 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 and used as a positive control, with participants excluded if they correctly identified $> 80\%$ of the plates (i.e. made fewer than three errors) [40]. 12 plates consisted of standard Ishihara plates that were red-shifted or blue-shifted so that the number should be correctly identifiable by participants with red-green color deficiencies [41]). These plates were used to detect participants who falsely identified as red-green color blind, with participants excluded if they correctly identified $< 80\%$. After these additional exclusion criteria, 207

of 360 (58%) participants who self-identified as color blind were used for the main analysis.

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 [42]. This 93-color set was selected by [42] 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 neurotypical participants proceeded as follows. First, a fixation cross was presented in the centre of the screen for 250ms. Following this, the two stimuli were presented as solid-colored circles for 250 ms. 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.

Atypical 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.

Catch trials involved no presentation of colored stimuli 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.

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 neurotypical participants, or −4/−3, −2/−1, 1/2, 3/4 for the atypical participants. At the cessation of these practice trials they are 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), which were presented in a random sequence. Neurotypical participants were allocated 162 unique color pairs. After providing a response for each color pair once, 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. Atypical participants were allocated 81 unique color pairs, which were also presented in a double pass manner for a total of 162 main trials and 10 catch trials.

THINGS data: similarity judgments of natural objects

Overview of dataset

The THINGS dataset [20] includes 4.70 million human similarity judgments that were obtained through online crowdsourcing for a collection of 1,854 naturalistic object images. These objects were selected from the THINGS database [18], which offers an extensive inventory of real-world objects encompassing both living and nonliving entities. A single representative image for each object was chosen and used in the experiment.

Experimental procedure

To quantify the similarity between objects, a triplet odd-one-out task was employed. In this task, participants ($N = 12,340$) were presented with three images and were required to determine which object was most dissimilar from the other two. The principal advantage of this task is that the third object provides a contextual framework for the other two objects, thereby highlighting the pertinent dimensions that make the two objects the most similar [19].

Estimation of embeddings and dissimilarity matrices

While the dissimilarity matrices are directly empirically obtained in the case of the color similarity data, we need to estimate the dissimilarity matrices in the case of THINGS data where pair-wise similarities between objects are not directly determined. Following the previous study [19], we first estimate the embeddings of 1854 objects from the odd-one-out similarity judgments so that those embeddings best predicted the human responses for odd-one-out task. Then, we computed the dissimilarity between two objects as the Euclidean distance between the embeddings of the two objects to obtain the dissimilarity matrix of 1854 objects.

The embeddings of 1854 objects were estimated by the following four steps used in the previous study [19, 20]. First, the embeddings of the 1854 objects were initialized with 90 randomly assigned dimensions ranging from 0 to 1. Second, the Euclidean distance between all pairs of the embedding vectors was

computed and considered as the dissimilarity between the embeddings,

$$D_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2, \tag{3}$$

where $\|\cdot\|_2$ is the L2 norm $\|\boldsymbol{z}\|_2 = \sqrt{\sum_i z_i^2}$. Conversely, the similarity between the embeddings was quantified as the negative Euclidean distance,

$$S_{ij} = -D_{ij}. \tag{4}$$

Third, using the similarity between the embeddings, we estimated the probability that a participant chooses image k as an odd object among the triplet (i, j, k), which is equivalent to the probability of choosing image i and j as the most similar object pair among three possible pairs. Here, the probability was estimated by the softmax function of the similarity between the embeddings of the pair (i, j),

$$\Pr[i, j] = \frac{\exp(S_{ij})}{\exp(S_{ij}) + \exp(S_{jk}) + \exp(S_{kl})}, \tag{5}$$

where S_{ij} is given by Eq. 4. Fourth, we updated the embeddings by minimizing the following loss function. For the l-th triplet in the dataset, let $(c_l^{(1)}, c_l^{(2)})$ denote the index pair chosen by a participant as the most similar pair. Then, the loss function is given by

$$L = -\sum_{l=1}^{n_{\text{train}}} \log \Pr[c_l^{(1)}, c_l^{(2)}] + \lambda \sum_{i=1}^{m} \|\boldsymbol{x}_i\|_1, \tag{6}$$

where n_{train} is the total number of triplets in training dataset, m is the number of the natural objects, and $\|\cdot\|_1$ denotes the L1 norm $\|\boldsymbol{z}\|_1 = \sum_i |z_i|$. The first term is the cross-entropy loss and the second term is the L1 norm regularization with the hyperparameter λ. The loss function was optimized by the Adam algorithm with an initial learning rate of 0.001, using a fixed number of 1000 epochs. The hyperparameter λ was optimized by 5-fold cross-validation.

Unsupervised alignment using Gromov-Wasserstein distance

In this section, we provide an overview of unsupervised alignment methods for aligning two qualia structures (sets of embeddings) by using Gromov-Wasserstein optimal transport. With this method, we can quantify the degree of similarity between the qualia structures and in what way those are similar or different by examining 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 or the object 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.

$$X = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \end{pmatrix}, \quad Y = \begin{pmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_n \end{pmatrix}. \tag{7}$$

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:

$$\min_P \min_Q \|X - QYP\|_F^2, \tag{8}$$

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., $\boldsymbol{x}_j \leftrightarrow \sum_i P_{ij} \boldsymbol{y}_i$), 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 [43], 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

$$\min_Q \|X - QY\|_F^2. \tag{9}$$

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 Eq. 8, 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 [25, 44]. Denoting the optimal transportation plan (the assignment matrix) by Γ^*, the problem to solve becomes

$$\min_{Q} \|X - QY\Gamma^*\|_F^2. \tag{10}$$

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) [23]. GWOT is an unsupervised alignment technique that can find correspondence between two point clouds (embeddings) in different domains based on internal distances within the same 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 or objects and D_{ij} denotes the subjective rating of dissimilarity between the i-th and j-th color or object.

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

$$\min_{\Gamma} \sum_{i,j,k,l} (D_{ij} - D'_{kl})^2 \Gamma_{ik} \Gamma_{jl}. \tag{11}$$

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 $\boldsymbol{q}$. As for the distributions $\boldsymbol{p}$ 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

$$\boldsymbol{x}_j \leftarrow \sum_{i=1}^{n} \Gamma_{ij} \boldsymbol{y}_i. \tag{12}$$

Then, this mapping is subsequently used for finding the rotation matrix Q in Eq. 10.

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 [24, 28].

$$\min_{\Gamma} \sum_{i,j,k,l} (D_{ij} - D'_{kl})^2 \Gamma_{ik} \Gamma_{jl} + \epsilon H(\Gamma), \tag{13}$$

where $H(\Gamma)$ is the entropy of a transportation map.

To find good local optimums, we conducted hyperparameter tuning on ϵ in Eq. 13 by changing it from 0.02 to 0.16 with equal spacing (20 different values of ϵ) for the color similarity data and from 1.0 to 10.0 for the THINGS data. We chose the value of ϵ, where the optimized transportation plan minimises the Gromov-Wasserstein distance without the entropy-regularisation term (Eq. 11) following the procedure proposed by a previous study [26].

Initialization of transportation plan

To further facilitate the optimization of the transportation plan, we also implemented random initialization in addition to hyperparameter tuning. We used the built-in function (entropic_gromov_wasserstein) from the open source Python Optimal Transport library (POT) [45] for solving the entropic Gromov-Wasserstein problem (Eq. 13). In the built-in function, the transportation plan is initialized to uniform distribution when the marginals $\boldsymbol{p}$ and $\boldsymbol{q}$ are both uniform distributions. However, we observed that the optimization occasionally converged to bad local minimum with a large GWD. To prevent this, we found that it is effective to randomly initialize the transportation plan. We initialized the transportation plan 10 times for the color similarity data and 2 times for the THINGS data. Then, we selected the resulting transportation plan with the smallest GWD. 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 $\sum_{ij} \Gamma_{ij} = 1$.

Evaluation of unsupervised alignment

To assess the degree of similarity between the two qualia structures in the unsupervised setting, we computed k-nearest matching rate in the embedded space after the unsupervised alignment described in the previous section. Given the two sets of embeddings X and Y, we computed the Euclidean distance between all the pairs of the embeddings of X and Y. For each embedding of X (or Y), we checked whether the same embedding in Y (or X) is within the top k closest embeddings. If the embeddings of the same color or object from X and Y are within the k-nearest neighbor, we considered that the corresponding color or object is correctly matched. We computed the k-nearest matching rate as the percentage of embeddings collectedly matched based on the above criterion.

Data availability

Data from the THINGS-data is available at https://osf.io/f5rn6/.

Code availability

Code for the behavioral experiments is available at https://osf.io/9xwr2/.

References

[1] D. C. Dennett, Quining Qualia in *Consciousness in Modern Science*, A. J. Marcel, E. Bisiach Eds. (Oxford University Press, 1988).

[2] D. Rosenthal, "Quality spaces and sensory modalities" in *Phenomenal Qualities: Sense, Perception, and Consciousness*, P. Coates, S. Coleman Eds. (Oxford University Press, 2015).

[3] H. Lyre, Neurophenomenal structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness* **2022**, (2022).

[4] S. B. Fink, L. Kob, H. Lyre, A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences* **2**, (2021).

[5] N. Tsuchiya, H. Saigo, A relational approach to consciousness: categories of level and contents of consciousness. *Neuroscience of Consciousness* **2021**, (2021).

[6] T. Nagel, What is it like to be a bat? *The philosophical review* **83**, 435-450 (1974).

[7] D. J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, 1996).

[8] J. Kleiner, Mathematical Models of Consciousness. *Entropy* **22**, 609 (2020).

[9] A. Y. Lee, Modeling mental qualities. *Philosophical Review* **130**, 263-298 (2021).

[10] H. Lau, M. Michel, J. E. LeDoux, S. M. Fleming, The mnemonic basis of subjective experience. *Nature Reviews Psychology* **1**, 479-488 (2022).

[11] C. Tallon-Baudry, The topological space of subjective experience. *Trends in Cognitive Sciences* **26**, 1068-1069 (2022).

[12] R. Malach, Local neuronal relational structures underlying the contents of human conscious experience. *Neuroscience of consciousness*, **2021**, niab028 (2021)

[13] C. E. Helm, Multidimensional Ratio Scaling Analysis of Perceived Color Relations. *Journal of the Optical Society of America* **54**, 256-262 (1964).

[14] A. Tversky, Features of similarity. *Psychological review* **84**, 327 (1977).

[15] R. N. Shepard, L. A. Cooper, Representation of colors in the blind, color-blind, and normally sighted. *Psychological science* **3**, 97-104 (1992).

[16] G. P. Epping, E. L. Fisher, A. M. Zeleznikow-Johnston, E. M. Pothos, N. Tsuchiya, A Quantum Geometric Framework for Modeling Color Similarity Judgments. *Cognitive Science* **47**, e13231 (2023).

[17] A. M. Zeleznikow-Johnston, Y. Aizawa, M. Yamada, N. Tsuchiya, Are Color Experiences the Same across the Visual Field? *Journal of Cognitive Neuroscience* **2023**, 1-34 (2023).

[18] M. N. Hebart, A. H. Dickter, A. Kidder, W. Y. Kwok, A. Corriveau, C. Van Wicklin, C. I. Baker, THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one* **14**, 10, e0223792 (2019).

[19] M. N. Hebart, C. Y. Zheng, F. Pereira, C. I. Baker, Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behavior* **4**, 1173–1185 (2020).

[20] M. N. Hebart, O. Contier, L. Teichmann, A. H. Rockter, C. Y. Zheng, A. Kidder, A. Corriveau, M. Vaziri-Pashkam, C. I. Baker, THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife* **12**, e82580 (2023).

[21] J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–27 (1964).

[22] N. Kriegeskorte, R. A. Kievit, Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences* **17**, 401-412 (2013).

[23] F. Mémoli, Gromov-Wasserstein Distances and the Metric Approach to Object Matching. *Found Comput Math* **11**, 417–487 (2011).

[24] G. Peyré, M. Cuturi, Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends in Machine Learning* **11**, 355-607 (2019).

[25] D. Alvarez-Melis, T. Jaakkola, Gromov-Wasserstein Alignment of Word Embedding Spaces in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1881-1890 (2018).

[26] P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, R. Singh, SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport. *Journal of Computational Biology* **29**, 3-18 (2022).

[27] A. Thual, H. Tran, T. Zemskova, N. Courty, R. Flamary, S. Dehaene, B. Thirion, Aligning individual brains with Fused Unbalanced Gromov-Wasserstein. arXiv:2206.09398 [Neurons and Cognition] (2022).

[28] G. Peyré, M. Cuturi, J. Solomon, Gromov-wasserstein averaging of kernel and distance matrices in *Proceedings of ICML*, 2664-2672 (2016).

[29] B. D. Roads, B. C. Love, Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence* **2**, 76–82 (2020).

[30] N. Block, Wittgenstein and qualia. *Philosophical Perspectives* **21**, 73-115 (2007).

[31] G. Lee, The experience of left and right in *Perceptual Experience*, T. S. Gendler, J. Hawthorne Eds. 291-315 (Oxford University Press, 2006).

[32] L. D. Griffin, Similarity of psychological and physical colour space shown by symmetry analysis. *Color Research and Application* **26** 151-157 (2001).

[33] S. E. Palmer, Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences*, **22**, 923-943 (1999)

[34] M. A. Webster, E. Miyahara, G. Malkoc, V. E. Raker, Variations in normal color vision. II. Unique hues. *J. Opt. Soc. Am. A* **17**, 1545-1555 (2000).

[35] J. Locke, *An essay concerning human understanding* (Oxford University Press, 1689).

[36] D. Rosenthal, How to think about mental qualities. *Philosophical Issues* **20**, 368-393 (2010).

[37] A. S. Cowen, D. Keltner, Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences* **114**, e7900-e7909 (2017).

[38] L. Nummenmaa, R. Hari, J. K. Hietanen, et al. Maps of subjective feelings. *Proceedings of the National Academy of Sciences* **115**, 9198-9203 (2018).

[39] P. Gardenfors, *Conceptual spaces: The geometry of thought* (The MIT Press, 2000).

[40] J. Birch, Efficiency of the ishihara test for identifying red-green colour deficiency. *Ophthalmic and Physiological Optics* **17**, 403–408 (1997).

[41] A. Pouw, R. Karanjia, A. Sadun, A method for identifying color vision deficiency malingering. *Graefe's Archive for Clinical and Experimental Ophthalmology* **255**, 613–618 (2017).

[42] N. Saji, M. Imai, M. Asano, 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. *Cognitive Science* **44**, 12813 (2020).

[43] J. C. Gower, G. B. Dijksterhuis, *Procrustes Problems* (Oxford University Press, 2004).

[44] J. Alaux, E. Grave, M. Cuturi, A. Joulin, Unsupervised hyperalignment for multilingual word embeddings. in *Proceedings of ICLR* (2019).

[45] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N.T.H. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, Pot: Python optimal transport. *Journal of Machine Learning Research* **22**, 1–8 (2021).

Acknowledgments. GK and MO were supported by JST Moonshot R&D Grant Number JPMJMS2012. NT and MO were supported by Japan Promotion Science, Grant-in-Aid for Transformative Research Areas Grant Numbers 20H05710 (NT) and 20H05712 (MO). NT was supported by Australian Research Council (DP180104128, DP180100396). NT and AZ were supported

by National Health 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. We thank the authors of the THINGS data for providing us with early access to the THINGS data.

Supplementary Information

Supplementary Figures S1, S2, S3, S4
Supplementary Movies S1, S2