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

Genji Kawakita[1*], Ariel Zelenikow-Johnson[2*], Naotsugu Tsuchiya[2 3 4†], Masafumi Oizumi[5†]

1. Imperial College London, 2. Monash University, 3. National Institute of Information and Communications Technology, 4. Advanced Telecommunications Research Computational Neuroscience Laboratories, 5. The University of Tokyo

* These authors contributed equally to this work. † These authors contributed equally to this work.

Take-home messages

- We propose an unsupervised method for assessing the equivalence of qualia structures across individuals.
- Color qualia structures were consistent across groups of color-neurotypical participants, but not between color-neurotypical and color-blind groups.

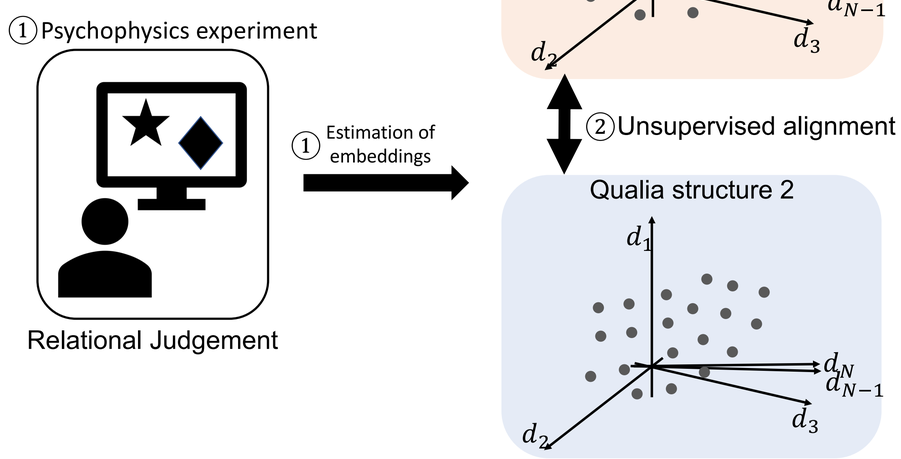
Introduction

- Whether one person's experience of "red" is equivalent to someone else's has long been considered unanswerable.
- 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 [Tsuchiya & Saigo 2021].
- One particular approach is to analyze reports of subjective similarities between sensory experiences. 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"

Qualia structure paradigm

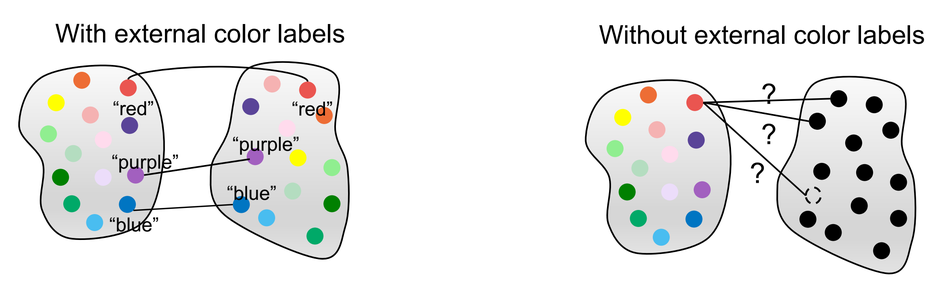
Qualia structure paradigm consists of 2 steps.

1. Collecting a large number of relationships between qualia through subjective reports to estimate the relational structure of these qualia (**qualia structure**).
2. Comparing qualia structures between participants without assuming correspondence between individual qualia

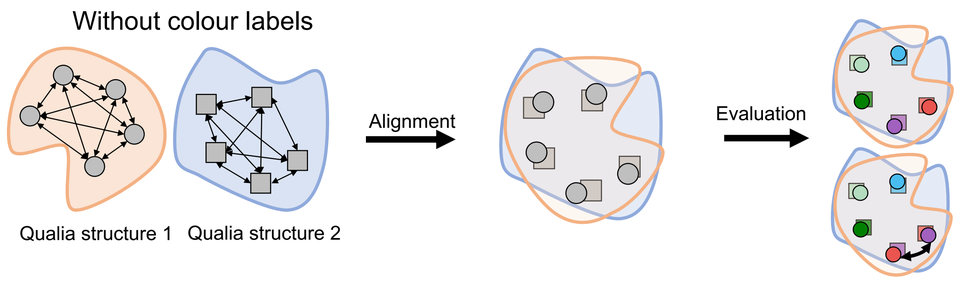


Unsupervised alignment of qualia structures

- To compare qualia structures, one might assume stimulus-level "extrinsic" correspondence, i.e., supervised alignment, which is not guaranteed. We therefore need to consider all possibilities of correspondence (e.g., my "red" can be your "blue"), i.e., unsupervised alignment.

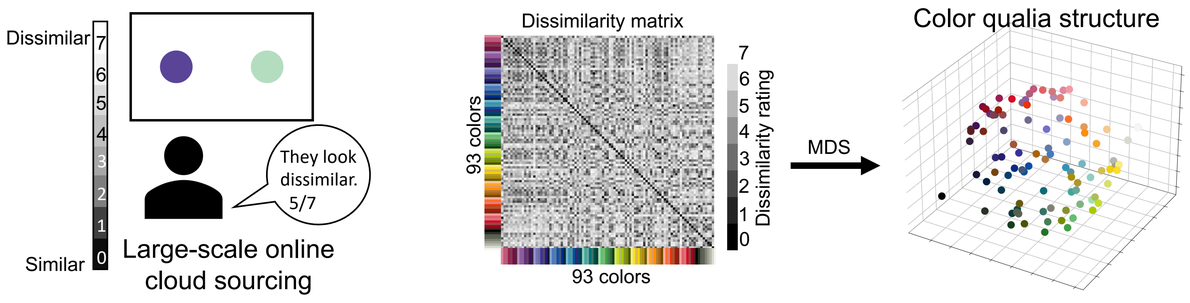


- In unsupervised alignment, we try to find the optimal matching between qualia structures based only on their internal relationships. After finding optimal alignment, we can then use external labels to evaluate how the embeddings of different individuals relate to each other.



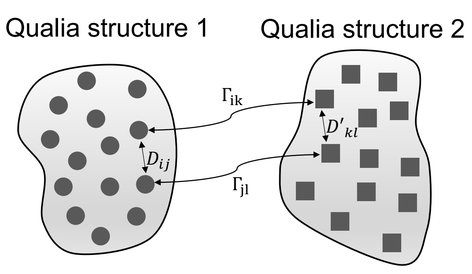
Method 1: Color similarity judgement task

- 426 color-neurotypical and 207 color-atypical participants were recruited to report similarities between pairs of colors.
- The similarity relationships were represented as a matrix D, from which embeddings of colors can be estimated as an approximation of a qualia structure using multidimensional scaling (MDS).



Method 2: Gromov-Wasserstein optimal transport

- We used Gromov-Wasserstein optimal transport (GWOT) [Memoli 2011] to assess the degree of equivalence between qualia structures without assuming any correspondence between experiences across individuals.
- GWOT finds optimal mapping (transportation plan),Γ, between point clouds in different domains that minimizes Gromov-Wasserstein distance (GWD), which can be computed solely based on the distance relationships of points "within" each domain without relying on correspondence "across" domains.
- GWOT has been successfully used in various topics including unsupervised translation of languages [Alvarez-Melis & Jaakkola 2018].

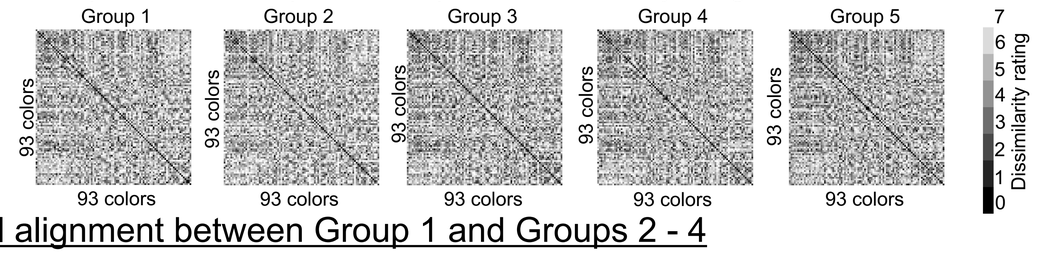


$$GWD = \min_{\Gamma} \sum_{i,j,k,l} \left| D_{ij} - D'_{kl} \right|^2 \Gamma_{ik} \Gamma_{jl}$$

Result 1: Color qualia structures were consistent across color-neurotypical groups

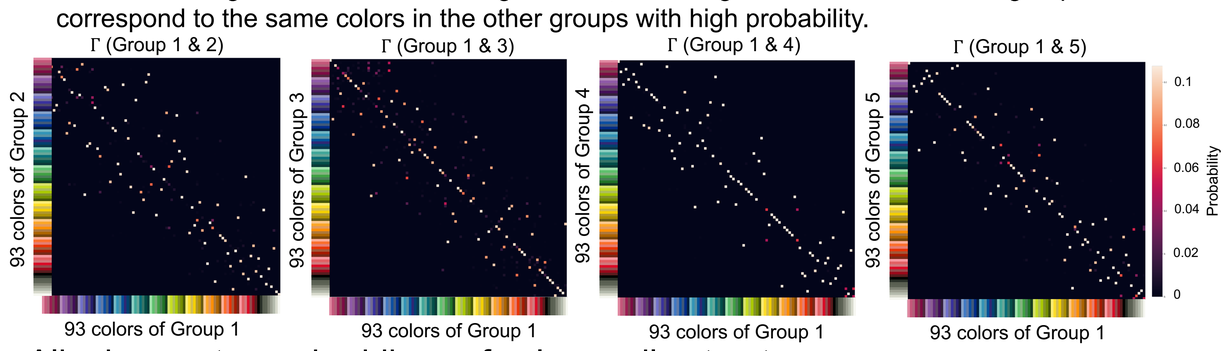
Color dissimilarity matrices for five groups

- We divided color pair similarity data 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.



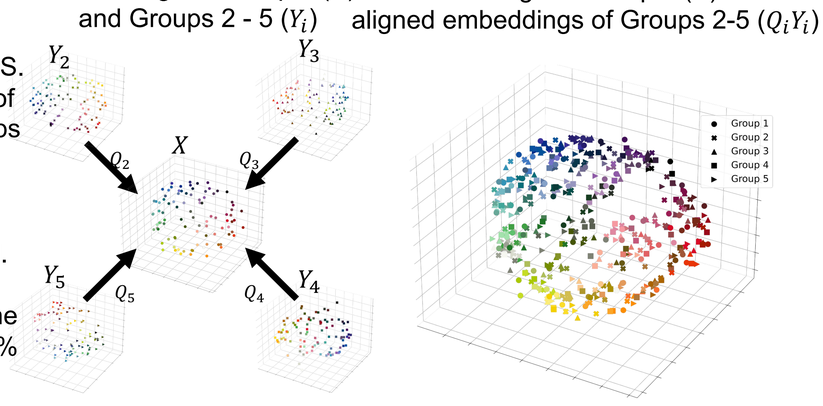
Optimal alignment between Group 1 and Groups 2 - 4

- We then computed GWD between pairs of groups to obtain optimized mappings Γ^* between the dissimilarity matrix of Group 1 and those of Groups 2 to 5.
- Most of the diagonal elements show high values, indicating that most colors in one group correspond to the same colors in the other groups with high probability.



Aligning vector embeddings of color quaila structures

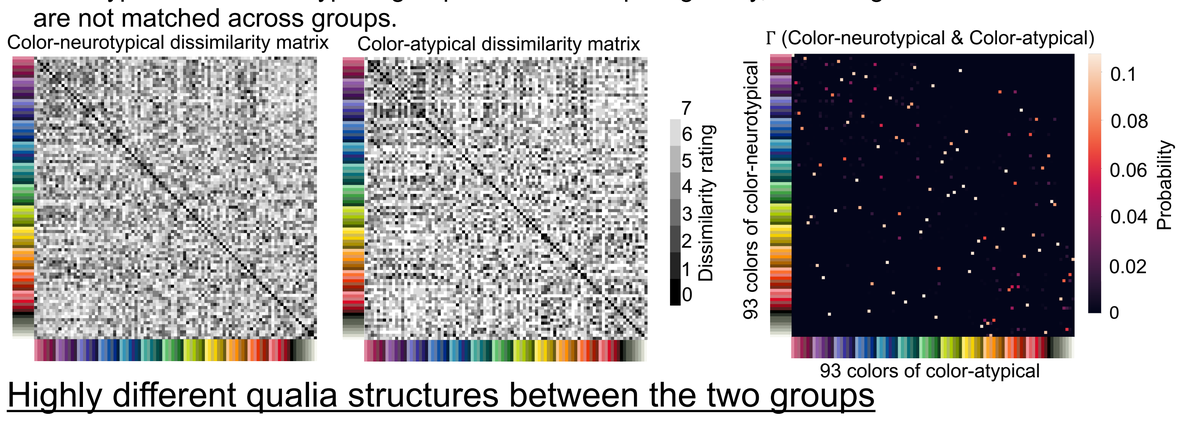
- We next performed unsupervised alignment of the vector embeddings of color qualia structures($X, \mathrm{Y_i}$) obtained by MDS.
- We found that the embeddings of similar colors from the five groups are located close to each other, indicating that similar colors are 'correctly' aligned by the unsupervised alignment method.
- We computed the average k-nearest color matching rate in the aligned space, which were 75.8% when k = 1, 90.8% when k = 3, and 94.2% when k = 5.



Result 2: Qualia structures of color-neurotypical and color-atypical groups were highly different

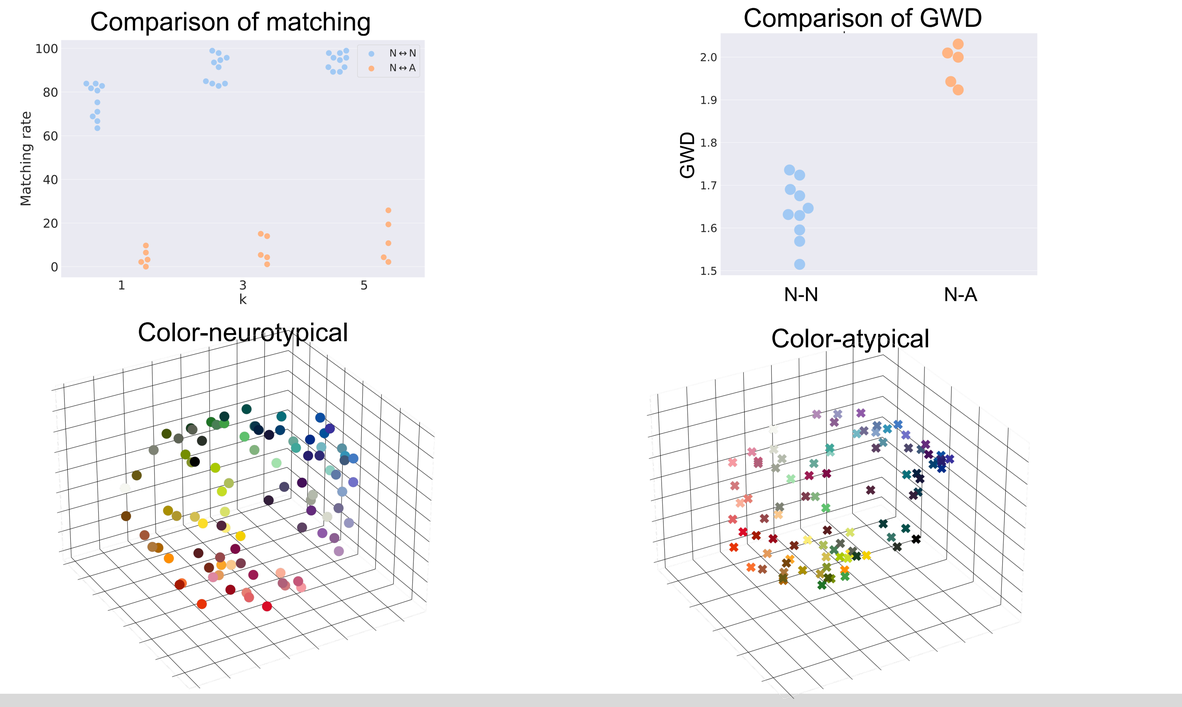
Optimal alignment between color-neurotypical and color-atypical groups

- To investigate whether we can align possibly different color qualia structures of color-atypical participants with those of color-neurotypical participants, we conducted color similarity judgement tasks on 207 color-atypical participants and obtained a dissimilarity matrix.
- Unlike alignment between color-neurotypical groups, optimized mapping Γ^* between color-neurotypical and color-atypical groups is not lined up diagonally, indicating that the same colors are not matched across groups.



Highly different qualia structures between the two groups

- Top k matching rate between Group 1-5 and color-atypical group is 4.3%, 8.0% and 12.5% when k = 1, 3, 5, respectively.
- GWD between color-neurotypical and color-atypical groups are significantly larger than any of the GWD values between color-neurotypical groups.
- 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.



Discussion

- By using an unsupervised alignment method, we were able to match the color qualia structures of different groups of participants based only on the way the color qualia relate to each other, without using any external color labels.
- While we focused only on color similarities, our method has the potential to be applied to a wide range of subjective experiences and different modalities (visual objects, emotion, semantic concepts, etc.).
- Our approach offers a novel and powerful tool for quantitatively exploring various aspects of subjective experiences and advancing our understanding of consciousness.

References

N. Tsuchiya, H. Saigo, A relational approach to consciousness: categories of level and contents of consciousness. Neuroscience of Consciousness 2021, (2021).
F. Memoli, Gromov-Wasserstein Distances and the Metric Approach to Object Matching. Found Comput Math 11, 417–487 (2011).
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).

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.