Inferring Individualized Color-Vision Distortions from fMRI Hue-Representation Geometry

Jinil Kim [1 2 3] **Albert Minkue Cho** [4] **Jungwoo Seo** [5] **Jiook Cha** [2 5 6]

Abstract

Color vision deficiency (CVD) distorts the pairwise relations between neural hue representations while categorical recognition is relatively preserved. Such structured distortions recur across health-related data and can act as phenotypes, yet interpreting these neural signatures and translating them into intervention remains challenging. Here we formulate fMRI-measured hue responses as structured neural representations and infer the low-dimensional distortions that explain their geometry. We quantify this distortion as an impairment of continuous hue structure—how well each hue can be reconstructed from the others—yielding a per-hue vulnerability profile. A complementary representational distance matrix (RDM) captures the relative distances between hue representations, identifying which hue pairs are pulled apart or pushed together relative to healthy controls. The framework then fits a two-parameter cortical model—one that distorts hue angles along the brain's two color-opponent axes—to these signatures by held-out predictive loss. We then analytically invert the fitted model into a stimulus-space color-correction filter that is exact for every displayed hue (residual $< 0.001°$). In two CVD individuals, this yields an individualized correction computed directly from each person's neural distortion signature; whether it improves perception is the framework's defined next test.

[1]Department of Linguistics, Seoul National University, Seoul, South Korea [2]Department of Psychology, Seoul National University, Seoul, South Korea [3]Department of Computer Science and Engineering, Seoul National University, Seoul, South Korea [4]Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea [5]Department of Brain & Cognitive Sciences, Seoul National University, Seoul, South Korea [6]Interdisciplinary Program of Artificial Intelligence, Seoul National University, Seoul, South Korea. Correspondence to: Jiook Cha <connectome@snu.ac.kr>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Many health-related data carry distorted structure that can act as a phenotype—systematic, repeatable patterns in how stimuli relate within a high-dimensional biological response space. Color vision deficiency (CVD), for example, produces distorted pairwise relations between neural hue representations even while categorical recognition remains relatively preserved. Other instances of such structured-phenotype data include geometric deformations of cardiac anatomy (Biffi et al., 2020) and directional shifts in gene-expression latent spaces under drug perturbation (Lotfollahi et al., 2019). Yet such phenotypes are challenging to interpret and, especially, to translate into individually-targeted intervention. When such structural distortion can be parameterized, it can be inverted for correction—a route absent from current generic correction approaches.

In a geometric framework, responses to stimuli can be summarized as a *representational geometry*, where pairwise distances capture what a neural population distinguishes (Kriegeskorte et al., 2008), including the cortical representation of color (Brouwer & Heeger, 2009; Bannert & Bartels, 2018). Despite this perspective, structured distortions have rarely been used to derive individualized perceptual corrections. The closest analogue, a cochlear simulator that inverts a hearing-impairment deformation model (Irino, 2023), targets simulation of sensory loss rather than its correction.

CVD arises from altered spectral sensitivity of retinal cone photoreceptors, affecting approximately 8% of males (Machado et al., 2009), and produces structured distortions in cortical color representations that vary across individuals, even within one diagnostic category (Tregillus et al., 2021). Current assistive approaches, including commercial spectral-notch filters, apply a fixed, generic correction. Recent evidence indicates that such filters alter color *appearance* but minimally affect color *discrimination* at threshold (Somers et al., 2024). Effective correction may therefore require individually-tailored strategies.

Categorical color recognition can remain comparatively preserved in CVD: dichromats and anomalous trichromats often name and categorize colors more like color-normal observers than their threshold losses would predict (Bonnardel,

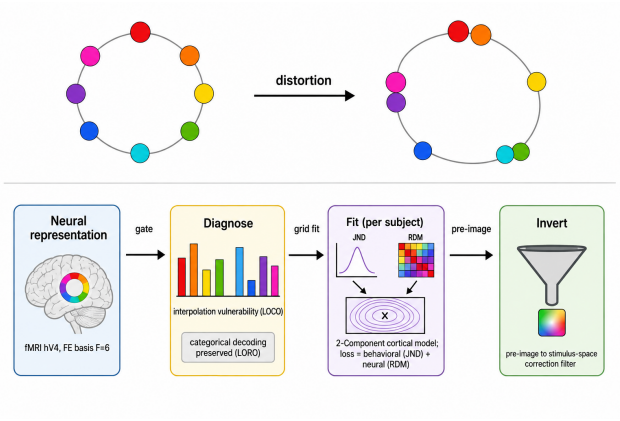

Figure 1. Structured distortion and the fit-and-invert pipeline. **Top**: CVD distorts the continuous geometry of hue representations—some neighboring hues collapse while others expand—even though classification accuracy is preserved. **Bottom**: The pipeline quantifies the structured distortion via an interpolation vulnerability profile and the pairwise representational distances (ΔRDM), fits a parsimonious (2-DOF) cortical distortion model, and analytically inverts it into a per-subject stimulus-space correction filter.

2006; Emery et al., 2021a). Continuous hue relationships, however, may still be altered, and classification accuracy is insensitive to this relational structure. In CVD, the pairwise geometry of stimulus-evoked fMRI responses is itself distorted in low-dimensional, parameterizable ways.

The unit of inference is each individual's distortion model, not the population mean. Our contributions are:

1. We formulate distorted neural geometry as a structured-distortion inference problem, quantified by an interpolation vulnerability profile and a pairwise representational-distance difference from healthy controls that expose continuous-geometry deficits invisible to categorical classification.
2. We infer a low-dimensional (2-DOF) cortical opponent-axis distortion model, selected among candidate retinal and retinal–cortical mechanisms by held-out predictive loss and structural adequacy.
3. We invert the fitted cortical model into subject-specific candidate corrections, individually tailored unlike current generic filters.

2. Methods

2.1. Data Acquisition

Ten participants (7 healthy controls [HC], 3 females, age 22.7 ± 2.5 years; 3 color-vision-deficient [CVD], all male: two deutan and one protan) were assessed with the 14-plate Ishihara test (Ishihara, 1917); all three CVD cases were confirmed. Given the small CVD sample, all CVD results are interpreted as individual case demonstrations, not population-level effects (Crawford & Howell, 1998).

Stimuli consisted of eight colors equally spaced from 0° to 315° in CIE $L^*a^*b^*$ space ($L^* = 75$, chroma distance 40). Participants performed a modified Rapid Serial Visual Presentation (RSVP) task (Brouwer & Heeger, 2009) across six runs, viewing 1.5 s color patches. Functional images (Siemens 3T; TR = 1.5 s; 2 mm isotropic; 24 oblique slices) were preprocessed with fMRIPrep and projected to retinotopically-defined ROIs (V1, V2, hV4) via the Wang probabilistic atlas at 50% threshold (Wang et al., 2015). Response amplitudes were estimated using a two-stage FIR-then-GLM procedure (Brouwer & Heeger, 2009) and Procrustes-aligned across runs.

Outside the scanner, we also measured color-discrimination thresholds (just-noticeable differences, JND) for eight hue pairs using adaptive 1-up/1-down staircases; these enter the fitting loss as CVD-to-HC ratios (Section A.3).

2.2. Structured Distortion Quantification

We compute two complementary read-outs of the distortion: an interpolation vulnerability profile in hV4, the established cortical locus of perceptual color representation (Brouwer & Heeger, 2009; Bannert & Bartels, 2018), and a pairwise representational distance matrix (RDM) in each ROI (V1–hV4). Both derive from a common encoding model of the hue representation.

We modeled hue-selective neural populations with half-wave rectified and squared sinusoidal basis functions (Brouwer & Heeger, 2009): $r_k(\theta) = \max(0, \cos(\theta - \theta_k))^2$ for $F = 6$ channels with preferred hues equally spaced at 60°. A weight matrix $W \in \mathbb{R}^{F \times V}$ mapping channels to V voxel responses was estimated by ridge regression with generalized-cross-validation regularization (Golub et al., 1979). This encoding model turns raw voxel data into a structured representation in which the relational geometry across hues is explicitly parameterized.

Two cross-validation schemes probe this representation. Leave-one-*run*-out (LORO) holds out one run—a single repetition of all eight hues—and tests whether each already-seen hue stays decodable across runs (category identity). Leave-one-*color*-out (LOCO) holds out one hue—all six of its repetitions—and tests whether that hue can be reconstructed by interpolating the seven remaining hues (relational geometry). The eight per-hue LOCO accuracies form an interpolation vulnerability vector $\mathbf{v} \in [0, 1]^8$, which exposes continuous-geometry deficits that categorical classification cannot see. We focus distortion inference on hV4, the only ROI supporting above-chance interpolation in HC.

The pairwise read-out is the RDM. For each participant, an 8×8 crossnobis distance matrix (Walther et al., 2016;

Nili et al., 2014) is computed in shared-response-model (SRM) aligned space (Chen et al., 2015), and the difference $\Delta\text{RDM} = \text{RDM}_{\text{CVD}} - \overline{\text{RDM}}_{\text{HC}}$ measures how that individual's pairwise geometry deviates from the HC mean. The vulnerability vector $\mathbf{v}$ (8-dim) and ΔRDM (28-dim) are the diagnostic signatures fitted in Section 2.4.

2.3. Candidate Distortion Mechanisms

We evaluate three candidate mechanisms for how a stimulus hue θ maps to a distorted angle θ' at distinct processing levels (notation in Section A.1): a retinal cone shift (Machado, 1-DOF (Machado et al., 2009); Section A.6), a retinal shift under cortical compensation (R+C, 1-DOF), and a cortical opponent-axis reorganization (2-Component, 2-DOF). These are competing hypotheses about the locus of distortion, not a nested escalation.

Retinal + cortical compensation (R+C; 1 DOF). The R+C model scales a fixed Machado retinal shift $\Delta\lambda$ by a single cortical compensation gain g (the only free parameter):

$$\delta\theta_{\text{RC}}(\theta) = (2 - g)\, \delta\theta_{\text{Machado}}(\theta; \Delta\lambda), \tag{1}$$

where g=1 is no compensation, g=2 exact compensation, and g>2 overcompensation that reverses the retinal shift past the undistorted hue (Tregillus et al., 2021).

Cortical opponent-axis model (2 DOF). The cortical model expresses the distortion as angular dilation along an S-cone axis (β_s) and a type-specific confusion axis (β_c):

$$\theta'(\theta) = \theta + \beta_s \cos(\theta - 90^\circ) + \beta_c \cos(\theta - \theta_{\text{conf}}), \tag{2}$$

with $\theta_{\text{conf}} = 16^\circ$ (protan) or 150° (deutan). The S-cone term is motivated by behavioral evidence for hue-angle reorganization in CVD (Emery et al., 2021b); the two axes are the cardinal S-cone direction (90°) and the subtype-specific confusion line (θ_{conf}), which together span the plane of candidate cortical distortions.

2.4. Fitting Loss and Model Selection

For a candidate distortion, the simulated geometry is obtained by applying $\delta\theta$ to the HC-derived representation and compared to the observed CVD geometry. Three loss families quantify the CVD–HC difference: a behavioral JND-ratio rank loss L_γ, a representational ΔRDM cosine loss L_{RDM}, and an hV4 LOCO loss L_{LOCO}. Because the atoms are on incommensurable scales (rank-based behavioral vs. cosine-based neural), each atom's loss is z-scored across the (β_s, β_c) grid before the atoms are combined into a composite; the minimizing grid point is taken as the fit.

Loss combinations and parameters were selected by held-out generalization rather than in-sample significance. Over the full candidate pool—all non-empty combinations of behavioral ($\gamma_{\text{focal}}/\gamma_{\text{all}}$), neural ΔRDM (per ROI, V1–hV4), and hV4 L_{LOCO} atoms, each composite normalized for atom count—each was ranked by the median held-out test composite loss (5-train/2-test HC resample, N=300), then by its test-loss IQR; combinations with grid-boundary saturation above 50% were rejected as structurally inadequate. A strict 7-fold leave-one-HC-out supplement tests whether the fitted distortion predicts each held-out HC's geometry better than the no-correction baseline ($\delta\theta = 0$).

Once fitted, the 2-Component transform is analytically inverted to a per-hue stimulus-space pre-image $\tilde{\theta}_k$, solving $\theta'(\tilde{\theta}_k) = \theta_k$ (Brent root-finding; Section A.3) and giving the correction $\delta\theta_k^{\text{filt}} = \tilde{\theta}_k - \theta_k$.

3. Results

3.1. A Structured Cortical Distortion

Discrimination preserved, interpolation impaired. A stimulus-space correction is well-posed only if the displayed hues remain neurally distinguishable yet their continuous arrangement is distorted. Both conditions hold. Cross-run (LORO) classification of the eight hues shows no detectable HC–CVD difference (pooled cross-subject decoding, $p = 0.668$), and both fit-recovered CVD participants exceed chance at every ROI: the hues are individually distinguishable at the cortical level. Cross-color (LOCO) interpolation, by contrast, is selectively impaired. Only hV4 supported above-chance interpolation in HC—measured as adjacent accuracy, the fraction of predictions within ± 1 hue step of the target (chance $3/8$)—at 0.47 ± 0.05 ($p = 0.044$), replicating Brouwer & Heeger (2009); there the protan participant fell significantly below chance (Sub-09 0.13, $p = 0.024$), and the deutan participant showed a marginal reduction in the same direction (Sub-08 0.25, $p = 0.082$). The deficit concentrated on the S-cone intermediate hues (blue, purple, magenta) in an uncorrected, exploratory per-hue test. The per-hue interpolation accuracies form each participant's vulnerability profile $\mathbf{v}$.

Pairwise geometry distorted at a distinct ROI per subject. The geometric basis of this deficit appears as elevated pairwise distortion that localizes to a different early-visual area in each participant (Figure 2). Sub-08 (deutan) shows significantly elevated representational disparity at V2 (Crawford–Howell $p = 0.040$), with no elevation at V1, V3, or hV4; Sub-09 (protan) shows it at V1 ($p = 0.007$), with no elevation elsewhere. The corresponding ΔRDM heatmaps reveal subject-specific distortion structure—elevated distances among S-cone-intermediate pairs for Sub-08, a broader reorganization of pairwise distances for Sub-09. That the same categorical-discrimination-preserving representation is geometrically distorted—and in an individually

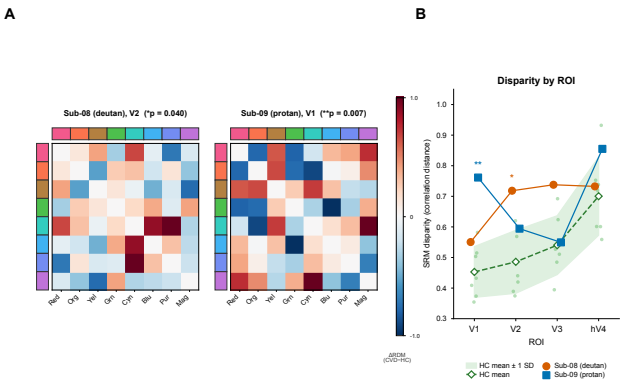

Figure 2. Pairwise color geometry is distorted at a distinct ROI in each CVD participant. **(A)** $\Delta\text{RDM} = \text{RDM}_{\text{CVD}} - \overline{\text{RDM}}_{\text{HC}}$ in SRM-aligned space at each subject's affected ROI (Sub-08 V2, Sub-09 V1); warm = CVD distances larger than the HC mean, cool = smaller. **(B)** Mean pairwise disparity per subject per ROI; HC band = group mean ± 1 SD ($n = 7$), individual HC leave-one-out as dots. Each CVD participant shows significantly elevated disparity at a single, distinct ROI (Sub-08 V2, $p = 0.040$; Sub-09 V1, $p = 0.007$; Crawford–Howell single-case, exploratory).

patterned way—is the structured-distortion phenotype this framework targets. A third CVD participant (Sub-10) is null at both the LOCO and SRM levels and does not enter distortion inference; they are reserved as a future-work probe of alternative representational dimensions.

3.2. An Individualized, Invertible Correction Filter

Model selection: structural modeling beats cone-shift. Directly modeling the structural distortion as a cortical opponent-axis reorganization is more adequate than the retinal cone-shift accounts, which are structurally inadequate as exact filters. The R+C model overcompensates, its single gain pinned near the grid ceiling and lacking any axis orthogonal to the confusion direction: Sub-08 trips the 50% saturation gate (saturation rate 100%, $g \to 3.0$), while Sub-09 (41%, $g \approx 2.95$) stays below the gate but is dominated on held-out loss. Its retinal-only limit (Machado, the $g=1$ case) is moreover non-invertible at protan severity, compressing three hues into a narrow arc so that they lose a unique pre-image and only 4 of 8 of Sub-09's hues admit an exact correction (Sections A.6 and A.7). Only the 2-Component model fits in the grid interior for both participants and is carried forward (Table 1); its per-subject loss combination (Sub-08: $\gamma_{\text{OY}} + L_{\text{RDM}}^{(\text{V2})}$; Sub-09: $\gamma_{\text{all}} + L_{\text{RDM}}^{(\text{V1})}$) and (β_s, β_c) were selected by held-out test composite loss, and the fit clears the no-correction null on all 7 leave-one-HC-out folds on the neural (ΔRDM) criterion—a null-comparison sanity check, since most candidate fits clear this baseline, rather than a discriminative ranking among them. We report the mechanism class—the sign quadrant of (β_s, β_c)—as descriptive: the averaged loss surface for the real CVD fits is

Table 1. Selected per-subject 2-Component fit (hV4). (β_s, β_c) in degrees report the mechanism class (sign quadrant), not point-resolved magnitudes. Test loss: median held-out composite (5/2 HC split, N=300; lower is better). All fits clear the no-correction null on $7/7$ leave-one-HC-out folds (a sanity check, not a discriminative ranking). Stability: parameter IQR over resamples or modal-bin share. Sub-09's (β_s, β_c) is metric-dependent (SRM basis gives $(+32, 0)$; Section A.5), so its filter is illustrative.

Subject	(β_s, β_c)	Test loss	Stability
Sub-08 (deutan)	$(+6, -42)$	-2.36 ± 2.15	IQR $(8, 2)$
Sub-09 (protan)	$(+2, +24)$	-1.54 ± 1.42	mode 87.7%

$2.1\times$–$5.5\times$ deeper than the HC null, while absolute magnitudes are not identifiable (Section A.5).

Sub-08 (deutan). The 2-Component model places Sub-08 in the deutan confusion-axis quadrant ($\beta_s > 0$, $\beta_c < 0$). The fit has the lowest held-out median test loss (-2.36 ± 2.15) and its confusion-axis sign is stable across reference sets (7-fold leave-one-HC-out $\beta_c \in [-46°, -38°]$, no crossing of zero); the neural and behavioral terms independently support the same deutan direction (neural $\Delta L = -0.406$, $7/7$ folds; behavioral $\Delta L = -13.8$, $5/7$ folds).

Sub-09 (protan). Under the production PCA-basis RDM loss—adopted for both participants for its parameter stability (Sub-09 resample mode share 87.7%, $263/300$, versus $57\%/64\%$ for the two SRM-basis variants)—the 2-Component model fits Sub-09 with a dominant confusion-axis rotation and a small S-cone component, clearing the no-correction null on all 7 folds on the neural criterion ($\Delta L = -0.472$). This protan fit is presented as illustrative: the confusion-axis sign is identified only under the PCA basis, and the dominant axis itself is basis-dependent—the argmin moves to $(\beta_s, \beta_c) = (+32, 0)$ under the alternative SRM-basis reduction, where the confusion-axis term vanishes in favor of an S-cone shift (Section A.5).

The neural term carries what behavior cannot. Fitting the behavioral and neural atoms separately shows that the neural geometry contributes beyond behavior. For Sub-09 the behavioral-only fit improves on the no-correction baseline in only $3/7$ leave-one-HC-out folds, whereas the neural ΔRDM term improves in all 7 and recovers a stable confusion-axis rotation, so the protan fit rests on the neural geometry alone. For Sub-08 the two terms agree, and the neural term sharpens the estimate.

Inversion to a per-subject filter. Each fitted transform admits an exact stimulus-space pre-image for all eight hues (residual $< 0.001°$), yielding a directly implementable 8-point correction lookup table per subject (mean $|\delta\theta| = 26.3°$ for Sub-08, $16.2°$ for Sub-09; Figure 3). The two filters differ by subtype in the direction of the dominant

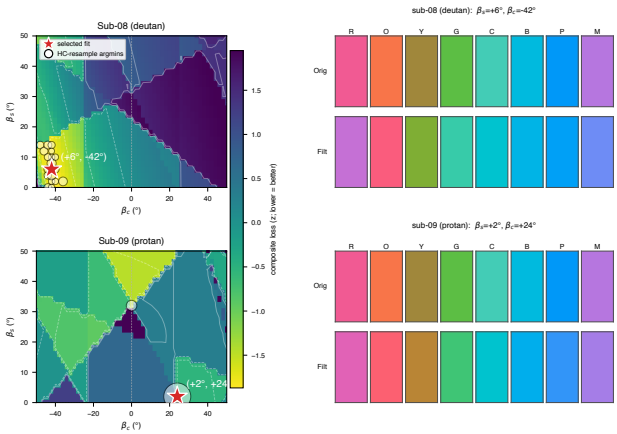

Figure 3. Per-subject 2-Component fit and inverted filter (hV4). Each row is one participant. **Left**: composite loss landscape over the (β_s, β_c) grid, with the selected fit (argmin, red star) and HC-resample argmins (white). **Right**: the inverted stimulus-space filter—displayed (Original) vs. corrected (Filtered) for the eight scanner hues. Sub-08 (deutan) $(+6°, -42°)$; Sub-09 (protan) $(+2°, +24°)$. Parameters report the mechanism class (sign quadrant), not point-resolved magnitudes (Section A.5). Observed vs. predicted ΔRDM fits are in Section A.2.

confusion-axis term ($\beta_c = -42°$ deutan vs. $+24°$ protan); each is built from that individual's own cortical geometry rather than a fixed subtype-average spectral shift.

4. Discussion

From structured distortion to an invertible correction. A CVD color deficit appears in these data not as a uniform loss of chromatic signal, but as an individually patterned geometric distortion of the cortical color representation. Categorical discrimination is preserved, while continuous-hue interpolation (hV4) and pairwise geometry (V2 in the deutan, V1 in the protan participant) are impaired. The LOCO vulnerability profile compresses high-dimensional voxel data into an 8-element geometry-sensitive summary. A composite loss that decomposes geometric fidelity into behavioral and pairwise neural components then yields a low-dimensional model that is both interpretable and analytically invertible. Grounding the fit in cortical geometry, rather than a behavioral threshold alone, is what makes the resulting filter individual: it is read from that participant's own representation rather than assigned from their diagnostic subtype. Future work could replace grid search with amortized inference (e.g., conditional normalizing flows) for real-time distortion estimation, and the approach may generalize to other conditions where sensory representations are distorted rather than absent.

Limitations. This is a proof-of-concept, and three considerations bound it. First, the fit-recovered CVD sample is $N = 2$ (with Sub-10 null); single-case statistics (Crawford & Howell, 1998) permit individual-level inference, but population- or subtype-level claims require replication, and establishing per-person rather than subtype-average correction specifically requires testing several individuals within one subtype. Second, what the fits support is mechanism-class consistency—the sign quadrant of (β_s, β_c)—not absolute magnitudes: grid optimization on 8 data points inflates the HC false-positive rate (Section A.5), and across pre-specified null tests $0/3$ candidate fits pass both null sources, with a per-axis noise floor of $\sim 20°/25°$, so the parameters are a low-dimensional descriptive embedding rather than physiological distortion magnitudes. Third, we have *not* shown that the correction improves perceptual discrimination; a behavioral advantage and broader validation remain to be established—the framework's immediate next step.

5. Conclusion

We present a framework that quantifies structured distortions in fMRI hue representations and analytically inverts a parsimonious cortical model into an individualized, stimulus-space correction filter. As a proof-of-concept in two CVD individuals, the inverted filter is exact for all displayed hues. Directly modeling the distortion is structurally more adequate than retinal cone-shift accounts. More broadly, recasting a sensory deficit as an invertible distortion of its own cortical representation—characterized one individual at a time—offers a template for personalized perceptual correction.

Impact Statement

The broader implication of this work is methodological: many health conditions leave structured rather than absent signatures in high-dimensional measurements. Mapping such signatures onto low-dimensional, invertible geometric features could turn descriptive phenotypes into potential, individualized intervention targets.

References

Bannert, M. M. and Bartels, A. Human V4 activity patterns predict behavioral performance in imagery of object color. *Journal of Neuroscience*, 38(15):3657–3668, 2018.

Biffi, C., Oktay, O., Tarroni, G., Bai, W., De Marvao, A., Doumou, G., Rajchl, M., Bedair, R., Prasad, S., Cook, S., O'Regan, D., and Rueckert, D. Explainable anatomical shape analysis through deep hierarchical generative models. *IEEE Transactions on Medical Imaging*, 39(6): 2088–2099, 2020. doi: 10.1109/TMI.2020.2964499.

Bonnardel, V. Color naming and categorization in inherited

color vision deficiencies. *Visual Neuroscience*, 23(3–4): 637–643, 2006. doi: 10.1017/S0952523806233558.

Brouwer, G. J. and Heeger, D. J. Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, 29(44):13992–14003, 2009.

Chen, P.-H., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J. V., and Ramadge, P. J. A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems*, volume 28, pp. 460–468, 2015.

Crawford, J. R. and Howell, D. C. Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist*, 12(4):482–486, 1998.

Emery, K. J., Parthasarathy, M. K., Joyce, D. S., and Webster, M. A. Color perception and compensation in color deficiencies assessed with hue scaling. *Vision Research*, 183:1–15, 2021a. doi: 10.1016/j.visres.2021.01.006.

Emery, K. J., Volbrecht, V. J., Peterzell, D. H., and Webster, M. A. Variations in normal color vision. VII. relationships between color naming and hue scaling. *Vision Research*, 183:1–12, 2021b.

Golub, G. H., Heath, M., and Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Irino, T. Hearing impairment simulator based on auditory excitation pattern playback: WHIS. *IEEE Access*, 11:78419–78430, 2023. doi: 10.1109/ACCESS.2023.3298673.

Ishihara, S. *Tests for Colour-Blindness*. Handaya, Tokyo, 1917.

Kriegeskorte, N., Mur, M., and Bandettini, P. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Lotfollahi, M., Wolf, F. A., and Theis, F. J. scGen predicts single-cell perturbation responses. *Nature Methods*, 16: 715–721, 2019. doi: 10.1038/s41592-019-0494-8.

Machado, G. M., Oliveira, M. M., and Fernandes, L. A. F. A physiologically-based model for simulation of color vision deficiency. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1291–1298, 2009.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4): e1003553, 2014. doi: 10.1371/journal.pcbi.1003553.

Somers, L. P., Franklin, A., and Bosten, J. M. Empirical tests of the effectiveness of EnChroma multi-notch filters for enhancing color vision in deuteranomaly. *Vision Research*, 218:108390, 2024. doi: 10.1016/j.visres.2024.108390.

Tregillus, K. E. M., Isherwood, Z. J., Vanston, J. E., Engel, S. A., and MacLeod, D. I. A. Color compensation in anomalous trichromats assessed with fMRI. *Current Biology*, 31(5):936–942, 2021.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200, 2016. doi: 10.1016/j.neuroimage.2015.12.012.

Wang, L., Mruczek, R. E. B., Arcaro, M. J., Bhatt, S., and Kastner, S. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, 25(10):3911–3931, 2015.

A. Supporting Analyses

A.1. Notation

Table 2. Key notation.

Symbol	Description
θ, θ'	Stimulus / distorted hue angle
$\mathbf{v} \in [0,1]^8$	LOCO vulnerability profile (per-hue interp. accuracy)
$W \in \mathbb{R}^{F \times V}$	Encoding weight matrix (F=6 channels $\rightarrow V$ voxels)
$\Delta\lambda$	Retinal cone-shift magnitude (nm)
$g \in [0,3]$	R+C cortical compensation factor (1: none; 2: full; $>$ 2: over)
β_s, β_c	S-cone / confusion-axis angular dilation ($^\circ$)
θ_{conf}	Confusion axis (16° protan; 150° deutan)
ΔRDM	Pairwise distance deviation from HC mean

A.2. Per-Subject ΔRDM Fits

Figure 4 shows, per fit-recovered participant, the observed and 2-Component-predicted ΔRDM (CVD minus HC mean) at the subject's affected ROI, with the observed-versus-predicted scatter over the 28 upper-triangle entries. The model recovers the correct *direction* of the dominant pairwise deviation on held-out folds, but the in-sample observed-versus-predicted cosine is modest (0.48 for Sub-09 protan at V1; 0.16 for Sub-08 deutan at V2): the 45° categorical quantization of the RDM loss makes the fit direction-consistent but magnitude-unresolved, consistent with the identifiability limits of Section A.5. Selection therefore rests on held-out generalization—clearing the no-correction null on all leave-one-HC-out folds (Section 3.2)—not on raw RDM cosine.

A.3. Implementation Details (Reproducibility)

Encoding model. The forward encoder is a half-wave-rectified squared-cosine basis with $F = 6$ channels at preferred hues spaced 60° apart (Brouwer & Heeger, 2009). The channel-to-voxel weight matrix W ($\hat{X} = CW^\top$) is fit by ridge regression with the regularization parameter selected by generalized cross-validation (Golub et al., 1979). Decoding reconstructs channel activations from a held-out voxel pattern as $\hat{\mathbf{c}} = W\mathbf{x}$ and returns the hue whose basis vector has the highest Pearson correlation with $\hat{\mathbf{c}}$; the same encoder is used for both cross-validation schemes below.

Cross-validation and metrics. *LORO* (discrimination): for each of the 6 runs, W is trained on the remaining 5 runs and the held-out run's 8 hues are decoded by nearest-neighbour matching; performance is 8-class exact accuracy (chance $1/8$). *LOCO* (interpolation): each of the 8 hues is held out in turn, W is trained on the remaining 7 hues pooled across all 6 runs (42-sample design), and the held-out hue is predicted in channel-output space; the primary metric is *adjacent accuracy*, the proportion of predictions within ± 1 hue step ($\pm 45^\circ$) of the target (chance $3/8$). The 8 per-hue adjacent accuracies constitute the vulnerability profile $\mathbf{v}$. *LOSO* (cross-subject generalization): W is trained on the remaining HC participants' SRM-aligned data and tested on the held-out participant.

Alignment and RDM. Run-level Procrustes alignment precedes a shared response model (SRM) (Chen et al., 2015); the reduced dimension k per ROI ($k = 4, 4, 3, 3$ for V1, V2, V3, hV4) was selected by leave-one-subject-out RDM reliability. Geometric disparity uses a leave-one-out-consistent HC reference (each fold's reference is the mean of the remaining 6 HC, against which both the held-out HC and the CVD participant are compared). The 8×8 crossnobis RDM (Walther et al., 2016) and $\Delta\text{RDM} = \text{RDM}_{\text{CVD}} - \overline{\text{RDM}}_{\text{HC}}$ are computed in this aligned space.

Statistics. Individual CVD participants are compared to the HC distribution with the Crawford & Howell (1998) modified single-case t ($df = 6$, one-tailed). Above-chance interpolation and HC–CVD comparisons use color-label permutation (1,000 iterations) and subject-label permutation (10,000 iterations) as appropriate. False-discovery-rate correction is applied within the pre-specified RDM color-pair and identifiability families; exploratory per-hue tests are reported uncorrected.

Code and data availability. The analysis code and two executable notebooks that regenerate every quantitative value reported in this paper—each compared against the manuscript in a per-number re-

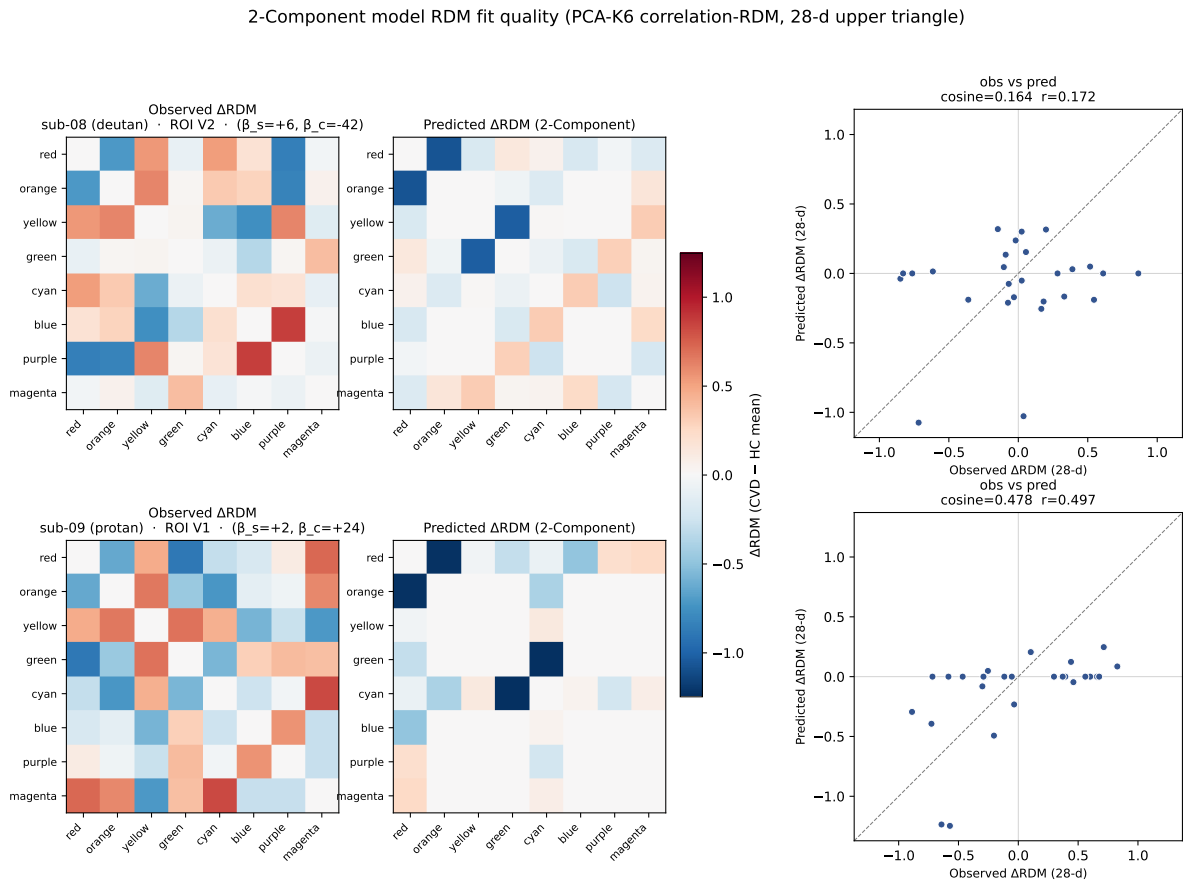

Figure 4. Per-subject ΔRDM fit (2-Component). For Sub-08 (deutan, V2) and Sub-09 (protan, V1): observed ΔRDM, predicted ΔRDM, and the observed-versus-predicted scatter over the 28 upper-triangle entries (cosine similarity annotated). ΔRDM is the CVD-minus-HC-mean pairwise distance.

production report—are available at https://github.com/Transconnectome/colorBlind_analysis (docs/ICML_workshop/repro_notebooks). The large per-resample fit outputs behind the modal-parameter cells are available from the authors on request.

A.4. Loss Atoms and Selection Procedure

Loss atoms. The fitting objective draws on three families: (i) L_γ, a rank loss between the model-predicted distortion magnitude at each hue pair and the observed behavioral JND ratio $\gamma = \gamma_{\text{CVD}}/\bar{\gamma}_{\text{HC}}$ (variants: a subtype-focal pair γ_{focal} or the aggregate γ_{all}); (ii) L_{RDM}, the cosine dissimilarity between the simulated and observed 28-element ΔRDM (computable per ROI); and (iii) L_{LOCO}, the MSE between the simulated and observed vulnerability profiles. Only L_{LOCO} couples to the forward encoder; the production fits use a behavioral + neural-RDM composite, so the encoder enters fitting only through the LOCO precondition gate.

Selection. Candidate combinations spanned all non-empty subsets of the behavioral, neural-ΔRDM (per ROI), and hV4 LOCO atoms; for each, the atoms' grid-z-scored losses were summed and scaled by $1/\sqrt{n}$ (for n atoms) so that combinations of different size are comparable. Per-subject loss combination and parameters were selected by a three-gate procedure over a 5-train/2-test HC resample ($N = 300$) and a strict 7-fold HC leave-one-out, using $(\beta_s, \beta_c) \in [-90°, 90°]^2$ and $g \in [0, 3]$ for R+C: (1) a *directional precondition* admitting only atoms whose CVD loss exceeds the HC leave-one-out distribution in the expected direction (signed Cohen's $d \geq 0.5$); (2) a *boundary-saturation* gate rejecting combinations with $\geq 50\%$ of resample solutions at the grid boundary; (3) the *held-out test composite loss* (median, then IQR) as the primary and secondary criteria. The selected combinations were $\gamma_{\text{OY}} + L_{\text{RDM}}^{(\text{V2})}$ for Sub-08 and $\gamma_{\text{all}} + L_{\text{RDM}}^{(\text{V1})}$ for Sub-09, with parameter

stability reported in Table 1.

A.5. HC Specificity and Identifiability

We applied the full hV4 fitting pipeline to all 7 HC participants, each substituted in turn into the CVD slot, to characterize the grid search's behavior under the null. Two findings frame the parameters as descriptive. First, the grid search over-fits 8 data points: most individuals—healthy or not—admit some configuration achieving a nominally good fit, and HC participants with poorer baseline interpolation show the largest apparent fitting gain—the signature of regression to the mean rather than a genuine, transferable correction (the gain is, by construction, the improvement over each HC's own baseline, so it anti-correlates with that baseline). Second, under the canonical (behavioral + RDM) loss, pre-specified null testing places $0/3$ candidate fits past both null sources (a per-realization Bonferroni test and a color-label permutation), the HC pseudo-CVD distance does not exceed all HC references (rank 0.875, not 1.0), and a $(0, 0)$ origin-recovery null reveals a per-axis uncertainty floor of $\sim 20°/25°$. A separate parameter-recovery family—two fit-recovered subjects $\times$ three recovery tests—likewise yields $0/6$ surviving FDR correction. What remains reportable is therefore the mechanism class (sign quadrant), whose averaged loss surface is 2.1×–5.5× deeper than the HC null, together with the cross-criterion consistency of each subject's fitted sign; the absolute (β_s, β_c) magnitudes are not physiologically interpretable. This calibration is an honest descriptive diagnostic, *not* a selection rule: the pipeline targets individual profiling of known CVD cases, not population screening, which is precisely why selection rests on held-out predictive loss and structural adequacy (Section A.4) rather than on any in-sample significance threshold.

A.6. Structural Adequacy of Candidate Mechanisms

The retinal R+C mechanism lacks a degree of freedom orthogonal to the confusion axis, so its single compensation factor g is driven toward the grid ceiling (Sub-08 $g \to 3.0$; Sub-09 $g \approx 2.95$). For Sub-08 this trips the boundary-saturation gate (100% of resample solutions at the bound, $\geq 50\%$). For Sub-09 the boundary rate (41%) stays below the gate, so R+C is not rejected there by saturation; it is instead dominated by the 2-Component fit on held-out composite loss (-0.86 vs. -1.54; median), with its gain still pinned near the ceiling. Parameter clustering at grid bounds is a recognized model-inadequacy signal. The 2-Component model, by contrast, fits in the grid interior for both subjects, with the fitted sign quadrant consistent across the ΔRDM atom computed in multiple ROIs (V1–hV4) for Sub-08; for Sub-09 the class is metric-dependent (Section A.5). These are descriptive structural-adequacy observations, not significance claims.

A.7. Correction Feasibility and the Inadequacy of Cone-Shift Models

The angular 2-Component transform is bijective by construction, so an exact stimulus-space pre-image exists for every displayed hue: all 8 colors invert exactly for both subjects (Sub-08 mean $|\delta\theta| = 26.3°$; Sub-09 $16.2°$; residual $< 0.001°$), giving a directly implementable per-subject lookup table.

The retinal cone-shift accounts, by contrast, are structurally inadequate as filters. The R+C model is rejected by overcompensation boundary saturation, its gain pinned at the grid ceiling ($g \to 3.0$; Section A.6). Its retinal-only limit (Machado, the $g=1$ case) is moreover non-invertible for the protan participant: at that subtype severity ($\Delta\lambda = 13.5$ nm) its arc-compression maps green ($135°$), cyan ($180°$), and blue ($225°$) into a narrow $\sim 26°$ band (the eight shifted hues span only $\sim 96°$), so these three lose a unique pre-image and only 4 of 8 hues admit an exact pre-image, exactly where Sub-09 is most vulnerable. This arc-compression is a structural property of the retinal-shift form, independent of the fitting criterion, and disqualifies the retinal family for the stimulus-space filter task. That modeling the structural distortion directly inverts exactly for every hue, where the cone-shift models do not, is the core adequacy argument for the cortical account.

A.8. Model Comparison: Full Table

Table 3. Candidate-mechanism comparison by structural adequacy and held-out evidence. *Adequate?* marks whether the fit lies in the grid interior; R+C is driven to the g ceiling, failing the saturation gate for Sub-08 (100%) and, for Sub-09 (41%, below the gate), dominated by the 2-Component on held-out loss. Held-out evidence (2-Component only): median test composite loss and the number of strict 7-fold leave-one-HC-out folds beating the no-correction baseline ($\delta\theta$=0) on the ΔRDM criterion. Machado and R+C are the rejected/dominated alternatives; no in-sample significance values enter selection.

Subject	Model	DOF	(β_s, β_c) / g	Adequate?	Test loss	RDM folds
Sub-08	Machado	1	$\Delta\lambda$	dominated	—	—
	R+C	1	$g\rightarrow$3.0 (bdy 100%)	**no**	—	—
	2-Comp	2	$(+6, -42)$	**yes**	-2.36 ± 2.15	**7/7**
Sub-09	Machado	1	$\Delta\lambda$	dominated	—	—
	R+C	1	$g\approx$2.95 (bdy 41%)	**no**	—	—
	2-Comp	2	$(+2, +24)$	**yes**	-1.54 ± 1.42	**7/7**
Sub-10	—	—	—	null	—	—

R+C lacks a degree of freedom orthogonal to the confusion axis. Sub-10 is null at the LOCO and SRM levels.