Ancestral neural computations constrain the evolution of novel computations in simulated color vision networks

Nathan P. Buerkle[1,2] and Stephanie E. Palmer[1,2,3]

[1]Graduate Program in Neurobiology
[2]Department of Organismal Biology and Anatomy
[3]Department of Physics
The University of Chicago, Chicago, IL

For correspondence: nathan.p.buerkle@gmail.com
sepalmer@uchicago.edu

Abstract

Evolution is a constrained process that does not necessarily lead to globally optimal phenotypes, but instead can bias evolution towards more readily accessible phenotypes. Here, we use standard machine learning algorithms to simulate convergent evolution of tetrachromatic color vision from a trichromatic ancestor both within and across distinct phylogenetic lineages. Analysis of network performance shows that all trichromatic networks were capable of evolving tetrachromacy, but some networks were more amenable to evolutionary modification than others. Comparisons of hidden layer computations between groups of tetrachromatic networks that either shared or did not share the same trichromatic "ancestor" shows that ancestry severely constrains evolution into a restricted and predictable portion of the theoretically available computational state space. Overall, our simulations of color vision evolution suggest that phylogenetic history is an important aspect of the functional organization of neural circuits.

Keywords: neural computation, evolution, color vision, machine learning

Introduction

Animals exhibit a rich diversity of behaviors and perceptual capacities that evolve as adaptive responses to ecological pressures. However, the evolutionary processes that give rise to behavioral variability do not necessarily occur as isolated, *de novo* searches for optimal solutions (Fawcett et al., 2013; Parker and Smith, 1990). Instead, factors such as phylogenetic history (Hale, 2014), development (Smith et al., 1985), and evolvability (Wagner and Altenberg, 1996) have the capacity to constrain and bias evolutionary trajectories. Thus, locally optimal phenotypes may be readily accessible, while global optimums may be difficult or impossible to reach. Accounting for these constraints and studying the mechanistic basis of adaptation has been invaluable for better understanding evolutionary patterns for morphological traits (Brakefield, 2011), protein sequences (Storz, 2016), and gene regulatory networks (Stern, 2013). Direct application of these concepts to the brain and behavior, in contrast, has been relatively limited (but see Ding et al., 2019; Katz, 2007; Seeholzer et al., 2018; Tosches, 2017).

Due to the complex and highly integrated organization of neural circuits, evolutionary constraints are thought to be especially strong for the brain and behavior (Katz, 2011; Tierney,

1995). Explicit and implicit assumptions of optimal neural coding are common throughout neuroscience, but existing network configurations could constrain and bias new computations into non-intuitive, locally optimal solutions. For example, the well-studied jamming avoidance response in weakly electric fish is elegant in its computational implementation, but substantially simpler algorithms are theoretically possible (Heiligenberg and Rose, 1985; Rose, 2004). Comparative work suggests that ancestral neural circuit organization is an important factor biasing evolution towards the observed, relatively complex computation (Matsubara and Heiligenberg, 1978; Rose and Canfield, 1991). Thus, an appreciation and consideration of phylogenetic history and its role in shaping neural circuit structure and function could lead to greater insights into principles defining how nervous systems process information and generate behavior.

Motivated by the diversity in photoreceptor spectral tuning across insects, color vision represents an attractive system for studying the evolution of neural computation. The ancestral insect eye was most likely comprised of ultraviolet (UV), blue, and green photoreceptors capable of trichromatic color vision (Briscoe and Chittka, 2001). A common adaptation, most notably in numerous butterflies, is the addition of a fourth, red-sensitive photoreceptor that expands color vision to tetrachromatic (Koshitaka et al., 2008; Stavenga and Arikawa, 2006; Zaccardi et al., 2006). No ecological or selective pressures have been identified to explain which insects do and do not have tetrachromatic vision, raising the possibility that evolutionary constraints influence and possibly impede color vision expansion (Briscoe and Chittka, 2001). Evolving a new photoreceptor is genetically simple, and the sensory periphery is generally thought to be evolutionarily flexible (Bendesky and Bargmann, 2011; Briscoe, 2002; Frentiu et al., 2007). A novel photoreceptor alone, however, might not confer expanded and improved color vision. Instead, making use of the novel color dimension likely requires changes to processing circuits.

The relatively simple and well-described opponent coding mechanism underlying color vision makes it especially amenable to evolutionary questions. Opponent neurons compare different photoreceptor responses with spectrally antagonistic excitation and inhibition and arise early in visual circuits (Lee et al., 1989; Paulk et al., 2009). In theory, tri- and tetrachromatic vision require only 2 and 3 unique channels, respectively (Chittka, 1992, p. 1; Hurvich and Jameson, 1957; Vorobyev and Osorio, 1998), and this simplicity can facilitate comparisons across species (Fig. 1A). Moreover, despite this low dimensional computation, the available computational state space is large, as numerous unique combinations of opponent channel can lead to perceptually equivalent color discrimination (Chittka, 1992; Vorobyev and Osorio, 1998). Assuming discrete all-or-nothing inputs to an opponent neuron, this leads to an upper bound of 15 and 560 unique ways to implement tri- and tetrachromatic vision, respectively. Relaxing this assumption from discrete to continuous inputs leads to an effectively infinite computational state space for color opponent computation. This lack of constraint on how to implement opponent coding means that similarities between independent origins of tetrachromatic vision can more confidently be ascribed to evolutionary rather than computational constraints.

For this study, we adopted a theoretical approach to simulate the evolution of tetrachromatic vision from a trichromatic ancestor. By generating numerous trichromatic ancestors and numerous independent origins of tetrachromatic vision from each one, we were able to simulate convergent evolution both within and across distinct phylogenetic lineages. Comparisons of

network performance, the overarching computational structure of each network, and homologous hidden units between networks that either shared or did not share the same trichromatic ancestor allowed us to assess how phylogenetic history might constrain the evolution of neural computation. Analyses showed that some trichromatic networks were more amenable to color vision expansion than others and that ancestral computations biased tetrachromatic computations into a predictable and severely restricted portion of the available computational state space. Overall, our simulation results were broadly consistent with existing behavioral and neurophysiology data, color vision theory, and conceptual ideas in evolutionary biology.

Color Vision Model

Feed-forward, 3-layer neural networks were trained to discriminate monochromatic inputs using a standard backpropagation learning algorithm (Fig. 1, see methods for details). The input layer represented initial phototransduction in the eye with UV, blue, green, and red photoreceptors modeled as Gaussian filters on the input light spectrum (Fig. 1B). The output layer was a filter bank of narrowly tuned Gaussians wavelength "categories"(Fig. 1C) that mimicked wavelength selective neurons found in both insects (Kien and Menzel, 1977; Swihart, 1972) and primates (Schein and Desimone, 1990; Zeki, 1980). Networks had a single hidden layer with a sigmoid nonlinearity. Our analyses focused on networks with 30 hidden units, but results were qualitatively similar regardless of hidden layer size (Fig. S1). Networks responded to monochromatic light stimuli, with input layer responses scaled by a random, multiplicative luminance factor that removed stimulus brightness as a learnable cue (Fig. S1C).

Using this simple network design, we simulated color vision evolution using a two-stage training procedure (Fig. 1B). First, a network was initialized with random starting weights in all layers and pre-trained for trichromatic vision with an input layer that had UV, blue, and green photoreceptors. This trained network, mimicking a trichromatic ancestor, was then "mutated" by converting a subset of green photoreceptors to red, matching the known evolutionary history of butterfly photoreceptors (Briscoe, 2008). This mutated network was then retrained for tetrachromatic vision 100 unique times, which we viewed as biologically analogous to 100 species that independently evolved tetrachromatic vision from the same trichromatic ancestor. We generated and evolved 100 trichromatic networks to simulate distinct phylogenetic lineages (e.g. butterflies, spiders, birds, etc.), for a total of 10,000 evolved networks.

If ancestry constrains evolutionary trajectories, we expect evolved networks sharing the same trichromatic starting weights to be more similar than networks with different trichromatic starting weights. These potential constraints could be unique to a trichromatic ancestor, but they could also be more broadly applicable to any set of starting weights. To control for this possibility, we trained and analyzed an additional set of *de novo* networks. These networks lacked a pre-training step and were instead trained for tetrachromatic vision directly from random starting weights. Mirroring the evolved networks, we generated 100 sets of random starting weights and trained each starting point 100 unique times, again for a total of 10,000 *de novo* networks. Because of large sample sizes, nearly every statistical comparison was highly significant using a standard p value, so we instead adopted a stricter, effect-size based significance criterion of Cohen's $d > 0.5$ for most analyses.

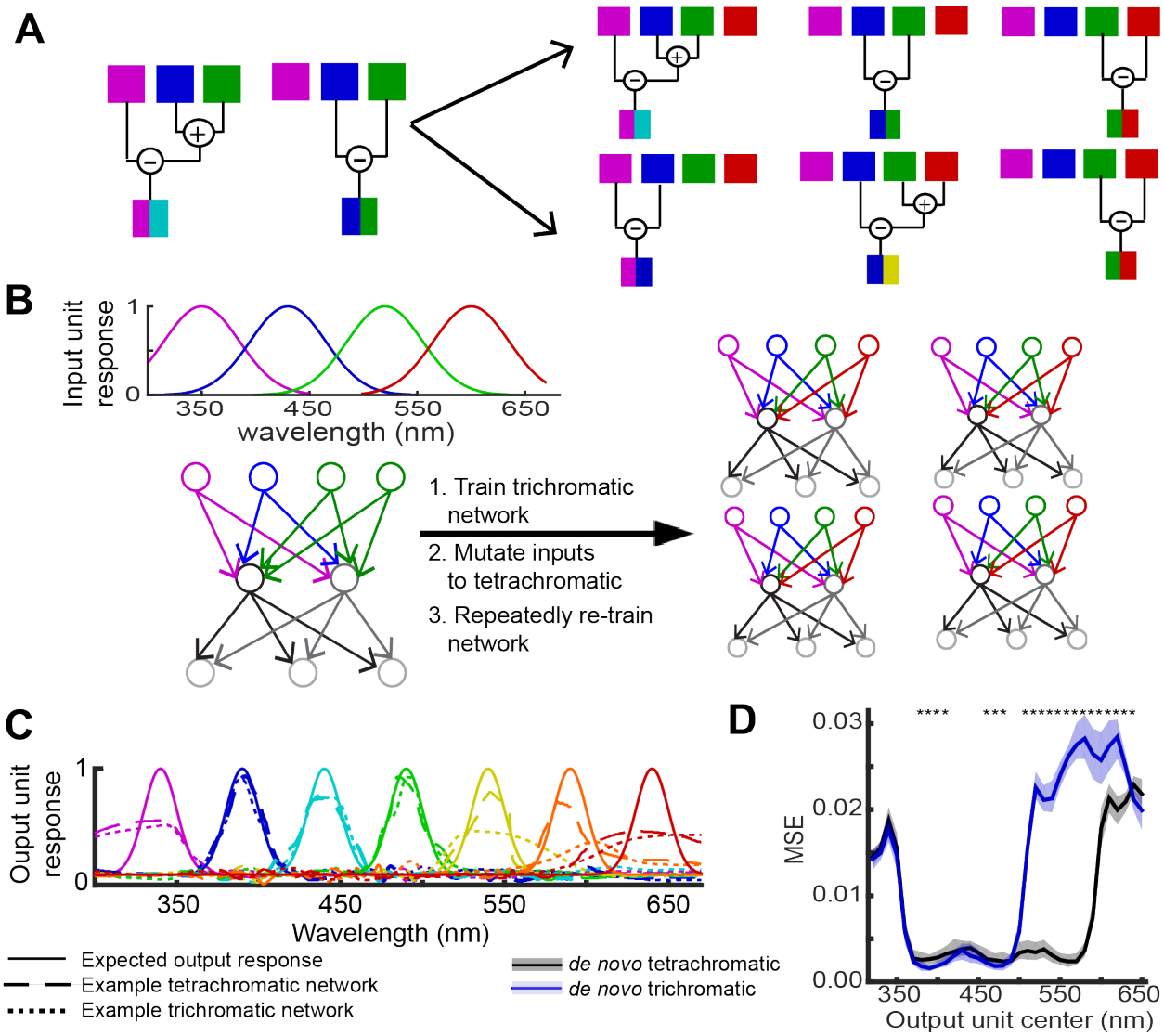

Figure 1: Network design

A. In theory, trichromatic color vision requires three photoreceptor types and two color opponent channels. The left shows one of numerous potential pairs of opponent channels, corresponding to UV vs. (B+G) and B vs. G. Tetrachromatic vision requires a fourth photoreceptor type and a third opponent channel (right). To evolve tetrachromatic computations from a trichromatic ancestor, a network could inherit the original channels and layer on a third (top) or evolve a completely novel set of opponent channels (bottom).

B. Color vision evolution was simulated using feed-forward networks and a two stage training procedure. The input layer comprised photoreceptors tuned to UV, blue, green, or red light and responded to monochromatic light stimuli (top). A trichromatic network lacking a red input was first trained from random starting weights (left). A green photoreceptor was then "mutated" to red and independently retrained multiple times using unique training data (right). Overall, we trained 100 trichromatic networks and "evolved" each one 100 times for a total of 10,000 evolved networks that simulated convergent evolution both within and across distinct phylogenetic lineages. A similar set of 10,000 *de novo* tetrachromatic networks were trained directly from random starting weights.

C. The output layer for each network was a filter bank of 34 narrowly tuned Gaussian curves that evenly tiled visual space. Tuning curves were generated by measuring the response of each output unit to monochromatic stimuli between 300 and 670 nm in 1 nm intervals. Depicted here are output tuning curves for a subset of output units for a representative trichromatic and tetrachromatic network. Performance was measured for each output unit individually as the MSE between the observed and expected response.

D. Network performance was measured for tri- (n=100) and tetrachromatic (n=100) networks trained from random starting weights. Shading shows the 25th and 75th percentile across all networks. Due to large sample sizes, we adopted a stricter, effect size based significance criterion throughout the paper, with asterisks indicating Cohen's d > 0.5.

Results

Tetrachromatic networks had an expanded range of good color vision

We first validated our network design and training protocol by comparing the performance of *de novo* tri- and tetrachromatic networks trained from random starting weights. Because trichromatic networks lacked a red photoreceptor, we expected the largest performance differences for long wavelength stimuli. Thus, rather than measuring overall network performance across the visual spectrum, we opted to analyze each output unit individually. For each output unit, we constructed a tuning curve by measuring its response to wavelengths spanning the full visual spectrum and averaged over different luminance factors (Fig. 1C). Performance was then defined as the mean squared error (MSE) between the expected and observed response.

Performance generally followed the Fisher Information of the input layer (Fig. S1A), which generally matches behavioral discrimination thresholds in insects and vertebrates (Koshitaka et al., 2008; Neumeyer, 1986; Valois and Jacobs, 1968; von Helversen, 1972). Output units that were tuned to wavelengths where input layer photoreceptors intersect typically had the best performance (i.e. lowest MSE). Slight decreases in performance between these Fisher Information peaks were primarily due to decreased output unit amplitude (Figs. 1C,S2). By design, tetrachromatic networks performed generally well across the entire visual range, while long wavelength output units for trichromatic networks performed poorly (Fig. 1D). This impaired performance arose because trichromatic networks lacked a red photoreceptor such that only green photoreceptors responded to long wavelength light. This made luminance a confounding variable, as a dim stimulus at the center of green tuning elicits the same response as a bright stimulus on the tail. In the absence of luminance variation, tri- and tetrachromatic networks performed similarly well for all output units (Fig. S1B). Notably, for short and middle wavelength output units, trichromatic networks performed slightly but significantly better than tetrachromatic networks, which likely reflected a smaller range of learnable stimuli but an equal number of hidden units.

Mutating the input layer impaired network performance

De novo networks demonstrated that tetrachromatic networks have an expanded range of color vision but did not address questions about color vision evolution. To simulate evolution, the input layer of fully trained, *de novo* trichromatic networks was "mutated" by converting a subset of green photoreceptors to red (Fig. 1B). This mutation, prior to any retraining, resulted in a mismatched tetrachromatic input layer and trichromatic hidden layer. This situation likely mirrors biology, where a novel photoreceptor necessarily precedes changes to central brain circuitry. Output units tuned to short wavelengths were unaffected by the mutation, while performance was significantly impaired for middle and long wavelength output units (Fig. 2A). Interestingly, the effect of mutation was similar across all networks (Fig. 2A, note error bar size), suggesting that changes to peripheral sensing caused similar disruptions to sensory perception regardless of the specific computations implemented by each network.

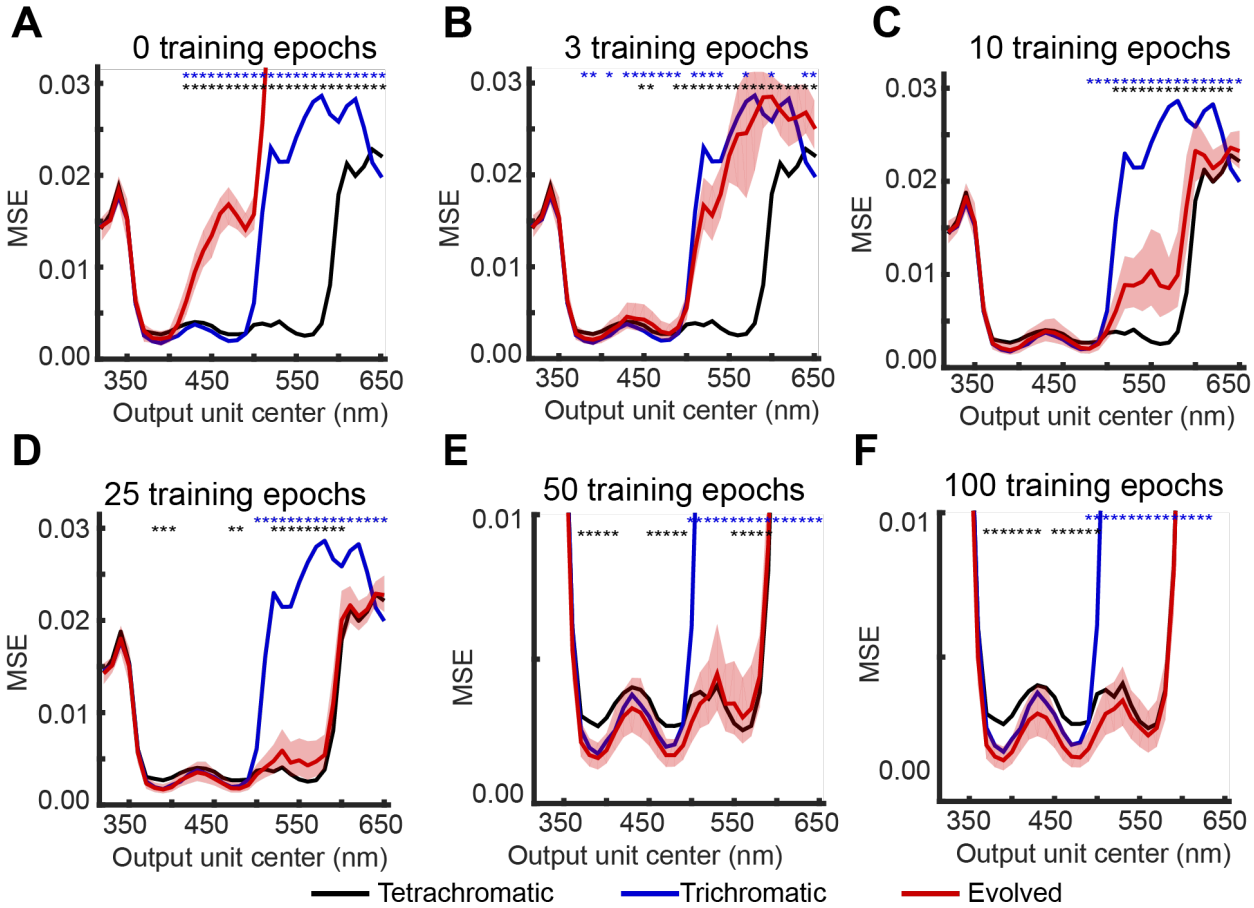

Figure 2: Evolved network performance over training time
The performance of evolved networks was tracked over training time and compared to *de novo* tri- and tetrachromatic networks trained from random starting weights after A) 0, B) 3, C) 10, D) 25, E) 50, and F) 100 training epochs. 50 training epochs matches the total training time of the *de novo* networks. Shading shows the 25th and 75th percentile and was omitted for *de novo* networks for clarity. Blue and black asterisks show significant differences between evolved and either tri- or tetrachromatic networks, respectively, with Cohen's d > 0.5.

The impaired MSE performance of mutant networks could arise through a variety of changes to the output unit tuning curves. To uncover which specific changes affected performance, we fit Gaussians to output unit tuning curves and compared the fit parameters between networks (Fig. S2). For *de novo* tri- and tetrachromatic networks, the tuning centers, widths, and amplitudes broadly matched the target. For mutant networks, output units maintained narrow, single peaked tuning curves with widths and amplitudes that generally matched the original trichromatic network. Tuning centers, in contrast, shifted by 5-10 nm for middle wavelength output units and was the primary source of the MSE impairment. These relatively minor shifts suggest that the effect of introducing a novel photoreceptor on color vision may be smaller than the MSE performance metric initially revealed, with mutant networks likely retaining trichromatic color vision. However, even small perceptual deficits could lead to selection against the expression of a novel photoreceptor in a natural population.

Every network successfully gained tetrachromatic performance

Peripheral sensory systems are thought to be an especially flexible site for evolutionary change that can lead to adaptive behavior without large-scale changes to central processing (Bendesky

and Bargmann, 2011). The seemingly minor fitness cost associated with the introduction of a novel photoreceptor in our mutant networks was consistent with peripheral structures being evolutionarily labile. However, a novel red photoreceptor alone was insufficient to expand the range of color vision in the network, indicating that further modifications to color computation were necessary. To ask whether ancestral trichromatic circuits were capable of adapting to a mutated input layer, we next re-trained mutant networks and tracked performance over training time (Fig. 2).

Networks compensated for the impaired performance of mutant networks quickly, but improvement to tetrachromatic vision proceeded slowly. Within 3 re-training epochs, evolved network performance broadly matched the performance of the original trichromatic networks (Fig. 2B). Long wavelength output units continued to improve over training, becoming effectively tetrachromatic within 25-50 training epochs, which matched the total training time of *de novo* networks (Fig. 2D,E). Interestingly, short and middle wavelength output unit performance matched trichromatic networks, which was significantly better than tetrachromatic networks. However, this improvement came at the expense of long wavelength output unit performance, which remained slightly but significantly worse than *de novo* networks. Extending training to 100 epochs improved this performance to match the long wavelength output unit performance of *de novo* networks (Fig. 2F).

Different starting networks learn at different rates

With a sufficient number of retraining epochs, every evolved network performed similarly to *de novo* tetrachromatic networks, indicating that trichromatic ancestry did not prohibit the eventual evolution of tetrachromacy. However, rather than strictly prohibiting an adaptive phenotype, evolutionary constraints are primarily thought to affect how accessible an adaptive phenotype is (Smith et al., 1985). For our networks, this could mean that some trichromatic network configurations were amenable to modification while others resisted adapting to the novel red photoreceptor. To investigate this possibility, we compared the performance of networks derived from the same trichromatic starting point after 10 retraining epochs.

Visual inspection of network performance after 10 training epochs suggested that starting weights influenced learning rate (Fig. 3A). Some trichromatic networks facilitated evolution, with the average performance approaching tetrachromatic performance after 10 epochs (Fig. 3A). Other trichromatic networks impeded evolution, with descendant networks showing minimal performance improvements beyond the original trichromatic network (Fig. 3A). Starting weights appeared to similarly affect learning for *de novo* networks trained from random starting weights. Regardless of starting weights, *de novo* networks tended to perform well at Fisher Information peaks, while performance between peaks were more variable (Fig. 3A). This result was reminiscent of animal behavior, where discrimination thresholds near Fisher Information maxima are often similar across species, while thresholds between these regions of best discrimination vary more substantially (Koshitaka et al., 2008; Telles et al., 2016).

We quantified these effects by separating networks into groups of networks that either shared or did not share the same starting weights, resulting in four classes: 1) evolved networks with the

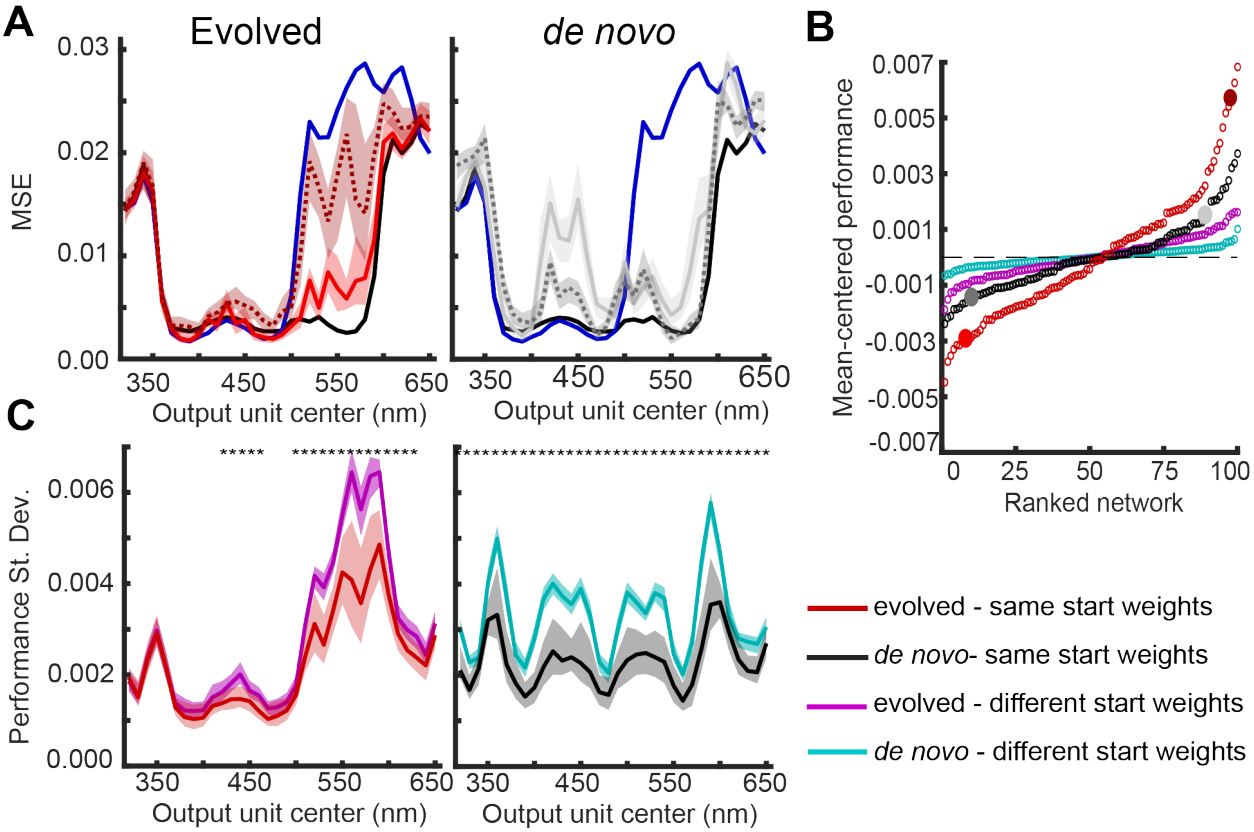

Figure 3: Starting weights influence learning rates

A. Networks were separated into groups of 100 that shared the same starting weights. Each panel shows the average performance after 10 training epochs for two groups of evolved networks trained from trichromatic starting weights (left) and *de novo* networks trained from random starting weights (right). Shading represents the 25th and 75th percentile of performance across the 100 networks sharing the same starting weights..

B. Performance after 10 training epochs for each group of 100 networks that either shared or did not share the same starting weights was reduced to a single value by averaging across networks and long wavelength output units. Distributions for each class were mean-centered and compared using a two-tailed Kolmogorov-Smirnov test. All pairwise comparisons were significantly different ($p < 0.001$). Large filled circles correspond to the networks shown in panel A.

C. The standard deviation of performance was calculated for each output unit within a group of 100 networks. Shading shows the 25th and 75th percentile across the 100 groups within a class. Asterisks denote significant differences with Cohen's $d > 0.5$.

same trichromatic starting weights, 2) evolved networks with different trichromatic starting weights, 3) *de novo* networks with the same random starting weights, and 4) *de novo* networks with different random starting weights. Since we started with 100 sets of starting weights, each class comprised 100 unique groups with each group containing 100 networks.

Group performance was summarized by a single value: the average performance across all networks and all long wavelength output units (> 500 nm), within a group. The distribution of this performance metric for each class was then mean-centered and compared to the other classes (Fig. 3B). These distributions were significantly different for every pairwise comparison between classes (Kolmogorov-Smirnov test, $p < 0.001$). The performance distribution for groups of evolved networks sharing the same starting weights was the most variable, suggesting that trichromatic ancestry imposed stronger constraints than random starting weights would have predicted.

Group differences in mean performance could emerge in two different ways. First, every network derived from the same set of starting weights could follow a similar learning trajectory. Alternatively, networks sharing the same starting weights could have a broad range of performance, with group differences caused by variability in the proportion of networks that learn quickly or slowly. To distinguish between these options, we calculated performance variability for each output unit within a group of 100 networks. For both evolved and *de novo* networks, groups of networks sharing the same starting weights showed significantly less performance variability than groups with different starting weights (Fig. 3C). However, the distribution of variability differed substantially between evolved and *de novo* networks. Because evolved networks were pre-trained to perform well for short and middle wavelength output units, variability was mostly confined to long wavelength output units. *De novo* network variability was more evenly distributed across all output units, although it was lower for output units corresponding to Fisher Information maxima.

Together, network performance results showed that starting weights affect the early evolutionary response to a novel photoreceptor but do not strictly prohibit evolution. Mutant networks compensated for the modified input layer quickly, but learning new computations was comparatively slow. Given our gradient descent learning algorithm, this result was expected with enough training epochs. Interestingly, pre-training networks with a trichromatic input layer facilitated evolved network performance that exceeded expectations set by networks trained from random starting weights. This result matches ideas both in biology and machine learning about curriculum learning, where learning through the iterative addition of complexity facilitates improved learning outcomes (Bengio et al., 2009). In this way, constraints imposed by an ancestral circuit might be beneficial in some evolutionary contexts.

Starting weights constrained the overarching computational structure of the hidden layer

Every network successfully evolved tetrachromacy, which we viewed as biologically analogous to convergent evolution both within and across distinct phylogenetic lineages. A convergent phenotype alone, however, is not sufficient to conclude that starting weights constrained the network. Instead, this might reflect similar selective pressures (i.e. the same training cost function). Differences in learning rate hinted at a starting point constraint, but a rigorous test required assessing the degree of similarity in the underlying computational implementation of color vision across networks. Thus, we next turned to analyzing the hidden layer, focusing on the weights connecting the input layer to the hidden layer. The high dimensionality of the hidden layer posed an analytical challenge, which we first approached using principal component analysis (PCA) to examine the overarching structure of hidden layer computations.

Consistent with color vision theory (Chittka, 1992; Hurvich and Jameson, 1957; Vorobyev and Osorio, 1998), PCA reduced the dimensionality of the hidden layer from 30 hidden units to 3 color opponent channels (i.e. eigenvectors) that described 48.5 ± 12.2, 27.6 ± 5.5, and 16.0 ± 5.7 percent of the variance (92.1% in total) in hidden unit computations. Each channel had a UV, blue, green, and red component that consistently showed opponent interactions, as indicated by a combination of positive and negative input weights (Fig. 4A,S3). To assess computational similarities, networks were separated into groups of 100 that either shared or did not share the

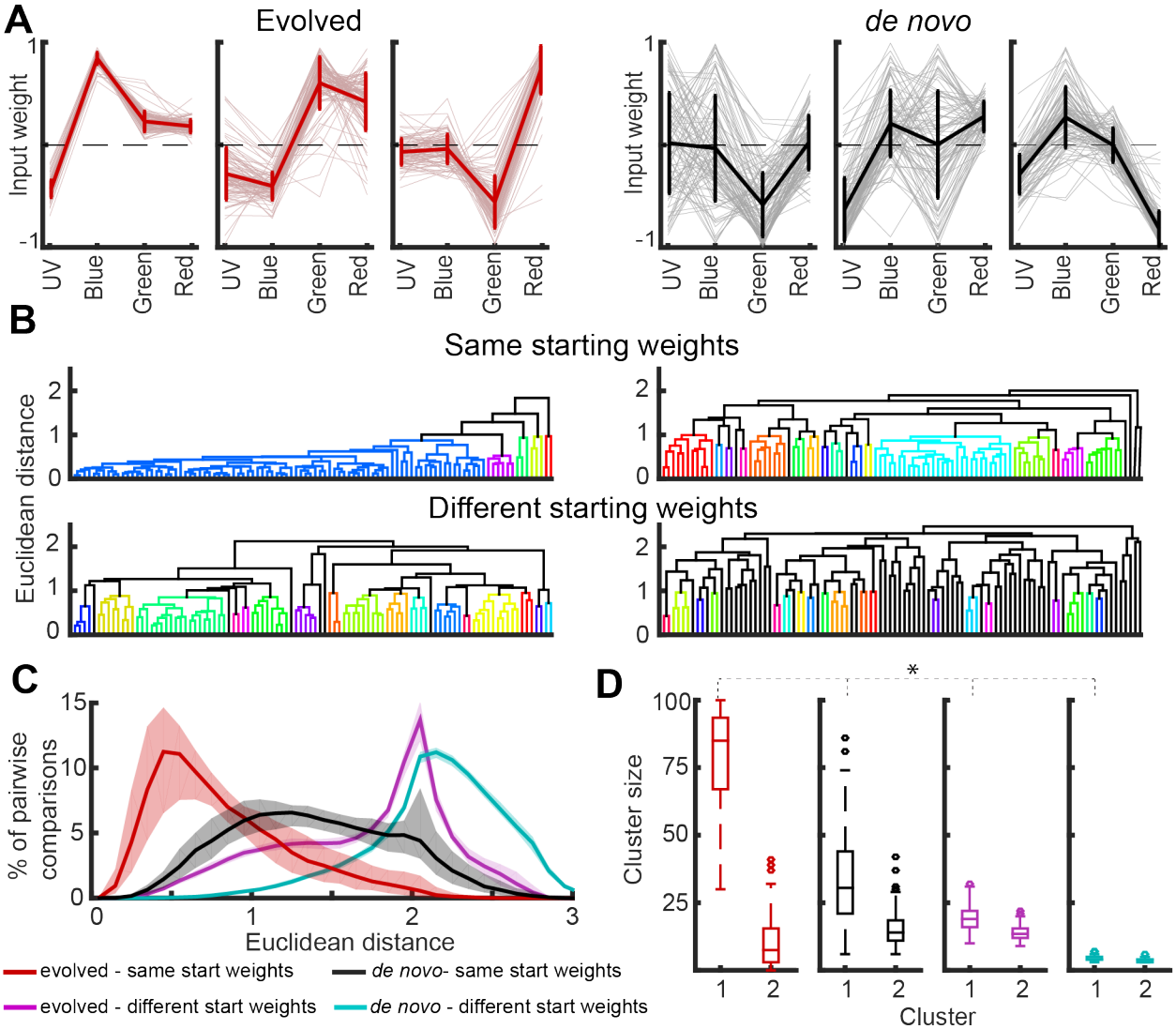

Figure 4: Starting weights constrain the computational structure of the hidden layer
A. PCA on the weights connecting the input layer to the hidden layer revealed 3 opponent channels (i.e. eigenvectors) that described the overarching computational structure of a network. Input weights with opposite signs are indicative of an antagonistic opponent interaction between photoreceptor types. The left three panels show the opponent channels for 100 evolved networks derived from the same trichromatic starting point. The right three panels similarly show the opponent channels for 100 *de novo* networks derived from the same random staring weights. Thick lines and error bars are mean ± standard deviation.
B. Opponent channels for groups of 100 networks that either shared or did not share the same starting weights were analyzed using hierarchical clustering. Depicted are representative dendrograms for each of the four classes of networks. The top row corresponds to the networks shown in panel A. Different colors represent different clusters with a clustering threshold of 1.0.
C. Euclidean distances between opponent channels were measured for every pair of networks within a group of 100 and binned in intervals of 0.1. Shading shows the 25th and 75th percentile.
D. Groups of networks were clustered using a distance threshold of 1.0. Boxplots show the distribution of cluster sizes for the two largest clusters for each class of groups.

same starting weights. The three opponent channels for each network were concatenated into a single 12 dimensional vector (3 channels X 4 input weights) and analyzed using hierarchical clustering (Fig. 4).

We first analyzed groups of *de novo* networks where each network had different random starting weights. Networks in these groups used diverse sets of opponent channel, spanning the full range

of color opponent combinations (Fig. 4B,S3A). Networks rarely clustered together, with the largest cluster containing between 3 and 7 networks out of 100 (Fig. 4D). We also looked at the distribution of Euclidean distances between every pair of networks in a group, and consistent with small cluster sizes, pairwise distances tended to be large (Fig. 4C). Moreover, this distribution of pairwise distances closely matched a null distribution created with opponent channels generated randomly (Fig. S3C, Jensen-Shannon Divergence = 0.02). This result suggested that our network design and training protocol were not biased and utilized the full computational state space.

We next compared groups of *de novo* networks sharing the same random starting weights to the previously analyzed groups with different random starting weights. The distribution of pairwise distances for these groups was shifted substantially towards zero (Fig. 4C, JSD = 0.31). These smaller distances led to significantly increased clustering, with the largest cluster containing an average of 33.8 ± 17.1 networks (Fig. 4D, Cohen's d = 2.4). Thus, starting weights did constrain and bias network evolution, but *de novo* networks also retained some diversity in how tetrachromatic computations were implemented.

Trichromatic starting weights constrained computations significantly more than random starting weights. Groups of evolved networks sharing the same starting weights typically converged on only a few computational motifs. Pairwise distances were shifted further towards 0 compared to *de novo* networks with the same starting weights (Fig. 4C, JSD = 0.19), and the largest cluster contained 79.2 ± 17.7 networks (Fig. 4D, Cohen's d = 1.9). Only 2.5 ± 2.1 clusters were needed to account for 90 out of the 100 evolved networks in a group, whereas *de novo* networks sharing the same random starting weights required 12.3 ± 7.9 clusters (Cohen's d = 1.7). Together, these results showed that starting weights constrained network computations, and trichromatic weights imposed especially strong constraints that biased networks into a severely restricted region of the available computational state space.

Evolved networks inherit the original opponent channels and consistently add the same novel channel

Analyzing groups of evolved networks with different trichromatic starting weights showed evolutionary trajectories for color computation were predictable. The largest cluster had 19.4 ± 4.6 networks, and 13.7 ± 2.9 clusters were required to account for 90 networks (Fig. 4D). Thus, these groups were most comparable to, but significantly less similar than groups of *de novo* networks with the same random starting weights (Cohen's d = 1.2). For these two classes, the pairwise distance distribution was visually consistent with the clustering differences (Fig. 4C), but the distributions were statistically similar (JSD = 0.05). Interestingly, this pairwise distance distribution matched the distance distribution of the original trichromatic networks visually and statistically (Fig. S3B, JSD = 0.01). This similarity was surprising because one might assume the smaller dimensionality of trichromatic networks that have only two opponent channels with three input weights would lead to smaller distances.

To understand how this similarity arose, we analyzed each opponent channel individually rather than as the concatenated group of three (Fig S3D). We first noticed that the new, third channel for nearly every evolved network had green vs. red opponency regardless of the specific

trichromatic starting weights (Fig. S3A), and hierarchical clustering of the third channel confirmed this similarity (Fig. S3D). In contrast, the first and second channel matched previous results (Fig. 4), with groups sharing the same starting weights more similar than groups with different starting weights. Calculating the Euclidean distance between these first two evolved channels and the original two trichromatic channels they evolved from allowed us to ask how much training affected opponent channel tuning. Evolved network distances (0.33 ± 0.20) were significantly smaller than *de novo* network distances (0.55 ± 0.20, Cohen's d = 1.1) or evolved networks with shuffled starting channels (0.66 ± 0.28, Cohen's d = 1.3).

Overall, these results indicated that networks evolve tetrachromatic vision by inheriting the original two opponent channels and adding a third, orthogonal channel specifically implementing green vs. red color opponency. This result is consistent with how networks performed, as limited disruption to short wavelength computations led to limited disruptions in short wavelength performance. Thus, maintenance of this performance may play a causal role in biasing tetrachromatic computations. *De novo* networks, in contrast, were freer to vary and find unique computational solutions that spanned the full computational state space.

Learning targets a specific subset of evolved network hidden units

The opponent channels revealed by PCA describe the combined tuning of a population of individual hidden units. Similar opponent channels between networks sharing the same starting weights could emerge from similar modifications to hidden unit tuning. This would be consistent with an adaptive hotspot hypothesis of evolution that proposes circuit nodes can vary in how evolutionarily flexible they are (Ding et al., 2019; Katz, 2011). Alternatively, evolved networks could exhibit a wide diversity of tuning despite opponent channel similarities. For example, distantly related South American and African electric fish often use similar electro-sensation computations, but the specific anatomical locus of these computations can vary substantially (Rose, 2004).

To distinguish between these options in our networks, we compared the tuning of individual hidden units before and after training using the cosine distance. This distance metric specifically measures changes in opponent tuning while ignoring changes in overall weight size (Fig. 5A). Because each trichromatic network was independently evolved 100 times, this analysis generated a distribution of distances for each hidden unit (Fig. 5b). For each trichromatic network, its 30 hidden units were ranked in order of the median distance or the distance variability across the 100 evolved networks. In parallel, the same analysis was conducted on the *de novo* networks and results were compared between evolved and *de novo* networks (Fig. 5).

In contrast to opponent channel results, evolved network cosine distances were significantly larger (Fig. 5D) and more variable (Fig. 5E) than *de novo* networks. Large cosine distances for 10-15 hidden units per network were typically associated with the development of green vs. red opponency in hidden units that originally had small weights and weak or no opponency (Fig. 5C, yellow and magenta). In contrast, hidden units with the smallest cosine distances always had large input weights and clear opponent tuning (Fig. 5C, green). Performing the same analysis on networks with 50 hidden units similarly found 10-15 hidden units with especially large distances

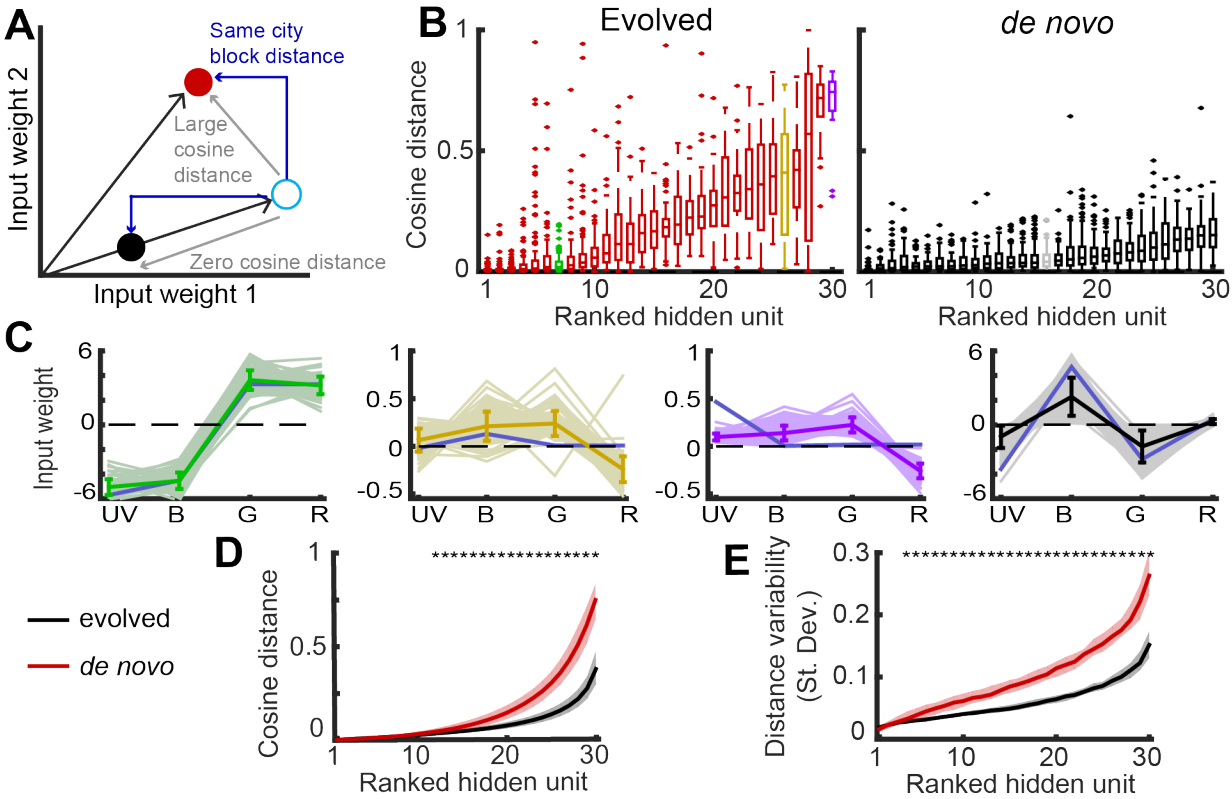

Figure 5: Training leads to larger changes in the tuning of evolved network hidden units

A. Schematic diagram shows the two types of distance used to measure hidden unit changes. The total length of the blue arrows is the city block distance, and the distance of each filled circle to the open circle is the same. The cosine distance measures the relative change in input weights while ignoring differences in overall weight magnitude. Because the black circle and open circle lie on the same vector, the cosine distance is 0, while the distance between red and the open circle is large.

B. Tuning changes for each hidden unit were measured as the cosine distance between the input weights of a hidden unit before training and the input weights after training. Boxplots show the distribution of cosine distances for 100 evolved (left) and *de novo* (right) networks derived from the same starting weights and ranked in order of median distance. Colored boxes correspond to plots in panel C.

C. Each panel shows the tuning of an example hidden unit, with colors corresponding to the colored boxes in panel B. The solid line and error bars are the mean ± standard deviation. The blue line shows the tuning of the original, untrained hidden unit. Note differences in the scale of the y-axis.

D. Cosine distance distributions were measured for each of the 100 starting networks and ranked in order of increasing cosine distance. Shading shows the 25th and 75th percentile. Asterisks show significant differences between evolved and *de novo* networks with Cohen's d > 0.5.

E. Same as D, but rather than median cosine distance, the standard deviation of the cosine distance distribution was measured.

(Fig S4). The extra hidden units compared to 30 hidden unit networks had the smallest distances and varied little between networks. This result was broadly consistent with an adaptive hotspot explanation for network evolution. Many hidden units were rigidly constrained and did not vary, and tuning curves of these hidden units suggest an important role in short wavelength discrimination. A small subset of hidden units, in contrast, was flexible and adapted to expand and improve color vision. As expected, green vs. red opponency appeared in these hidden units, but surprisingly, these computationally flexible hidden units could vary substantially relative to each other.

Hidden unit outputs are robust to input variability

In addition to having large cosine distances, evolved network hidden units also showed high distance variability (Fig. 5B,E). Differences in the specific opponent tuning of homologous hidden units could be an important factor in how different networks function, but the variation could also reflect network computations that were robust to changes. To distinguish between these possibilities, we generated tuning curves for the output response of the hidden layer. Our previous analyses focused on the input weights connecting each of the four input layer photoreceptors to the hidden layer, biologically analogous to looking at the dendrites of a neuron. Before connecting to the filter bank output layer, however, hidden unit responses are passed through a sigmoid non-linearity that transforms responses to values between 0 and 1, which is biologically analogous to the axon. We measured output tuning curves for wavelengths spanning the full visual range and compared the tuning of homologous hidden units using hierarchical clustering (Fig. 6).

The sigmoid non-linearity typically made hidden unit responses robust to differences in input tuning. In general, hidden unit outputs exhibited binary 0 or 1 responses with relatively sharp transition zones (Fig. 6A). Because of this non-linearity, large differences in input tuning could lead to only minor changes in output tuning. One exception to this effect, however, occurs when an input weight flips between positive and negative. In these instances, minor differences in input tuning between two hidden units can be amplified into large differences in output tuning (Fig. 6a, red). For evolved network hidden units, output response differences between homologous hidden units were almost exclusively observed for long wavelength stimuli where green vs. red opponency was developing. Differences between homologous *de novo* hidden units, in contrast, could exhibit these zero-crossing effects across the entire visual spectrum.

Surprisingly, despite larger and more variable cosine distances for input tuning, output responses for homologous hidden units in evolved networks were more similar than *de novo* networks. Visual inspection of the pairwise distance distribution suggested that output responses for evolved network hidden unit were substantially more similar than for *de novo* networks, but statistically, the distributions were relatively similar (JSD = 0.04, Fig. 6B). Homologous hidden units derived from the same trichromatic hidden unit were then clustered with a distance threshold of 3.0. For each trichromatic network, its 30 hidden units were ranked in order of how many of its evolved network homologs were in the largest cluster. The same clustering analysis was performed on *de novo* networks and then compared to the evolved networks.

Within a network and its 30 hidden units, individual hidden units varied substantially in how similar the output response was after training (Fig. 6C). The output response of some hidden units was rigidly constrained, with the largest cluster containing that hidden unit for every network. Other hidden units showed substantial variability, with as few as 8 networks contained in the largest cluster. A comparison of evolved and *de novo* networks showed that, although evolved network input responses were more variable, hidden unit output responses were less variable than *de novo* networks (Fig. 6C). The largest cluster for evolved networks was significantly larger than *de novo* networks for 27 of the 30 hidden units (Cohen's d > 0.5), with 19 having a Cohen's d > 1.5. Overall, these results broadly match the opponent channel results

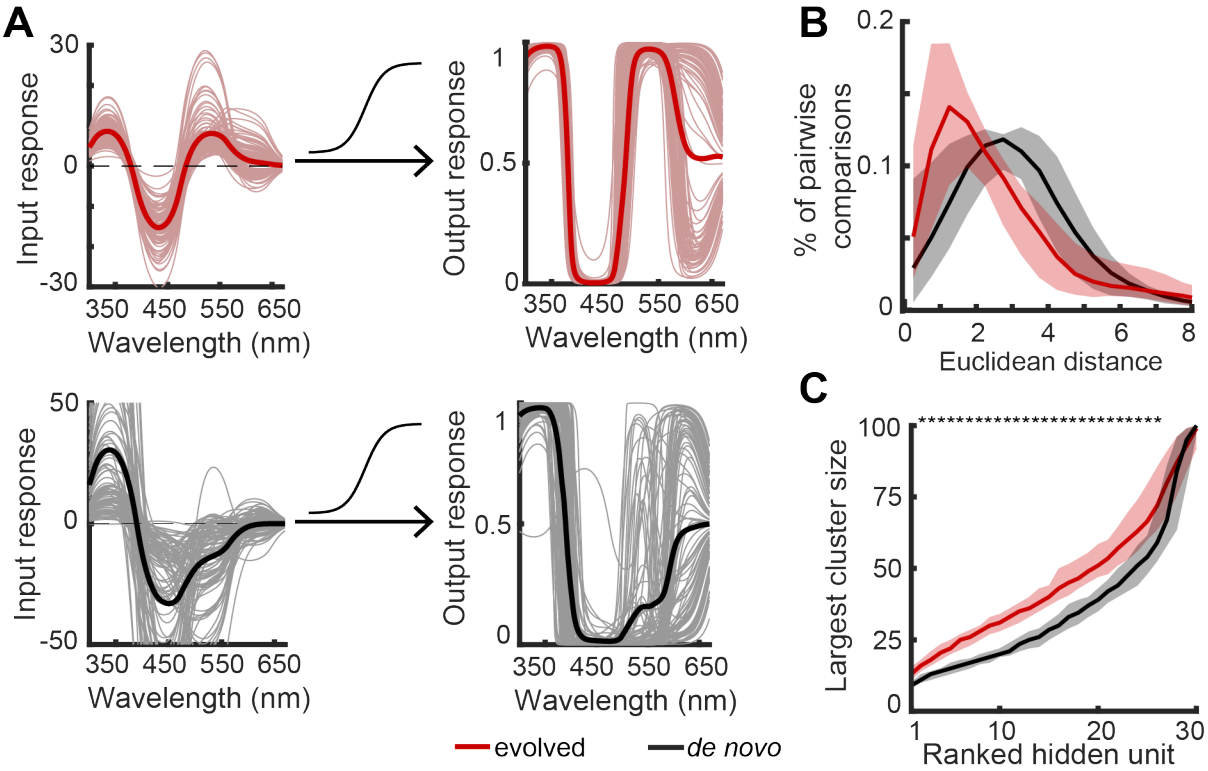

Figure 6: Output responses for evolved network hidden units are robust to changes in input tuning
A. Input responses to a hidden unit, which were analyzed in prior figures, are passed through a sigmoid nonlinearity before connecting to the output layer. Hidden layer responses were calculated for monochromatic stimuli from 300 to 670 nm in 1 nm increments. The left panel shows the input response of two example hidden units. The right panel shows the output response of the same hidden units.
B. The output responses of homologous hidden units were compared using hierarchical clustering. Data shown here is combined across 100 trichromatic networks that each had 30 hidden units. Shaded areas show the 25th and 75th percentile across all hidden units.
C. For each trichromatic network, hidden units were ranked in order of cluster size for the homologous hidden units in the evolved networks. Output responses were clustered with a distance threshold of 3.0. Shaded areas show the 25th and 75th percentile. Asterisks show significant differences between evolved and *de novo* networks with Cohen's d > 0.5.

showing that evolved networks were more similar than *de novo* networks. Importantly, our PCA analysis was independent of the subsequent sigmoid non-linearity, suggesting that the opponent channel analysis captured computationally relevant tuning differences.

De novo networks vary connection strength more than evolved networks

In addition to changes in opponent tuning, hidden units could also learn by increasing or decreasing their overall strength of connection. Proportional changes to input weight sizes appeared to be the primary way that connection patterns for *de novo* network hidden units changed through training. In contrast to evolved networks, hidden units for *de novo* networks had small cosine distances, but city block distances were significantly larger and more variable (Fig. 5C,7A). To examine this weight magnitude difference between evolved and *de novo* networks in greater detail, we quantified connection strength as the sum of the absolute value of the four input weights to each hidden unit. Connection strength of a hidden unit was then compared before and after training as well as between homologous hidden units (Fig. 7).

For all networks, training almost exclusively led to decreases in hidden unit connection strength (Fig. 7B). Random starting weights were large and varied little between hidden units. Training these random starting weights with a tetrachromatic input layer decreased connection strength, leading to a relatively uniform distribution. Networks with only a trichromatic input layer decreased connection strengths substantially more, with 10-12 out of 30 hidden units having connection strengths close to 0. Evolved networks derived from these trichromatic networks suppressed weights further, with 16-18 hidden units having connection strengths near zero.

This extra decrease in connection strength for trichromatic networks was reminiscent of L2 regularization, which is commonly used in machine learning to prevent over-fitting. This effect was specifically due to trichromatic networks having long wavelength output units that were, by design, unable to successfully learn (Fig. S5). This apparent regularization effect could potentially explain why trichromatic networks outperformed tetrachromatic networks for short and middle wavelength output units (Fig. 2). By restricting the available input space, fewer connections needed to be learned, and could potentially be optimized with fewer training examples. We investigated this by generating a new set of *de novo* tetrachromatic networks with an L2 regularization term explicitly added to the training protocol (Fig. S5). As expected, these networks had smaller connection strengths, but network performance failed to improve beyond *de novo* networks without L2 regularization.

Training decreased connection strengths, but it was unclear whether homologous hidden units consistently converged on similar connection strengths. To investigate connection strength variability, we compared relative connection strength from before and after training. For each starting network, hidden units were ranked from 1 to 30 in order of increasing connection strength. Similarly, we assigned ranks to hidden units for each trained network and asked if starting rank was correlated with trained rank (Fig. 7C). For both evolved and de novo networks, starting rank was significantly correlated with trained rank ($p < 0.001$). Driven primarily by hidden units with large connection strengths, the evolved network correlation ($r = 0.72 \pm 0.10$) was significantly stronger than the de novo network correlation ($r = 0.43 \pm 0.12$, Cohen's $d = 1.9$). Together with the cosine distance analyses, this result indicates that evolved and *de novo* networks differ in which features of the connection pattern are flexible and change to support learning.

Hidden unit opponent tuning and connection strength combine to give rise to opponent computations

Combined with the small cosine distances for *de novo* networks, this result raised the possibility that connection strength rather than opponent tuning was the important factor driving opponent channel similarities. In other words, the observed opponent channels might disproportionately reflect the tuning of a small number of hidden units with large connection strengths rather than real variability in overarching network computation. Since *de novo* network connection strengths varied more, the apparent increase in opponent channel similarities for evolved networks might not be due to constraints on the opponent tuning. To control for this possibility, we removed the influence of connection strength by normalizing the connection strength of each hidden unit to a unit vector. Using these normalized hidden units, we then recalculated the opponent channels and performed the same analyses as seen in Figure 4.

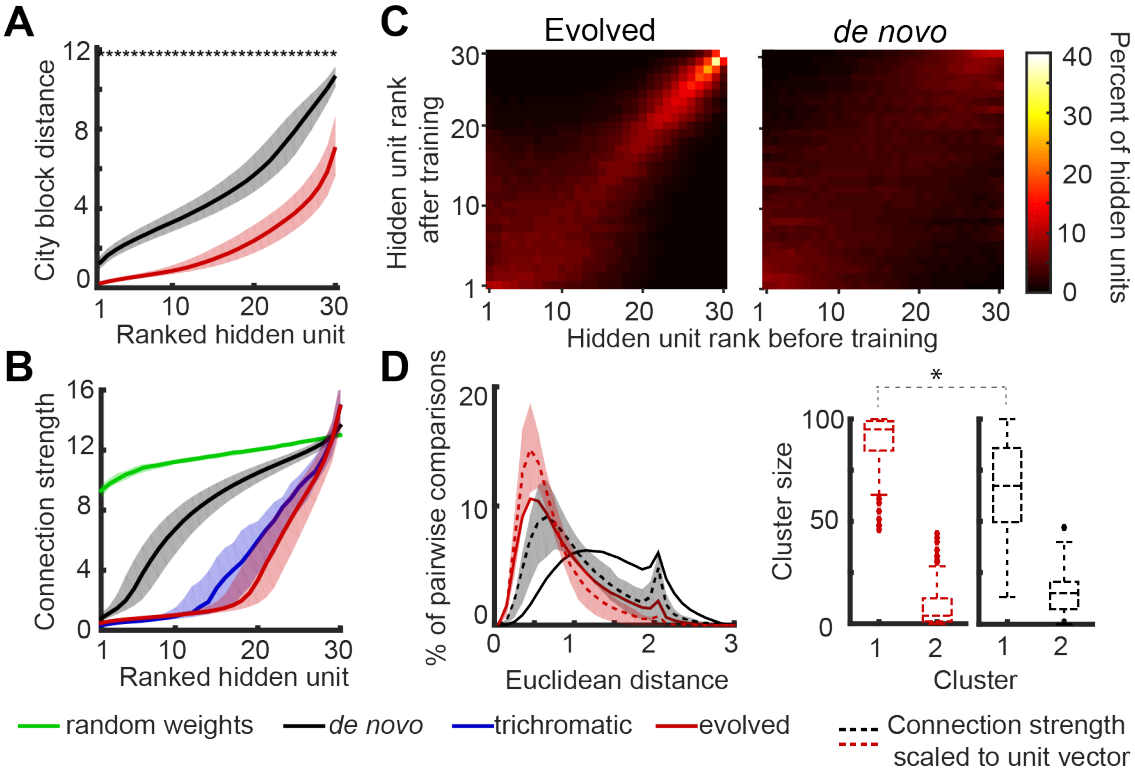

Figure 7: Hidden unit connection strength affects opponent channels and differs between evolved and *de novo* networks
A. City block distances were measured for each hidden unit before and after training. Distances for the 30 hidden units in each network were then ranked in order of increasing distance and compared between evolved and *de novo* networks. Shading shows 25th and 75th percentiles, and asterisks show significant differences between evolved and *de novo* networks with Cohen's d > 0.5.
B. Connection strength was measured as the sum of the absolute value of the input weights to a hidden unit. Hidden units were ranked in order of increasing connection strength for each network. Shading shows the 25th and 75th percentile.
C. The connection strength rank was calculated for each hidden unit before and after training. Since each starting network was independently trained 100 times, the x-axis comprises 100 networks while the y-axis has 10,000 networks. Percentages in the heat map correspond to each column independently. The correlation between pre-training and post-training rank was significantly bigger for evolved (r = 0.72 ± 0.10) networks compared to *de novo* networks (r = 0.43 ± 0.12, Cohen's d = 1.2).
D. The potential influence of connection strength on opponent channels was removed by normalizing the connection strength of each hidden unit to a unit vector. This normalization increases opponent channel similarities for both evolved and *de novo* networks, with evolved remaining significantly more similar. Cluster sizes were calculated with a distance threshold of 1.0 to match Fig. 4. Decreasing the distance threshold increases the differences between evolved and *de novo* networks.

Normalizing hidden unit connection strength led to increases in opponent channel similarities for both evolved and *de novo* networks (Fig. 7D). For *de novo* networks, the largest cluster nearly doubled in size, and the distribution of pairwise distances shifted towards 0 to approximately match evolved networks that did not have normalized connection strengths. Normalizing connection strength led to more modest increases in opponent channel similarities for evolved networks, although this may be due to ceiling effects since evolved networks were already highly similar. Regardless, evolved network opponent channels remained significantly more similar than *de novo* networks (Fig. 7D, pairwise distance distribution: JSD = 0.10; cluster size: Cohen's d = 1.2).

Together, these results showed that the opponent channels observed in Figure 4 were created through a combination of connection strength and opponent tuning. Even though *de novo* network hidden units have small cosine distances reflecting small changes in opponent tuning, these small changes across a population of hidden units combined to have a large influence on

the opponent channels as well as the output tuning curves. Evolved network computations, in contrast, were constrained by the computations of the ancestral trichromatic network. A subset of hidden units, however, was relatively unconstrained and free to vary. This flexibility may be important for allowing the network to traverse the computational space and lead to the observed differences in learning rate.

Discussion

Phylogenetic history has the capacity to constrain and bias future evolutionary trajectories, and these effects may be especially strong for complex neural circuits and the computations they implement (Katz, 2011; Smith et al., 1985). To examine the role of ancestry on neural computation, we simulated the convergent evolution of tetrachromatic vision both within and across distinct phylogenetic lineages using a simple neural network and training protocol. In our networks, trichromatic ancestry affected the rate of learning (Fig. 3) and also biased color vision computations into a restricted range of the available computational state space (Fig. 4). Overall, our results broadly show that trichromatic ancestry constrains the evolution of tetrachromatic vision, which can potentially explain why there is no clear ecological explanation describing which butterflies do and do not have tetrachromatic vision (Briscoe and Chittka, 2001). These findings suggest that evolutionary history is an important factor that influences the functional organization of neural circuits.

Expanding color vision from trichromatic to tetrachromatic necessarily begins with the evolution of a new photoreceptor. Consistent with the sensory periphery being evolutionarily labile and relatively immune to deleterious pleiotropic effects (Bendesky and Bargmann, 2011; Tosches, 2017), mutating the input layer led to minimal performance impairments in our model (Fig. 2,S2). Output unit tuning shifted, but the output layer maintained a structured organization of narrowly tuned units likely capable of trichromatic vision. Additionally, improved luminance discrimination can potentially offset any color vision impairments (Mancuso et al., 2009). Thus, a peripheral adaptation that creates a mismatched tetrachromatic eye and trichromatic brain is unlikely to play an important role in facilitating or impeding evolution.

Peripheral adaptation did not impair color vision but it also did not improve it, effectively making the mutation a neutral variant with respect to fitness. Neutral variants present at low frequency in a population are more likely to go extinct than drift to fixation, so the rate of circuit modification is likely an important factor determining whether a population retains a red photoreceptor. Every network successfully evolved tetrachromacy, but the rate of adaptation varied significantly between different trichromatic networks (Fig. 3). Networks that learn slowly may be rigidly constrained with bottlenecks that prevent modification. In contrast, fast-learning networks might have flexible computational structures with many accessible adaptive mutations or evolutionary pathways to choose from.

The degree of flexibility in hidden unit input tuning likely contributes to learning rates. The non-linear transformation of hidden unit input responses typically permitted input weights to vary substantially without affecting output responses. Input weight sign changes were the major exception and were important for the development of green vs. red opponency (Fig. 6). This computational dynamic could facilitate targeted changes to network computations that have

limited effects on existing computations necessary for trichromatic vision. Although we were unable to explicitly test this idea directly, hidden units in different trichromatic networks may vary in how flexible they are. If a network is less capable of decoupling the targeted development of green vs. red opponency from existing trichromatic computations, tetrachromacy evolution could become difficult and slow.

Despite the computational flexibility of single hidden units, starting weights strongly constrained the overarching opponent channel computations of a network. Consistent with color vision theory (Chittka, 1992; Hurvich and Jameson, 1957; Vorobyev and Osorio, 1998), *de novo* networks used diverse opponent channels that spanned an effectively infinite computational state space. Evolved networks, in contrast, were generally restricted to the original trichromatic channels and the addition of a specific green vs. red channel. Thus, comparisons across phylogenetic lineages gave the appearance of unconstrained computational flexibility, but evolutionary trajectories were predictable and strongly constrained by ancestral history. This is similar to hemoglobin evolution in high altitude birds, where protein sequence variability across independent evolutionary events suggests many paths to adaptation (Natarajan et al., 2016). However, accounting for phylogenetic history and existing neutral sequence variability revealed a constrained pattern of evolution.

Networks learned using a series of discrete training epochs comprised of numerous monochromatic stimuli. How individual training epochs relate to developmental or generational time in real species could vary substantially between different taxa. In mammals, activity dependent development plays a significant role in the functional organization of the cortex, so the entirety of training time could occur within single individuals, which is supported by experimental data in two species. Squirrel monkeys have two opsin genes, with the X-linked opsin having a green and red allele. Males and homozygous females are obligate dichromats, but because of X-inactivation, heterozygous females have both green and red photoreceptors. Behavioral tests have indirectly shown evidence of trichromatic color vision in these females (Jacobs, 2009). Rodents do not have natural allelic variation for the X-linked opsin, but experimental introduction of a red allele led to similar behavioral results (Jacobs et al., 2007), although the computational basis of this behavior and whether these animals exhibited true trichromatic vision remains uncertain (Makous, 2007). Together, these results suggest mechanisms of cortical development have sufficient flexibility to code for trichromatic vision when the periphery allows for it. Nonetheless, genetics do have a significant role in behavior and its interaction with activity dependent development remains unclear.

In contrast to the mammalian cortex, development of the insect nervous system appears to be more strongly genetically specified (Brochtrup and Hummel, 2011; Hiesinger et al., 2006). For these animals, training epochs could correspond to individual adaptive mutations occurring over the course of generations. The effect of single epochs on performance varied substantially from large to small, which mirrors the exponential distribution of effect size that genetic loci have on behavioral traits. While genetically determined wiring is dominant in insects, a degree of activity dependent development has been observed in *Drosophila* (Golovin et al., 2019; Kaneko et al., 2017; Vonhoff and Keshishian, 2017). The extent and function remains unclear, but the effect does appear to primarily affect connection strength rather than circuit structure. This limited

plasticity could relate to the first training epochs, where trichromatic performance is restored in the absence of substantial changes to circuit computation.

In theory, color vision has an infinite number of perceptually equivalent computational solutions (Chittka, 1992; Vorobyev and Osorio, 1998), and thus, the computational state space is relatively symmetrical and lacks a globally optimal implementation. One benefit of this simplicity is that it isolates the effects of evolutionary constraints and allows us to more confidently conclude that similarities between networks sharing the same starting weights and differences between networks with different starting weights were caused by ancestral network configurations. However, the relative simplicity and feed-forward organization of color vision computations may not be representative of most neural circuits.

Circuits and computations for other animal behaviors and perceptions can be substantially more complex, with computational landscapes that may have numerous local minima that vary in performance and can be separated by barriers that vary in how difficult they are to traverse. These rugged landscapes could introduce a complex and dynamic interaction between optimality and constraint. Having shown here that phylogenetic history does influence network evolution and computation, identifying a more complex computation suitable for evolutionary modeling would be valuable for investigating this relationship.

Acknowledgements: We would like to thank Joe Lombardo, Ben Hoshal, and Siwei Wang for useful comments and discussion on previous versions of this manuscript. This work was supported by the University of Chicago Big Ideas Generator, the Alfred P. Sloan Foundation (SEP), and an NSF CAREER award 1652617 (SEP).

Author Contributions: NPB conceived of the project, generated the models, and completed network analysis and comparisons. SEP supervised project development and analysis. NPB and SEP wrote the paper together.

References

Bendesky A, Bargmann CI. 2011. Genetic contributions to behavioural diversity at the gene–environment interface. *Nat Rev Genet* **12**:809–820. doi:10.1038/nrg3065

Bengio Y, Louradour J, Collobert R, Weston J. 2009. Curriculum LearningProceedings of the 26th Annual International Conference on Machine Learning, ICML '09. New York, NY, USA: ACM. pp. 41–48. doi:10.1145/1553374.1553380

Brakefield PM. 2011. Evo-devo and accounting for Darwin's endless forms. *Philosophical Transactions of the Royal Society B: Biological Sciences* **366**:2069–2075. doi:10.1098/rstb.2011.0007

Briscoe AD. 2008. Reconstructing the ancestral butterfly eye: focus on the opsins. *Journal of Experimental Biology* **211**:1805–1813. doi:10.1242/jeb.013045

Briscoe AD. 2002. Homology Modeling Suggests a Functional Role for Parallel Amino Acid Substitutions Between Bee and Butterfly Red- and Green-Sensitive Opsins. *Mol Biol Evol* **19**:983–986. doi:10.1093/oxfordjournals.molbev.a004158

Briscoe AD, Chittka L. 2001. The evolution of color vision in insects. *Annual review of entomology* **46**:471–510.

Brochtrup A, Hummel T. 2011. Olfactory map formation in the Drosophila brain: genetic specificity and neuronal variability. *Current Opinion in Neurobiology*, Developmental neuroscience **21**:85–92. doi:10.1016/j.conb.2010.11.001

Chittka L. 1992. The colour hexagon: a chromaticity diagram based on photoreceptor excitations as a generalized representation of colour opponency. *J Comp Physiol A* **170**:533–543. doi:10.1007/BF00199331

Ding Y, Lillvis JL, Cande J, Berman GJ, Arthur BJ, Long X, Xu M, Dickson BJ, Stern DL. 2019. Neural Evolution of Context-Dependent Fly Song. *Current Biology* **29**:1089-1099.e7. doi:10.1016/j.cub.2019.02.019

Fawcett TW, Hamblin S, Giraldeau L-A. 2013. Exposing the behavioral gambit: the evolution of learning and decision rules. *Behav Ecol* **24**:2–11. doi:10.1093/beheco/ars085

Frentiu FD, Bernard GD, Sison-Mangus MP, Brower AVZ, Briscoe AD. 2007. Gene Duplication Is an Evolutionary Mechanism for Expanding Spectral Diversity in the Long-Wavelength Photopigments of Butterflies. *Mol Biol Evol* **24**:2016–2028. doi:10.1093/molbev/msm132

Golovin RM, Vest J, Vita DJ, Broadie K. 2019. Activity-Dependent Remodeling of Drosophila Olfactory Sensory Neuron Brain Innervation during an Early-Life Critical Period. *J Neurosci* **39**:2995–3012. doi:10.1523/JNEUROSCI.2223-18.2019

Hale ME. 2014. Mapping Circuits beyond the Models: Integrating Connectomics and Comparative Neuroscience. *Neuron* **83**:1256–1258. doi:10.1016/j.neuron.2014.08.032

Heiligenberg W, Rose G. 1985. Neural correlates of the jamming avoidance response (JAR) in the weakly electric fish Eigenmannia. *Trends in Neurosciences* **8**:442–449. doi:10.1016/0166-2236(85)90157-2

Hiesinger PR, Zhai RG, Zhou Y, Koh T-W, Mehta SQ, Schulze KL, Cao Y, Verstreken P, Clandinin TR, Fischbach K-F, Meinertzhagen IA, Bellen HJ. 2006. Activity-Independent Prespecification of Synaptic Partners in the Visual Map of Drosophila. *Current Biology* **16**:1835–1843. doi:10.1016/j.cub.2006.07.047

Hurvich LM, Jameson D. 1957. An opponent-process theory of color vision. *Psychological Review* **64**.

Jacobs GH. 2009. Evolution of colour vision in mammals. *Phil Trans R Soc B* **364**:2957–2967. doi:10.1098/rstb.2009.0039

Jacobs GH, Williams GA, Cahill H, Nathans J. 2007. Emergence of Novel Color Vision in Mice Engineered to Express a Human Cone Photopigment. *Science* **315**:1723–1725. doi:10.1126/science.1138838

Kaneko T, Macara AM, Li R, Hu Y, Iwasaki K, Dunnings Z, Firestone E, Horvatic S, Guntur A, Shafer OT, Yang C-H, Zhou J, Ye B. 2017. Serotonergic Modulation Enables Pathway-Specific Plasticity in a Developing Sensory Circuit in Drosophila. *Neuron* **95**:623-638.e4. doi:10.1016/j.neuron.2017.06.034

Katz PS. 2011. Neural mechanisms underlying the evolvability of behaviour. *Phil Trans R Soc B* **366**:2086–2099. doi:10.1098/rstb.2010.0336

Katz PS. 2007. Evolution and development of neural circuits in invertebrates. *Current Opinion in Neurobiology*, Development **17**:59–64. doi:10.1016/j.conb.2006.12.003

Kien J, Menzel R. 1977. Chromatic properties of interneurons in the optic lobes of the bee. *J Comp Physiol* **113**:17–34. doi:10.1007/BF00610451

Koshitaka H, Kinoshita M, Vorobyev M, Arikawa K. 2008. Tetrachromacy in a butterfly that has eight varieties of spectral receptors. *Proc R Soc B* **275**:947–954. doi:10.1098/rspb.2007.1614

Lee BB, Martin PR, Valberg A. 1989. Sensitivity of macaque retinal ganglion cells to chromatic and luminance flicker. *The Journal of Physiology* **414**:223–243. doi:10.1113/jphysiol.1989.sp017685

Makous W. 2007. Comment on "Emergence of Novel Color Vision in Mice Engineered to Express a Human Cone Photopigment." *Science* **318**:196–196. doi:10.1126/science.1146084

Mancuso K, Hauswirth WW, Li Q, Connor TB, Kuchenbecker JA, Mauck MC, Neitz J, Neitz M. 2009. Gene therapy for red–green colour blindness in adult primates. *Nature* **461**:784–787. doi:10.1038/nature08401

Matsubara J, Heiligenberg W. 1978. How well do electric fish electrolocate under jamming? *J Comp Physiol* **125**:285–290. doi:10.1007/BF00656862

Natarajan C, Hoffmann FG, Weber RE, Fago A, Witt CC, Storz JF. 2016. Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science* **354**:336–339. doi:10.1126/science.aaf9070

Neumeyer C. 1986. Wavelength discrimination in the goldfish. *J Comp Physiol* **158**:203–213. doi:10.1007/BF01338563

Parker GA, Smith JM. 1990. Optimality theory in evolutionary biology. *Nature* **348**:27–33. doi:10.1038/348027a0

Paulk AC, Dacks AM, Phillips-Portillo J, Fellous J-M, Gronenberg W. 2009. Visual Processing in the Central Bee Brain. *J Neurosci* **29**:9987–9999. doi:10.1523/JNEUROSCI.1325-09.2009

Rose GJ. 2004. Insights into neural mechanisms and evolution of behaviour from electric fish. *Nat Rev Neurosci* **5**:943–951. doi:10.1038/nrn1558

Rose GJ, Canfield JG. 1991. Discrimination of the sign of frequency differences bySternopygus, an electric fish without a jamming avoidance response. *J Comp Physiol A* **168**:461–467. doi:10.1007/BF00199605

Schein SJ, Desimone R. 1990. Spectral properties of V4 neurons in the macaque. *J Neurosci* **10**:3369–3389. doi:10.1523/JNEUROSCI.10-10-03369.1990

Seeholzer LF, Seppo M, Stern DL, Ruta V. 2018. Evolution of a central neural circuit underlies Drosophila mate preferences. *Nature* **559**:564–569. doi:10.1038/s41586-018-0322-9

Smith JM, Burian R, Kauffman S, Alberch P, Campbell J, Goodwin B, Lande R, Raup D, Wolpert L. 1985. Developmental Constraints and Evolution: A Perspective from the Mountain Lake Conference on Development and Evolution. *The Quarterly Review of Biology* **60**:265–287. doi:10.1086/414425

Stavenga DG, Arikawa K. 2006. Evolution of color and vision of butterflies. *Arthropod Structure & Development*, Origin and Evolution of Arthropod Visual Systems (Part I) **35**:307–318. doi:10.1016/j.asd.2006.08.011

Stern DL. 2013. The genetic causes of convergent evolution. *Nature Reviews Genetics* **14**:751–764. doi:10.1038/nrg3483

Storz JF. 2016. Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics* **17**:239–250. doi:10.1038/nrg.2016.11

Swihart SL. 1972. The neural basis of colour vision in the butterfly, Heliconius erato. *Journal of Insect Physiology* **18**:1015–1025. doi:10.1016/0022-1910(72)90038-8

Telles FJ, Kelber A, Rodríguez-Gironés MA. 2016. Wavelength discrimination in the hummingbird hawkmoth Macroglossum stellatarum. *Journal of Experimental Biology* **219**:553–560. doi:10.1242/jeb.130484

Tierney AJ. 1995. Evolutionary implications of neural circuit structure and function. *Behavioural Processes*, Cognition and Evolution **35**:173–182. doi:10.1016/0376-6357(95)00041-0
Tosches MA. 2017. Developmental and genetic mechanisms of neural circuit evolution. *Developmental Biology*, Development and evolution of sensory cells and organs **431**:16–25. doi:10.1016/j.ydbio.2017.06.016
Valois RLD, Jacobs GH. 1968. Primate Color Vision. *Science* **162**:533–540. doi:10.1126/science.162.3853.533
von Helversen O. 1972. Zur spektralen Unterschiedsempfindlichkeit der Honigbiene. *J Comp Physiol* **80**:439–472. doi:10.1007/BF00696438
Vonhoff F, Keshishian H. 2017. Activity-Dependent Synaptic Refinement: New Insights from Drosophila. *Front Syst Neurosci* **11**. doi:10.3389/fnsys.2017.00023
Vorobyev M, Osorio D. 1998. Receptor noise as a determinant of colour thresholds. *Proceedings of the Royal Society of London B: Biological Sciences* **265**:351–358. doi:10.1098/rspb.1998.0302
Wagner GP, Altenberg L. 1996. Perspective: Complex Adaptations and the Evolution of Evolvability. *Evolution* **50**:967–976. doi:10.1111/j.1558-5646.1996.tb02339.x
Zaccardi G, Kelber A, Sison-Mangus MP, Briscoe AD. 2006. Color discrimination in the red range with only one long-wavelength sensitive opsin. *Journal of Experimental Biology* **209**:1944–1955. doi:10.1242/jeb.02207
Zeki S. 1980. The representation of colours in the cerebral cortex. *Nature* **284**:412–418. doi:10.1038/284412a0

Methods

Network design

Color vision networks were designed as 3 layer, feed forward networks that responded to monochromatic light stimuli. Preliminary networks were generated using a broad array of tuning parameters for both the input and output layers. The particular set of tuning parameters used in this study was chosen because it maximized performance differences between tri- and tetrachromatic networks. The input layer, simulating photoreceptors, had Gaussian tuning curves with 50 nm standard deviations centered on 350 (UV), 430 (blue), 520 (green), and 600 nm (red). Responses were scaled by a luminance factor between 1 and 100 and passed through a saturating nonlinearity that capped the response at 50. A total of 34 output units evenly tiled the visual range, from 320 to 650 nm in 10 nm increments. Each output unit had a standard deviation of 7.5 nm. The hidden layer had between 10 and 50 hidden units, with analyses focused on 30 hidden units. Hidden unit responses were passed through a sigmoid nonlinearity.

Networks were trained with the Levenberg-Marquardt backpropagation algorithm, which is the Matlab default. Alternative learning algorithms failed to perform well. Training samples were monochromatic light stimuli paired with a luminance factor. Wavelengths varied from 300 to 670 nm in 5 nm increments, and luminance factors varied continuously from 1 to 100. Each training epoch had 400 novel pairs of wavelength and luminance factor. Using this large training set minimized the influence of training data on learning trajectories and effectively isolated the

effect of starting weights. Severely reduced training sets with 10 nm wavelength increments and 6 discrete luminance factors (228 total stimuli) resulted in only minor performance decrements.

Circuit evolution was simulated using a two stage training process. First, we trained trichromatic networks with a UV, blue, and two green input unit units. Starting weights were set with the Nguyen-Widrow initialization algorithm. The two green input units had identical starting weights, making them function as a single input. After 50 training epochs, one of the green input units was mutated to red, and the network was trained for 100 more epochs. The un-mutated, trichromatic network was also trained for 50 more epochs (100 total), and these networks were used for all analyses. We trained 100 trichromatic networks using different random starting weights. Each trichromatic network was independently evolved 100 times using different sets of training data, for a total of 10,000 networks. In preliminary analyses we compared networks with different starting weights and the same training data. Results did not differ from networks with different starting weights and different training data.

As a comparison for evolved networks, we also trained a set of *de novo* networks. These networks were not pre-trained with a trichromatic input layer, but were otherwise identical to the evolved networks. Instead, these networks were directly trained for tetrachromatic vision from random starting weights. Mirroring the evolved networks, we generated 100 sets of random starting weights, again using the Nguyen-Widrow initialization algorithm. Each random network was trained 100 independent times using a tetrachromatic input layer. All analyses on these networks were performed in the same way as the evolved networks.

Network performance

We measured tuning curves for each of the 34 output units and compared them to the target response. Output unit responses were calculated for stimuli between 300 and 670 nm in 1 nm increments. Tuning curves were made by averaging the response across 10 luminance factors ranging from 10 to 100 in steps of 10. Performance was then defined as the MSE between the observed and expected tuning curve. Results were qualitatively similar when calculating the MSE for any single luminance factor. Overall performance decreased slightly without averaging, but the shape of the output units tuning curves was unchanged.

Opponent channels

The overarching computational structure of the hidden layer was analyzed by using PCA to extract the opponent channels. For this analysis, each hidden unit was an observation, and the UV, blue, green, and red input weights were the variables. PCA reduced the dimensionality to 4 eigenvectors, the first three of which consistently had color opponent interactions. The fourth typically had all positive input weights indicating a luminance channel and was not analyzed further. All opponent channels were scaled to unit vectors to make them comparable, but the sign was arbitrary. To account for this, we calculated every pairwise correlation between an opponent channel for one network and all other networks in a group of 100. Opponent channels with at least half of the correlations less than zero had its sign flipped. We tried several different sign flipping procedures, and this correlation based method maximized similarities between networks.

Similarities and differences between networks were measured using agglomerative hierarchical clustering. We compared both single channels and the full set of three opponent channels concatenated together. We also generated groups of random opponent channels as a null hypothesis. These opponent channels were made with a random number generator, scaled to a unit vector, and sign flipped using the same correlation procedure. There was no prior expectation on how the opponent channels should be ordered, and efforts to maximize similarities found no procedure better than ordering and labeling opponent channels according to percent of variance explained. We clustered networks using Euclidean distances between networks and a complete linkage algorithm, which maximized cophenetic correlations ($c > 0.85$). The clustering threshold was set at 1.0, but results did not depend on the precise threshold. In addition to cluster size, we also examined the distribution of all pairwise distances within a group of 100 networks. Pairwise distances were binned in distance intervals of 0.1 for visualization purposes and intervals of 0.01 for analysis since this bin size maximized distribution entropy. The distributions of pairwise distances were compared between group types using the Jensen-Shannon divergence, a metric related to the Kullback-Leibler divergence.

Hidden unit analysis

Homologous hidden units were compared between networks that shared the same set of starting weights. For cosine and city block distances, the input weights from each of the input layer photoreceptors were used to create a 4-dimensional representation of the hidden unit tuning. To generate tuning curves, hidden units responses were measured for monochromatic stimuli ranging from 300 to 670 nm in 1 nm increments and averaged over 10 luminance factors. For the input tuning curve, the input layer response to a stimulus was convolved with the input weights to a hidden unit. To generate the output tuning curve, these input tuning curves were passed through a sigmoid nonlinearity. Output tuning curves were compared and clustered between homologous hidden units using Euclidean distances and a complete linkage function.

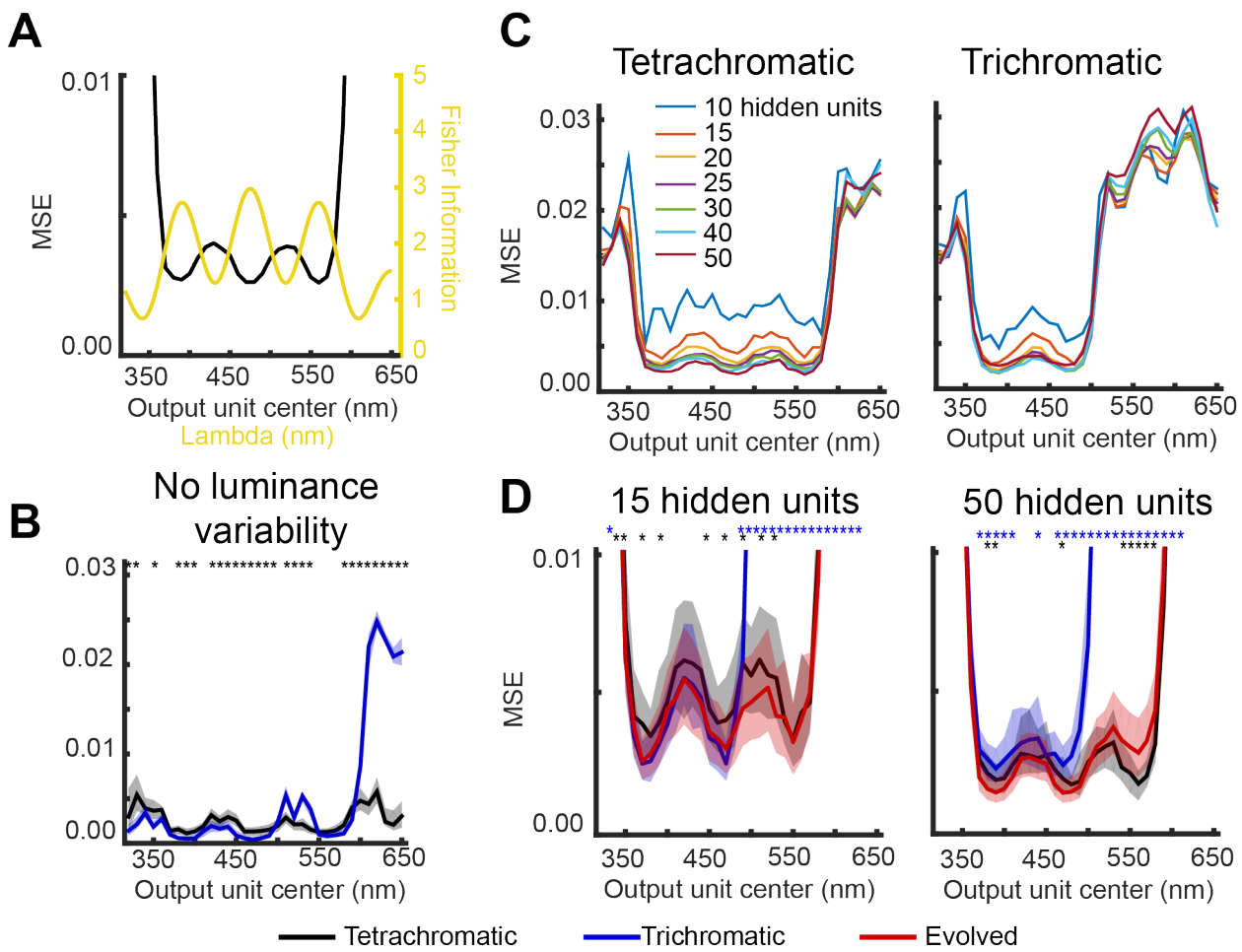

Figure S1

A. Black shows the performance of de novo tetrachromatic networks, showing a zoomed in version of the same plot in Fig. 1D. Yellow shows the Fisher Information of the input layer for networks with UV, blue, green, and red photoreceptors.

B. Tri- and tetrachromatic networks were trained with a training set of monochromatic stimuli that lacked luminance variability. Although differences between tri- and tetrachromatic networks were observed (asterisks indicate significance with Cohen's d > 0.5), networks perform similarly and lack clear differences for long most long wavelength output units.

C. Performance of de novo networks with variable hidden layer size. Performance begins to asymptote around 15 hidden units, and regardless of hidden layer size, long wavelength output units perform poorly for trichromatic networks.

D. Networks with 15 or 50 hidden units were created using the two-stage training procedure and were compared to de novo networks. Data shows the performance after 100 training epochs. Shading shows the 25th and 75th percentile of performance. Asterisks mark output units that were significantly different between evolved and either tri- (blue) or tetrachromatic (black) networks with Cohen's d > 0.5.

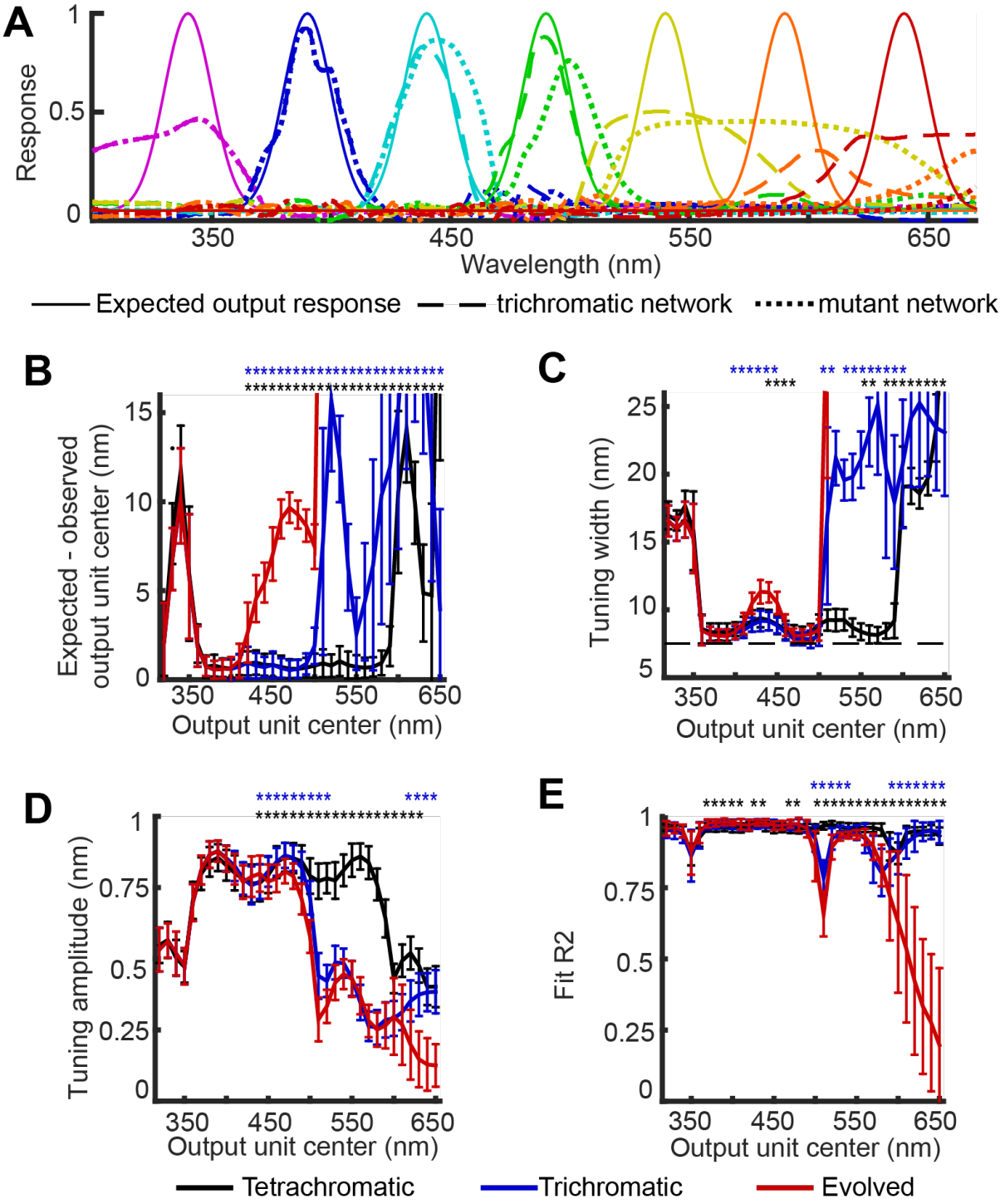

Figure S2

A. Example output tuning curves for a trichromatic network and the same network after mutating the input layer to tetrachromatic.

B-E. Instead of calculating the MSE, we instead fit Gaussians to each output unit tuning curve. Each panel shows a comparison of the fit parameters, with blue and black asterisks indicating significant differences between evolved and de novo tri- or tetrachromatic networks, respectively, with Cohen's d > 0.5. In panel B, the trichromatic dip at 550 nm is observed because all long wavelength output units exhibited broad tuning centered on the inflection point of the green photoreceptor. In panel C, the dotted line denotes the expected tuning width.

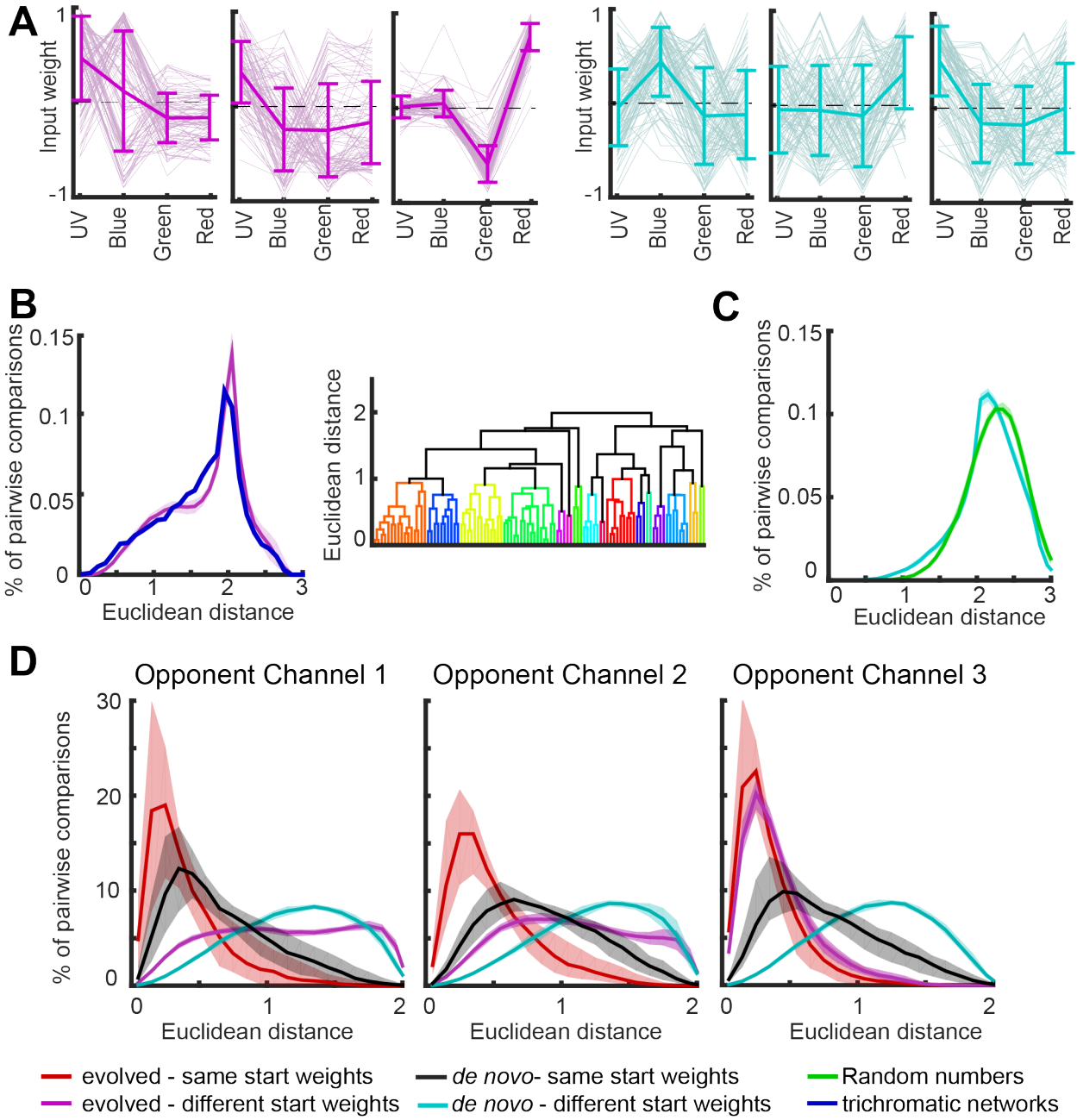

Figure S3

A. Example opponent channels for a group of 100 evolved (left) and de novo (right) networks that each had a different set of starting weights. Note the green vs. red opponency for channel 3 for all evolved networks regardless of the ancestral network.

B. (left) The distribution of pairwise distances was computed for the trichromatic networks that were subsequently evolved and compared to groups of evolved networks with different trichromatic starting weights. (right) Dendrogram shows the computational similarities between the 100 trichromatic networks that were subsequently evolved.

C. A null distribution of opponent channel similarities was created generating opponent channels using random numbers and scaling them to a unit vector to match the PCA analysis. The pairwise distance distribution is shown in comparison to the distribution for groups of de novo networks with different random starting weights.

D. Opponent channel similarities were examined for each individual opponent channel rather than as the concatenated group of 3. For opponent channel 3, note the shift towards 0 for groups of evolved networks with different starting weights to overlap substantially with groups of evolved networks with the same starting weights.

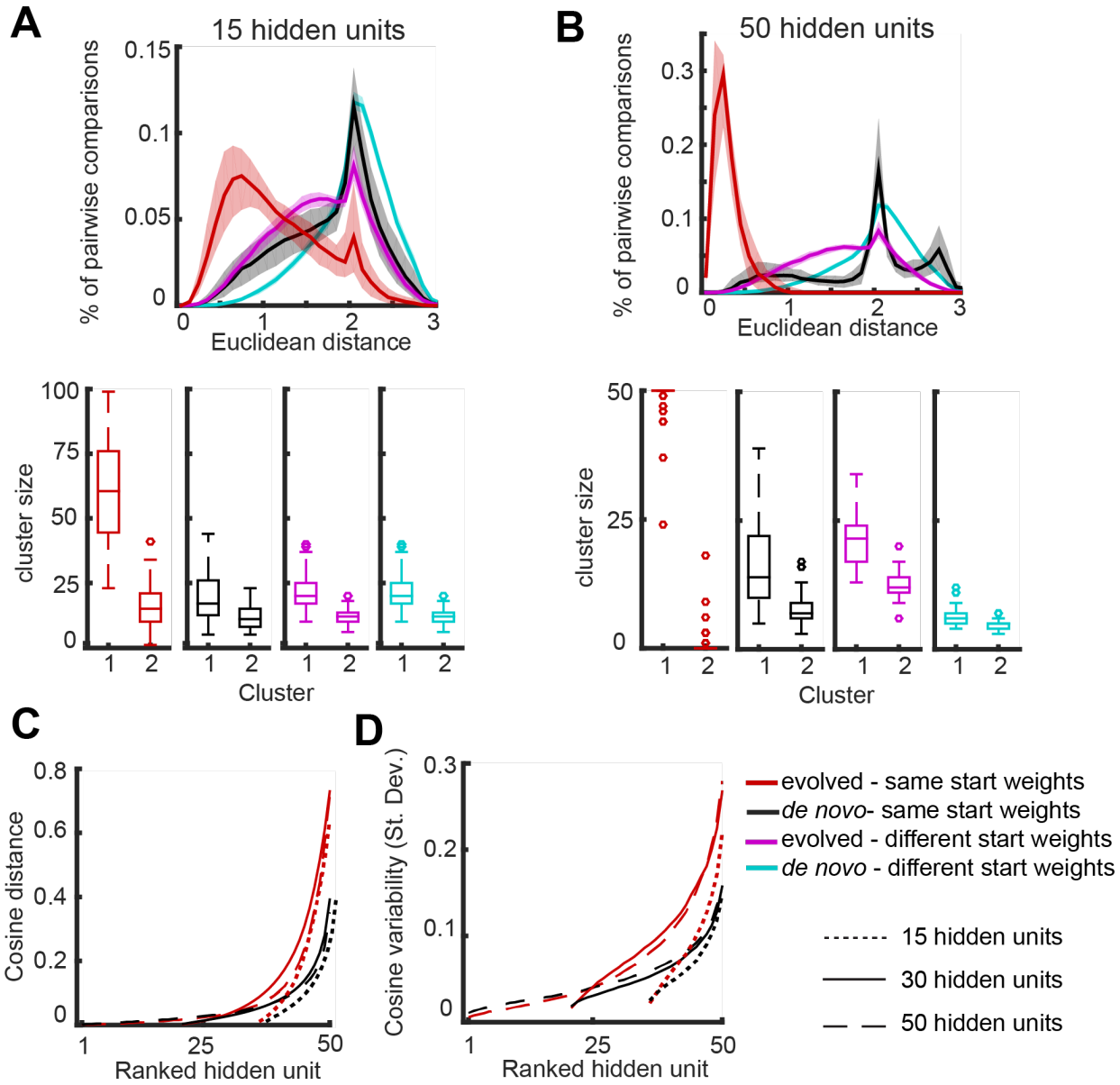

Figure S4

A. Opponent channels were analyzed for networks with only 15 hidden units.

B. Opponent channels were analyzed for networks with 50 hidden units. Note that due to the time required to train these networks, only 50 trichromatic networks were generated and each was evolved 50 independent times.

C. Cosine distances for ranked hidden units are shown for networks 15, 30, and 50 hidden units. Data are aligned so that rank 50 is the largest distance for each network, such that networks with 15 and 30 hidden units only extend to ranks 36 and 21, respectively.

D. Same as C, except showing the variability in cosine distances.

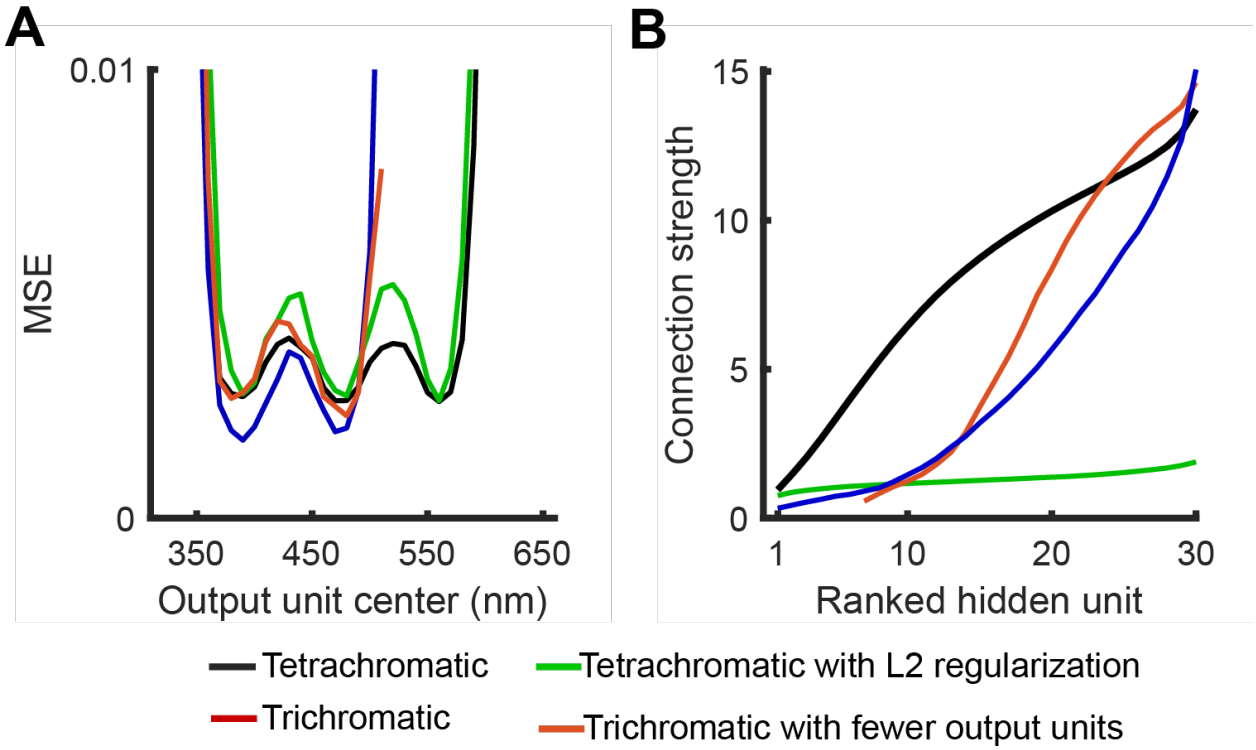

Figure S5

A. Trichromatic networks with only 24 rather than 34 output units were made as well as de novo tetrachromatic networks that had an L2 regularization term added to the training protocol. Neither of these changes were able to reproduce the improved performance of trichromatic networks compared to tetrachromatic networks.

B. Connection strength was measured for each of the four network types.