A Critical Analysis of Color Neurons in Vision Models

Fernando Aguilar-Canto [*1] **Hiram Calvo** [1] **Ricardo Menchaca Méndez** [2]

Abstract

Unit-level interpretability in deep neural networks faces growing skepticism, and the question of whether any single artificial neuron can be rigorously understood remains open. We propose three explicit criteria for validating unit-level interpretations—completeness, falsifiability, and relevance—and apply them to color-selective neurons, a class of units long reported in the literature and studied with increasing sophistication, but not previously validated under explicit, simultaneous multi-criteria standards. Using a protocol that combines causal ablation, controlled synthetic stimuli, natural image categorization, and input-level interventions, we identify and validate five units across three architectures (ResNet50, ViT-B/16, ViT-L/32). These units span excitatory, bimodal, and inhibitory selectivity to distinct spectral regions, and each satisfies all three criteria. Our results demonstrate that unit-level understanding is attainable under rigorous standards, and we argue that interpretability can benefit from reframing interpretation as falsifiable hypothesis testing.

1. Introduction

Contemporary unit-level interpretability of deep neural networks faces skepticism at four different levels: 1) the *interpretation* of the results yielded by certain methods (Gale et al., 2020), 2) the *validity* of the methods themselves, usually tied to top activations (Geirhos et al., 2023; Bareeva et al., 2024; Nanfack et al., 2024; Oikarinen & Weng, 2024), 3) the *relevance* of the most interpretable units (Morcos et al., 2018), and 4) whether the unit level constitutes a *proper* tier of analysis in the first place (Huang et al., 2023).

Traditional neuroscience, from the pioneering work of Hubel and Wiesel (1959) through decades of single-unit studies (De Valois et al., 1966; Wurtz, 1969; Desimone et al., 1984; Hegdé & Van Essen, 2000; 2007; Ito & Komatsu, 2004; Quiroga et al., 2005; Ilg, 2008; Mormann et al., 2017), has relied on individual neurons to derive generalizations about the behavior of entire systems such as the visual pathways. Yet the relationship between biological and artificial networks remains debated (Nadler et al., 2023; Linsley et al., 2025), and the extent of their correspondence might not be clearly understood. In light of this tension we ask a deliberately minimal but consequential question: Is there at least one artificial neuron that we can claim to understand under sufficient standards?

To address this skepticism, we propose that any interpretation I_u of a unit u should satisfy three criteria:

1. *Completeness criterion*: The interpretation I_u must hold over a predefined and semantically coherent domain $\mathcal{D}$, not dependent on the explainer (for instance, activation maximization) nor exclusively on synthetic stimuli (following the empirical argument of Geirhos *et al.* (2023)). For example, if we claim that u responds to a specific body part of animals (as in Aguilar and Calvo (2023)), this claim should be evaluated across a dataset of animals, as well as across varied natural images containing that body part—not only on the top-k activating exemplars.

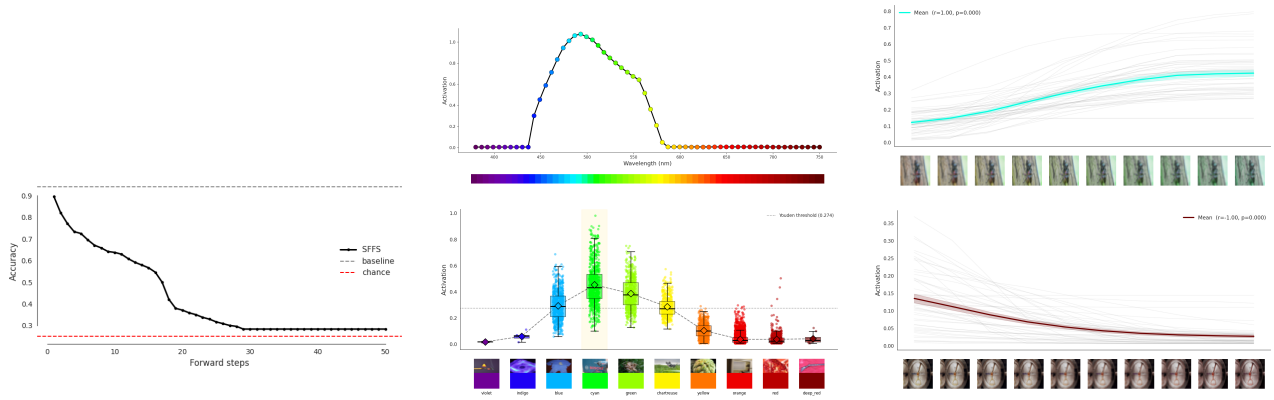
2. *Falsifiability criterion*: The claim made by I_u must be empirically testable. There must exist an experimental procedure that could reveal I_u to be false if it were incorrect, ruling out interpretations that are so flexible or post-hoc that they accommodate any observed response.

3. *Relevance criterion*: The unit u must be functionally relevant to the network or layer, even for simple tasks.

Color is a relevant dimension for deep neural networks (De, 2022; Mertens et al., 2023; Chiu et al., 2025). As our literature revision suggests (Section 3), color-selective units exist

[1]Laboratorio de Ciencias Cognitivas Computacionales, Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico [2]Laboratorio de Redes y Ciencia de Datos, Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico. Correspondence to: Fernando Aguilar-Canto <faguilarc2021@cic.ipn.mx>, Hiram Calvo <hcalvo@cic.ipn.mx>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. *Copyright 2026 by the author(s).*



(a) Relevance: SFFS ablation

(b) Synthetic tuning curve

(c) Chromatic tuning (natural)

(d) Shift toward 492 nm (preferred)

(e) Shift toward 743 nm (unpreferred)

Figure 1. **Validation of unit 38** (`layer1`, ResNet50). (a) SFFS confirms causal relevance. (b–c) Tuning curves from synthetic and natural images identify cyan as the preferred band, satisfying the Completeness criterion. (d–e) Color shift intervention confirms the Falsifiability criterion.

and evidence points toward their functional relevance; however, to our knowledge no prior work has proposed explicit criteria for unit-level validation, nor subjected individual color-selective neurons to simultaneous testing of completeness, falsifiability, and relevance. In this article, we apply the proposed criteria through a three-stage methodology (Section 4) that identifies causally relevant units, validates their selectivity across monochromatic[1] natural inputs, and tests the resulting interpretation via input-level intervention.

Our results (Section 5, see Figure 1) indicate that at least five units in earlier layers of ResNet50, ViT-B/16, and ViT-L/32 can be considered genuinely color-selective, demonstrating that rigorous hypothesis-testing frameworks can yield validated unit-level understanding.

2. Preliminary discussion

The three criteria presented above follow directly from the skepticism outlined in the introduction. Most unit-level interpretability techniques rely, explicitly or implicitly, on activation maximization (Erhan et al., 2009), which searches for hyperstimuli that drive a neuron's response to its extreme. Preliminary work by Gale *et al.* (2020) presented evidence against the idea of 'object detectors' in hidden layers (Zhou et al., 2014), since such units exhibited poor hit rates or high false-alarm rates. Recent work has drawn attention to the limitations of this approach (Geirhos et al., 2023; Bareeva et al., 2024; Nanfack et al., 2024), and modern concept-extraction methods tend to characterize the set of stimuli that maximizes the output, rather than studying the full activation range (Oikarinen & Weng, 2024). The completeness criterion is a direct response to this gap: it demands that an interpretation be evaluated over an independently defined domain, not over the exemplars selected by the explainer itself.

The falsifiability criterion addresses a related but distinct concern. Even when an interpretation is proposed, it may be stated in terms that are too vague to be tested—a problem raised by Huang *et al.* (2023), who argued that explanations should be evaluated through causal interventions. Complementary work by Borowski *et al.* (2021) and Zimmermann *et al.* (2021; 2024) has studied whether different explainers enable humans to predict the effects of interventions on unit activations. The falsifiability criterion formalizes the requirement that an interpretation must expose itself to empirical risk.

Finally, the relevance criterion is adopted directly from Morcos *et al.* (2018), who demonstrated that highly selective units are not necessarily the ones the network relies on most. Without a relevance check, interpretability research risks characterizing units that are easy to describe but functionally dispensable.

3. Related Work

Different general single-unit explainers have reported the existence of color-selective neurons. Feature visualization (Olah et al., 2020) has been used to propose a taxonomy of color neurons in the earlier layers of InceptionV1, classifying them into (single) color, multicolor, and color-contrast cells. Notably, visualizations have also been used to suggest the existence of multimodal color units in CLIP—neurons

[1] An image with a predominant color.

that fire in response to both text and image inputs related to a specific color, although they remain sensitive to the Stroop effect (Goh et al., 2021).

Network Dissection (Bau et al., 2017), in turn, relies on the Broden concept dataset, which includes specific categories for colors, allowing the explainer to assign color labels to individual units. Using this approach, the authors identified units associated with specific colors. Related techniques such as CLIP-Dissect (Oikarinen & Weng, 2022) have also been applied to label color neurons in vision transformers (Dorszewski et al., 2025). Bau *et al.* (2020) later extended this framework to study color as a category and proposed ablation as a method to measure the impact of individual units on classification for high-level features; however, color neurons were not explicitly examined in that causal analysis.

In-depth studies on color neurons have also been performed. Rafegas and Vanrell (2017; 2018) developed a framework to identify and classify color neurons into single (unimodal) and double, which are further classified into opponent and non-opponent. In particular, the selectivity index identifies color neurons by computing the ratio of activation of the top-k image patches in color and grayscale. Single and double color neurons are classified by using a Gaussian mixture to determine selectivity to one or two colors. Extending this methodology, Arias *et al.* (2025) categorized CLIP neurons into color, color-word, any-word, and multimodal cells using an additional color-label selectivity index to capture the Stroop effect. Bianco (2025) also applied the selectivity index to study DINOv2 and ViT features in the last layer, and proposed a transformation that enhances classification performance on color-related tasks from these features. Zini *et al.* (2022) used the selectivity index to show that a more realistic data augmentation increases the proportion of color-selective neurons, providing further indirect evidence of the functional relevance of these units.

In Flachot and Gegenfurtner (2018) the authors identified color neurons using color responsivity and analyzed their tuning relative to the first dimension of the RGB_{PCA} color space derived from the training data. Their approach used controlled synthetic stimuli to probe chromatic tuning independently of the explainer. They also proposed a classification of neurons according to the linearity of their tuning curves. In a subsequent study (Flachot & Gegenfurtner, 2021) the authors implemented color manipulations of individual optimal image segments to compute overall color sensitivity, hue selectivity, and chroma responsivity indices, and measured the impact of hue angle rotation on the classification accuracy of several networks including VGG-16. Critically, this study established that classification accuracy was highest when image segments matched the preferred hue of the corresponding units and degraded as the hue diverged, providing evidence that chromatic selectivity in individual units is functionally relevant to network behavior.

While Harris, Mihai, and Hare (2019; 2021) studied principally spectral and spatial opponent cells, this work is relevant since it performed control experiments to validate their claims. Most of the units identified as spectral opponent were also spatially opponent (double opponency). By randomizing the weights, they showed that these characteristics are learned. The authors also retrained using distorted colors to demonstrate that these features are a product of data statistics. Moreover, shuffling the color channels preserved spatial opponency while color opponency dropped substantially.

Although not directly related to the three criteria, other studies have explored joint representations of color and form (Taylor & Xu, 2021) or characterized individual unit responses to color and luminance (Bun & Horwitz, 2023).

3.1. Contributions

Does the previous literature satisfy all three interpretability criteria? General single-unit explainers such as Linear Explanations (Oikarinen & Weng, 2024) attempt to cover the completeness criterion and, according to recent evaluations (Oikarinen et al., 2025), achieve top performance in correlation-based metrics. However, other methods relying on correlation—such as Bills *et al.* (2023)—have been tested with interventions, failing to corroborate the hypothesized selectivity (Huang et al., 2023). We therefore conclude that, at this stage, no general explainer has covered all three criteria simultaneously. Nevertheless, this does not mean that serious attempts have not been performed for specific cases (Bau et al., 2020; Ukita, 2020; Cammarata et al., 2020).

In the case of color neurons, while no single work attempts to cover all three criteria, some of the cited articles provide partial evidence for individual requirements. Hue and color modifications—such as those performed by Flachot and Gegenfurtner (2021) and Harris, Mihai, and Hare (2021)—can be seen as steps toward the falsifiability criterion, although they were not explicitly formulated as hypothesis tests. Regarding the relevance criterion, some works indirectly confirm the importance of color-related cells (Zini et al., 2022; Bianco, 2025); however, color neurons have not been specifically evaluated in their natural representational context—that is, without transformations that may sacrifice performance on other dimensions to favor color processing. As for the completeness criterion, most studies rely on optimal examples / patches, or exclusively synthetic stimuli (Rafegas & Vanrell, 2017; 2018; Flachot & Gegenfurtner, 2018). Under the empirical argument of Geirhos *et al.* (2023), validation on synthetic stimuli alone is insufficient; completeness demands evaluation on natural images as well.

Taken together, the literature suggests that unimodal color neurons exist and are functionally relevant, but to our knowledge this is the first work to propose a set of three criteria–completeness, falsifiability, and relevance–for evaluating unit-level interpretations and to apply them in the context of color-selective cells in artificial networks. While related frameworks have been proposed for evaluating natural language explanations of units in language models (Huang et al., 2023), these have focused on observational and intervention modes without an explicit relevance criterion, and have not been applied to propose valid interpretations. We consider that completely understanding the role of a single unit is a valid and timely objective in light of the current uncertainty surrounding the unit level in mechanistic interpretability.

4. Methodology

The methodological framework consists of three primary steps: 1) identifying a set of units relevant for a color-related task, 2) validating the completeness criterion for each candidate unit, and 3) formulating and testing a specific interpretation for each cell. Although this process can be stated in general terms–similarly to Bau *et al.* (2020)–we believe it is necessary to adapt the procedure to the specific demands of color selectivity: interpretations at this level must be evaluated on their own terms.

4.1. Studied networks and layers

There is some consensus that color-selective neurons are more predominant in early layers (Rafegas & Vanrell, 2017; 2018; Dorszewski et al., 2025). We therefore focus on early layers, including one late layer for comparison. The studied networks and layers are:

1. ResNet50 (He et al., 2016) (`layer1`)
2. ViT-B/16 (Dosovitskiy et al., 2020) (`encoder.block.0.mlp[1]` and `encoder.ln`)
3. ViT-L/32 (Dosovitskiy et al., 2020) (`encoder.block.0.mlp[1]`)

Definition of a unit Throughout this article, the terms *unit*, *cell*, and *neuron* are used interchangeably. For vision transformers, we extract the activation of the MLP layer at the position of the classification token. For ResNet50, we apply global average pooling over spatial dimensions to obtain a single value per convolutional filter. In both cases, the layer's representation is treated as a vector of units.

4.2. Datasets

The methodology requires the construction of three datasets: a *spectral dataset* of synthetic stimuli, a *chromatic dataset* of natural images for the Completeness and Falsifiability criteria, and a smaller *probing dataset* to identify relevant units. All natural images are drawn from the ILSVRC2012 ImageNet validation set, using the train/val/test splits (35,000 / 5,000 / 10,000) introduced by Oikarinen and Weng (2024).

Probing dataset The probing dataset uses the training split for probe training and combines the validation and test splits for evaluation[2]. Each image is assigned to one of five initial color categories (red, yellow, green, blue) based on its dominant hue in HSV space. An image is assigned to a category only if the dominant hue accounts for at least 60% of its pixels and the mean saturation exceeds 0.5. Each class is then truncated to the size of the smallest class (139 images for training, 60 for testing) by random subsampling, yielding a balanced dataset. While this four-class HSV filtering is coarser than the ten-band spectral categorization used for completeness validation, it is sufficient for screening: a layer that cannot support linear separation of four well-separated hues is unlikely to encode fine-grained spectral information.

Spectral dataset Synthetic stimuli are used to locate the peak of chromatic selectivity, which serves as the primary basis for hypothesis formulation. Specifically, we generate solid-color images spanning 60 wavelengths evenly distributed across the visible spectrum (380–750 nm), converting each wavelength to RGB via the Bruton (1996) algorithm. The preferred wavelength of each unit is taken as the one eliciting the maximum mean activation across the synthetic set.

Chromatic dataset To validate the criteria on natural images, we construct a chromatic dataset from the test split of the ImageNet validation set using the same HSV-based categorization as the probing dataset, but now extended to ten spectral bands. Each image is converted to HSV space; images whose mean saturation falls below 0.3 are discarded as lacking a clear chromatic signal. The remaining images are assigned to one of ten spectral bands defined in Appendix A.1 (violet through deep red, each approximately 37 nm wide) based on the dominant hue range. An image is assigned to a band only if the band accounts for at least 50% of its pixels. Unlike the probing dataset, the chromatic dataset is not balanced, as the goal is to capture as many color images as possible for robust statistical comparison. Train / val / test splits are kept in the chromatic dataset.

[2] Using only the validation split yields insufficient samples per class ($\sim$15).

4.3. Relevance criterion

As discussed in Section 3.1, testing unit relevance requires avoiding interventions that limit the network's capabilities in unrelated dimensions. Ideally, relevance should be assessed by measuring the impact on global task performance. However, due to redundancy across layers, ablating individual units may not produce a measurable effect at the system level. For this reason, we evaluate relevance at the layer level: we probe the layer's capacity to solve a color classification task using a linear classifier trained on its activations. If the layer's representation fails to support accurate color classification, then it cannot be relied upon for that task–and, conversely, units whose ablation impairs this capacity are deemed relevant for color processing.

Candidate units are selected through Sequential Floating Forward Selection (SFFS) (Pudil et al., 1994): at each forward step, we individually ablate every remaining unit and select the one whose removal most degrades the probe's accuracy. Every few steps, a backward pruning phase removes units that have become redundant. This procedure directly measures the causal impact of each unit on color classification performance, rather than relying on proxy metrics such as probe weights or selectivity indices (Morcos et al., 2018; Hayne et al., 2022)[3].

4.4. Completeness criterion

Once the relevant units for the color classification task have been identified in a given layer, we validate the Completeness criterion by constructing tuning curves from both the spectral and chromatic datasets. The synthetic tuning curve identifies the preferred spectral band, which serves as the initial hypothesis. We then test whether this preference generalizes to natural images from the chromatic test set. Three conditions must be met for the Completeness criterion to be satisfied on the domain $\mathcal{D}$ of natural images with a discernible predominant color.

First, we test whether the unit's mean activation differs significantly across the ten spectral bands using a Kruskal–Wallis test on the test set. If significant, pairwise Mann–Whitney U-tests with Bonferroni correction identify which bands differ from the preferred band. As a complementary filter, we compute the Class-Conditional Mean Activation Selectivity (CCMAS) (Morcos et al., 2018) to discard units with weak absolute selectivity (≥ 0.1)[4].

Second, we compare the unit's activation for images in the preferred band against its activation for all remaining spectral images on the test set. We report the sensitivity index d' from signal detection theory, which quantifies how well the unit discriminates its preferred color from all other chromatic stimuli.

Third, following Gale *et al.* (2020), we evaluate whether the unit functions as a reliable detector by computing recall and precision on the test set at an activation threshold selected via the Youden index on the validation set, treating images from the preferred band as positives and all other spectral images as negatives. A unit is considered a strong detector if both recall and precision exceed 0.75. However, color-selective units may exhibit more complex behavior: a unit might respond strongly to multiple non-adjacent bands (multimodal selectivity) or show graded responses across perceptually adjacent bands. To capture this, we report pairwise precision and recall between the preferred band and every other band. If a unit achieves high precision against some but not all alternative bands, or responds selectively to a distinct set of bands, the interpretation is refined accordingly rather than discarded.

A unit that passes all three conditions–possibly with a refined interpretation–is considered to satisfy the Completeness criterion.

4.5. Falsifiability criterion

If the Completeness criterion is satisfied, we obtain a candidate interpretation I_u of the form:

> Unit u is selective to the spectral band defined in the HSV hue range $[h_1^\circ, h_2^\circ]$, centered at approximately λ nm.

To test this interpretation against the Falsifiability criterion, we design an input-level intervention on natural images.

We draw a random sample of natural images from the spectral dataset and apply a gradual color shift to each image, transforming its pixels toward the hypothesized preferred wavelength λ. The shift is applied in HSV space: the hue of each pixel is interpolated toward the target hue while saturation and value are held constant, preserving the spatial structure of the image. Shift intensity α is varied from 0

[3] A key difference from Hayne *et al.* (Hayne et al., 2022) is that we use ablation within the probe as a method for identifying relevance at the layer level, rather than using probing weights. This work itself uses ablation as a method to measure the impact. While end-to-end relevance is desirable, the well-documented presence of representational redundancy across units (Morcos et al., 2018; Bau et al., 2020) makes it difficult to detect the importance of individual low-level feature detectors through network-level ablation alone.

[4] As activations are not normalized and can take negative values, CCMAS is not strictly bounded to $[0, 1]$. Values exceeding 1 indicate that the mean activations for the preferred and non-preferred classes have opposite signs, which can occur in layers with unconstrained activation ranges (e.g., LayerNorm in ViTs). For such units, CCMAS should be interpreted with caution; throughout our analysis, we rely on d' and statistical tests as the primary selectivity metrics

(original image) to $\alpha_{\max}$ (maximum shift) in n steps, producing a sequence of progressively color-shifted versions of each image.

If the interpretation is correct, the unit's activation should increase monotonically as the shift toward λ intensifies. We quantify this via Spearman's rank correlation between shift intensity and mean activation across images. A significant positive correlation ($p < 0.05$, $r > 0$) supports the interpretation.

As a control, we repeat the procedure shifting toward wavelengths far from λ (e.g., the complementary wavelength or a distant spectral band). For these control shifts, the unit's activation should not increase and should preferably decrease, confirming that the unit is not merely responding to any color change but specifically to its preferred direction in color space.

An interpretation that passes the main shift test with a significant positive correlation and is not falsified by the control shifts is considered to satisfy the Falsifiability criterion.

5. Results

5.1. ResNet50 - `layer1`

SFFS ablation reduced accuracy from 0.942 to 0.283, identifying 29 relevant units. For conciseness, we analyze in detail the top-ranked unit (38) and unit 39, a bimodal red–blue detector (see Appendix B.1.2). Several other units showed synthetic tuning curves similar to that of unit 38, suggesting that the blue/cyan direction is broadly relevant for the probing task. However, not all SFFS-selected units passed the subsequent criteria: units 177, 24, and 96 failed pairwise Mann–Whitney tests against the remaining color classes, indicating that causal relevance to the probe task does not guarantee chromatic selectivity.

5.1.1. Unit 38: a cyan-selective neuron

Unit 38, identified by SFFS in the `layer1` of ResNet50, showed clear selectivity for the cyan band. The Kruskal–Wallis test confirmed significant variation across bands ($H = 1616.46$, $p < 0.001$). Cyan achieved the highest CCMAS (0.639) and d' (2.485), with significantly higher activation than eight of nine alternative bands. The exception was violet ($p = 0.006$ after Bonferroni correction, threshold ≈ 0.0056), which fell just short of significance, likely due to the very small sample size for violet ($n = 1$ in the test set). Global precision was 0.308 and recall 0.892, indicating graded selectivity across the blue–green region.

The shift experiment (Table 1) confirmed the cyan preference: wavelengths in the blue–green interval $[461, 555]$ nm produced significant positive correlations (all $r = 1.000$, $p < 0.001$), while distant wavelengths (indigo, orange, deep red) produced negative correlations ($r = -1.000$, $p < 0.001$). Tuning and shift curves for this unit are depicted in Figure 1.

Table 1. Shift experiment for unit 38 (ResNet50, `layer1`). Preferred band: cyan ($\sim$ 492 nm).

Target λ (nm)	r	p
430 (indigo)	-1.000	< 0.001
461 (blue)	1.000	< 0.001
486 (blue)	1.000	< 0.001
492 (cyan) (pref.)	1.000	< 0.001
Cyan (HSV)	1.000	< 0.001
499 (cyan)	1.000	< 0.001
524 (green)	1.000	< 0.001
555 (green)	1.000	< 0.001
618 (orange)	-1.000	< 0.001
743 (deep red)	-1.000	< 0.001
Image subsets (shift to cyan, 492 nm)		
Top-20 max cyan	1.000	< 0.001
Top-20 min cyan	1.000	< 0.001

5.2. ViT-B/16 - `encoder.block.0.mlp[1]`

We applied the SFFS procedure to the second MLP layer of the first encoder block (`encoder.block.0.mlp[1]`) of ViT-B/16. The probe achieved an initial accuracy of 79.88% on the four-class color categorization task. SFFS identified four units whose cumulative ablation most degraded this accuracy–2437, 1762, 283, and 2299–reducing probe performance from 79.88% to 24.40%. Figure 6 shows the accuracy decay across forward selection steps.

Of these four units, two (283 and 2299) exhibited clearly interpretable chromatic tuning. The other two (2437 and 1762), despite being identified as causally relevant by SFFS, showed weak color selectivity (CCMAS ≤ 0.012). We focus our subsequent analysis on units 283 and 2299, for which the three criteria could be rigorously evaluated. For presentation purposes, we describe extensively unit 283 in the main text and description of 2299 can be found at Appendix B.2.3.

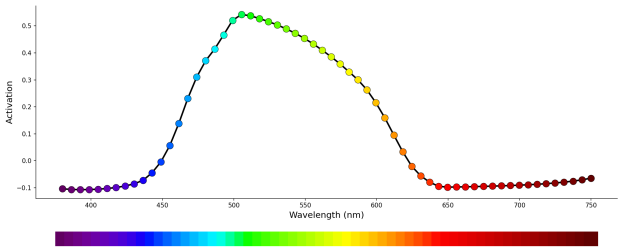
5.2.1. Unit 283: a green/yellow-selective neuron

Unit 283 showed a clear preference for the green spectral band. Table 6 summarizes the selectivity metrics. The Kruskal–Wallis test confirmed significant variation across bands ($H = 1432.85$, $p < 0.001$). The preferred band (green) achieved the highest CCMAS (0.408) and d' (2.451), and pairwise Mann–Whitney tests showed significantly higher activation than eight of the nine alternative bands. The only exception was chartreuse, the perceptually adjacent band, which did not differ significantly from green ($p = 0.088$), consistent with a graded response across neigh-

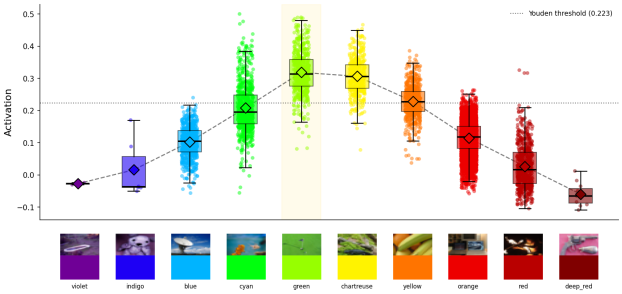
boring hues.

As a detector of the green band, unit 283 achieved perfect precision for violet, indigo, blue, red, and deep red, -but only moderate precision against cyan (0.722) and low precision against chartreuse and yellow (0.533 and 0.622). This pattern indicates that the unit responds strongly to green and its immediate neighbors on the hue circle, while reliably rejecting colors from the opposite side of the spectrum. The global precision of 0.336 at the Youden threshold confirms that the unit is not a clean detector of a single band, but rather exhibits partial selectivity localized to the green–yellow region of the spectrum. The full pairwise precision/recall matrix is provided in Appendix B.2.2.

Figures 2a–2b illustrate the quantitative results for unit 283. The synthetic tuning curve (Fig. 2a) reaches a peak at $\lambda = 505.42$ nm, consistent with the green band identification. The chromatic tuning curve on natural images (Fig. 2b) confirms this preference, with green and chartreuse bands eliciting the highest mean activations. The Youden threshold ($\theta = 0.2234$) admits images from green through orange while largely excluding blues and reds, consistent with the pairwise precision pattern reported above.



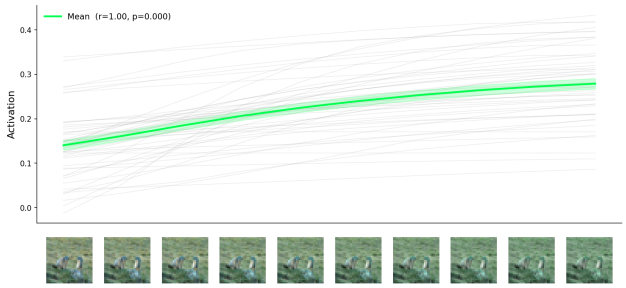
(a) Synthetic tuning curve.



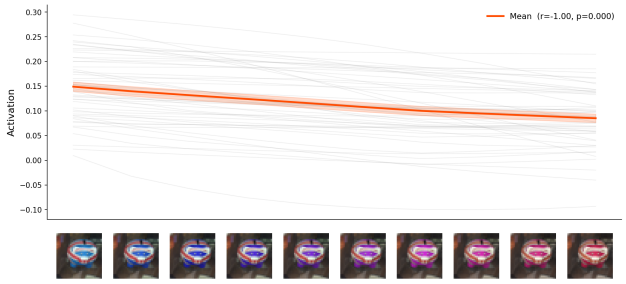
(b) Chromatic tuning curve (natural images).

Figure 2. Tuning curves for unit 283 (`encoder.block.0.mlp[1]` of ViT-B/16).

The results presented above provide strong evidence that unit 283 is color-selective for the green–yellow region of the spectrum. Table 8 summarizes the shift experiment used to evaluate the Falsifiability criterion. Shifting images toward



(a) Shift toward preferred wavelength ($\lambda = 505$ nm, green).



(b) Shift toward unpreferred wavelength ($\lambda = 630$ nm, orange-red).

Figure 3. Color shift intervention for unit 283 (`encoder.block.0.mlp[1]`, ViT-B/16). Gray lines show individual images; the colored line shows the mean activation across 50 randomly sampled images.

any wavelength in the interval $[474, 568]$ nm produced a significant increase in activation, with the maximum effect at the preferred wavelength of 505 nm (green). Outside this interval, shifts toward distant wavelengths had the opposite effect: activation *decreased* monotonically as the input moved toward 442 nm (blue-violet, $r = -0.952$) and 630 nm (orange-red, $r = -1.000$), exactly as the interpretation predicts.

Figure 3a shows the activation curves for 50 randomly sampled images from the chromatic test set, shifted toward 505 nm. Individual curves consistently trend upward with shift intensity, and the mean response rises steadily. Figure 3b shows the corresponding shifts toward 630 nm, where the individual and mean curves trend downward. Shift plots for the remaining wavelengths are provided in Appendix B.2.2.

5.3. ViT-B/16 - `encoder.ln`

We extended the analysis to `encoder.ln`, the final LayerNorm layer before the classification head of ViT-B/16. SFFS was unable to substantially impair the layer's color classification capacity: accuracy moved only from 0.767 to 0.658, suggesting that color information is distributed across

many units rather than concentrated in a few highly selective ones. Of 17 units flagged as relevant by SFFS, none satisfied the remaining criteria. Several units showed superficially high CCMAS values (e.g., unit 61: CCMAS $= 4.818$, but $d' = 0.083$; unit 189: CCMAS $= 1.272$, $d' = -1.469$, F1 $= 0$). In all cases, either the pairwise discrimination was limited to one or two bands or moderate scope, and in all cases, the shift experiment failed (see Appendix B.3 for additional details). This negative result is consistent with the view that deeper layers encode color entangled with semantic categories (Rafegas & Vanrell, 2017), making isolated chromatic selectivity difficult to validate at the unit level.

5.4. ViT-L/32 - `encoder.block.0.mlp[1]`

SFFS in `encoder.block.0.mlp[1]` of ViT-L/32 identified 25 relevant units, reducing accuracy from 0.838 to 0.608. Unit 1793, the top-ranked unit, proved to be a cyan-inhibitory neuron (see Appendix B.4.1 for full analysis).

6. Discussion

In this paper, we applied the proposed framework to critically analyze color neurons, arguably among the most basic visual features in computer vision. As discussed in the Related Work, color neurons have been extensively documented; using them as a starting point for rigorous validation is a natural choice.

Across three architectures—a convolutional network (ResNet50) and two Vision Transformers (ViT-B/16, ViT-L/32)—we identified units that are causally relevant to their layer's color representation and subjected them to the three criteria. We found units that satisfy all criteria: three excitatory unimodal neurons (green, orange, cyan), one bimodal neuron (blue–red), and one inhibitory neuron (suppressed by blue–cyan). However, we note that the strict 0.75 precision/recall threshold for a 'strong detector', while a useful reference following Gale *et al.* (2020), proved to be overly conservative for the present setting, where perceptual distances between color classes are not uniform. This criterion should be interpreted as a aspirational benchmark rather than a necessary condition; the converging evidence from the tuning curves, statistical tests, and shift experiments provides sufficient support for the reported interpretations.

The diversity of these profiles—spanning architectures, spectral regions, and excitatory/inhibitory modes—suggests that the protocol captures genuine functional specialization rather than artifacts of a particular model or color.

As a main finding, we consider that we have strong evidence that these units are correctly interpreted. The shift experiment, in particular, provides a causal test: activation increases when the input is moved toward the preferred wavelength and decreases when moved away. However, the domain of the interpretation is primarily restricted to images with a discernible predominant color. Global precision and F1 across all test images, including achromatic and ambiguous inputs, are relatively low, indicating that the interpretation should not be extended beyond $\mathcal{D}$ without further validation. However, the shift experiment generalizes across randomly sampled images from the test set, suggesting that the interpretation may extend beyond monochromatic stimuli.

In this sense, while Gale *et al.* (2020) framed unit selectivity as a binary detection problem, our results indicate that color-selective units can exhibit graded, multimodal, or inhibitory selectivity not captured by a single precision/recall threshold. The pairwise analysis introduced here provides finer characterization, revealing that a unit may be an excellent detector for certain spectral bands even if it responds to others.

Two clarifications are in order. First, validating one interpretation does not imply it is the only valid interpretation of the unit; a neuron may encode additional features beyond color. We do not address polysemanticity: a unit that passes all three criteria for color may still respond to non-chromatic features. Our framework validates a specific interpretation rather than claiming exhaustiveness. Second, the interpretation scheme presented here is not final and could be refined. For instance, some units may exhibit spatial selectivity in addition to chromatic tuning, a possibility our current stimuli are not designed to capture. The protocol, as implemented, favors "solid color" selectivity and may overlook more complex spatio-chromatic behaviors.

Finally, these findings suggest that some neurons behave as functions of their input in ways that can be formalized as testable hypotheses. Prior techniques, including feature visualization, can suggest such hypotheses; the present framework provides a means to test them rigorously. We believe interpretability could benefit from adopting a hypothesis-testing approach–slower, but constructive–in which claims about unit behavior are treated as falsifiable and subjected to explicit empirical criteria.

6.1. Limitations

This study has several explicit limitations. At a technical level, the color shift intervention may produce out-of-distribution samples (Grant & Tartaglini), and we did not control for potential confounds such as luminance or contrast variations introduced by the hue shift. The validation domain $\mathcal{D}$ is restricted to images with a discernible predominant color; the interpretations reported here should not be extended to achromatic or ambiguous inputs without further testing. Relevance is assessed at the layer level via a color classification probe, rather than through end-to-end ablation on the network's original task. While this choice

is motivated by the redundancy of low-level features (Morcos et al., 2018), it leaves open the question of whether the validated units are individually necessary for the model's object recognition performance. Finally, several spectral bands at the extremes of the visible spectrum (violet, indigo, deep red) contain very few samples in the test set, limiting statistical power for those bands.

7. Conclusion

We proposed three criteria for evaluating unit-level interpretations—completeness, falsifiability, and relevance—and demonstrated their application to color-selective neurons across three architectures. Five units satisfied all criteria, spanning excitatory, bimodal, and inhibitory selectivity. The results constitute, to our knowledge, the first rigorous validation of individual artificial neurons under explicit and demanding standards, and suggest that unit-level understanding is attainable despite current skepticism.

7.1. Future Work

The three criteria proposed here could be extended to domains beyond color, providing a general framework for evaluating unit-level interpretations—whether generated by automated explainers or by human inspection. Color was chosen as a deliberate starting point: its well-characterized perceptual structure makes the operationalization of completeness and falsifiability more direct than for most other visual attributes. Extending the framework to less easily parameterized concepts such as textures, object parts, or scene-level features will require concept-specific strategies for defining the domain $\mathcal{D}$ and designing falsifiable interventions. We leave these challenges for future work, but we believe the hypothesis-testing approach demonstrated here offers a template for rigorous interpretation across domains.

Software and Data

The code and documentation are available in the GitHub repository: https://github.com/Pherjev/Color.

Acknowledgments

The authors would like to thank the anonymous reviewers.

Funding

This work was made possible by the support of the Mexican government through SECIHTI, SNI, and the Instituto Politécnico Nacional (IPN) through IPN-SIP [grant number 2026-0078], IPN-COFAA, and EDI-IPN.

Impact Statement

This work advances mechanistic interpretability by providing rigorous criteria for validating unit-level interpretations. We hope to improve the diagnostic tools available to researchers and practitioners. However, as recent studies have shown, our understanding of unit-level behavior remains incomplete, and any interpretability method—including ours—may have limitations if over-relied upon. We encourage the community to build upon and critically evaluate the techniques presented here, ensuring that progress in interpretability aligns with robust scientific standards.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used DeepSeek to improve language, grammar, and stylistic clarity of the manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the final text.

References

Aguilar-Canto, F. and Calvo, H. Search of highly selective cells in convolutional layers with hebbian learning. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1455–1460. IEEE, 2023.

Arias, G., Baldrich, R., and Vanrell, M. Color in Visual-Language Models: CLIP deficiencies. *arXiv preprint arXiv:2502.04470*, 2025.

Bareeva, D., Höhne, M. M.-C., Warnecke, A., Pirch, L., Müller, K.-R., Rieck, K., Lapuschkin, S., and Bykov, K. Manipulating Feature Visualizations with Gradient Slingshots. *arXiv preprint arXiv:2401.06122*, 2024.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907375117. URL https://www.pnas.org/content/early/2020/08/31/1907375117.

Bianco, S. Enhancing color selectivity in foundation models for downstream color vision tasks. *Neurocomputing*, 645: 130471, 2025.

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Borowski, J., Zimmermann, R. S., Schepers, J., Geirhos, R., Wallis, T. S. A., Bethge, M., and Brendel, W. Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=QO9-y8also-.

Bruton, D. Color Science: RGB Values for Visible Wavelengths. Original at http://www.physics.sfasu.edu/astro/color/spectra.html; implementation available at https://www.eureca.de/5116-1-Bruton-color-mapping.html, 1996. Accessed: 06-05-2026.

Bun, L. M. and Horwitz, G. D. Color and luminance processing in V1 complex cells and artificial neural networks. *Color Research & Application*, 48(6):841–852, 2023.

Cammarata, N., Voss, C., Schubert, L., Goh, G., and Olah, C. Curve Detectors. *Distill*, 2020. doi: 10.23915/distill.00024.003. URL https://distill.pub/2020/circuits/curve-detectors/.

Chiu, M.-C., Wang, Y., Kim, D. E. G., Chen, P.-Y., and Ma, X. ColorSense: A Study on Color Vision in Machine Visual Recognition. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 681–697. IEEE, 2025.

De, K. Exploring Effects of Colour and Image Quality in Semantic Segmentation by Deep LearningMethods. *Electronic Imaging*, 35:1–10, 2022.

De Valois, R. L., Abramov, I., and Jacobs, G. H. Analysis of response patterns of lgn cells. *Journal of the optical Society of America*, 56(7):966–977, 1966.

Desimone, R., Albright, T. D., Gross, C. G., and Bruce, C. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4 (8):2051–2062, 1984.

Dorszewski, T., Tětková, L., Jenssen, R., Hansen, L. K., and Wickstrøm, K. K. From Colors to Classes: Emergence of Concepts in Vision Transformers. In *World Conference on Explainable Artificial Intelligence*, pp. 28–47. Springer, 2025.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Flachot, A. and Gegenfurtner, K. R. Processing of chromatic information in a deep convolutional neural network. *Journal of the Optical Society of America A*, 35(4):B334–B346, 2018.

Flachot, A. and Gegenfurtner, K. R. Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks. *Vision Research*, 182: 89–100, 2021.

Gale, E. M., Martin, N., Blything, R., Nguyen, A., and Bowers, J. S. Are there any 'object detectors' in the hidden layers of cnns trained to identify objects or scenes? *Vision Research*, 176:60–71, 2020.

Geirhos, R., Zimmermann, R. S., Bilodeau, B., Brendel, W., and Kim, B. Don't trust your eyes: on the (un) reliability of feature visualizations. *arXiv preprint arXiv:2306.04719*, 2023.

Goh, G., †, N. C., †, C. V., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. Multimodal Neurons in Artificial Neural Networks. *Distill*, 2021. doi: 10.23915/distill.00030. https://distill.pub/2021/multimodal-neurons.

Grant, S. and Tartaglini, A. R. Control and Predictivity in Neural Interpretability. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Harris, E., Mihai, D., and Hare, J. Spatial and Colour Opponency in Anatomically Constrained Deep Networks. *arXiv preprint arXiv:1910.11086*, 2019.

Harris, E., Mihai, D., and Hare, J. How Convolutional Neural Network Architecture Biases Learned Opponency and Color Tuning. *Neural Computation*, 33(4):858–898, 2021.

Hayne, L., Suresh, A., Jain, H., Kumar, R., and Carter, R. M. Much Easier Said Than Done: Falsifying the Causal Relevance of Linear Decoding Methods. *arXiv preprint arXiv:2211.04367*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hegdé, J. and Van Essen, D. C. Selectivity for Complex Shapes in Primate Visual Area V2. *The Journal of Neuroscience*, 20(5):RC61, 2000.

Hegdé, J. and Van Essen, D. C. A Comparative Study of Shape Representation in Macaque Visual Areas V2 and V4. *Cerebral cortex*, 17(5):1100–1116, 2007.

Huang, J., Geiger, A., D'Oosterlinck, K., Wu, Z., and Potts, C. Rigorously Assessing Natural Language Explanations of Neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 317–331, 2023.

Hubel, D. H. and Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

Ilg, U. J. The role of areas MT and MST in coding of visual motion underlying the execution of smooth pursuit. *Vision research*, 48(20):2062–2069, 2008.

Ito, M. and Komatsu, H. Representation of Angles Embedded within Contour Stimuli in Area V2 of Macaque Monkeys. *Journal of Neuroscience*, 24(13):3313–3324, 2004.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Linsley, D., Feng, P., and Serre, T. Better artificial intelligence does not mean better models of biology. *Trends in Cognitive Sciences*, 2025.

Mertens, L., Yargholi, E., Van den Stock, J., Op de Beeck, H., and Vennekens, J. Color-Dependent Prediction Stability of Popular CNN Image Classification Architectures. In *International Conference on Artificial Neural Networks*, pp. 491–503. Springer, 2023.

Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., and Botvinick, M. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*, 2018.

Mormann, F., Kornblith, S., Cerf, M., Ison, M. J., Kraskov, A., Tran, M., Knieling, S., Quian Quiroga, R., Koch, C., and Fried, I. Scene-selective coding by single neurons in the human parahippocampal cortex. *Proceedings of the National Academy of Sciences*, 114(5):1153–1158, 2017.

Nadler, E. O., Darragh-Ford, E., Desikan, B. S., Conaway, C., Chu, M., Hull, T., and Guilbeault, D. Divergences in color perception between deep neural networks and humans. *Cognition*, 241:105621, 2023.

Nanfack, G., Fulleringer, A., Marty, J., Eickenberg, M., and Belilovsky, E. Adversarial Attacks on the Interpretation of Neuron Activation Maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4315–4324, 2024.

Oikarinen, T. and Weng, T.-W. CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks. *arXiv preprint arXiv:2204.10965*, 2022.

Oikarinen, T. and Weng, T.-W. Linear Explanations for Individual Neurons. *arXiv preprint arXiv:2405.06855*, 2024.

Oikarinen, T., Yan, G., Kulkarni, A., and Weng, T.-W. Beyond Top Activations: Efficient and Reliable Crowdsourced Evaluation of Automated Interpretability. *arXiv preprint arXiv:2506.07985v2*, 2025.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. An Overview of Early Vision in InceptionV1. *Distill*, 2020. doi: 10.23915/distill.00024.002. https://distill.pub/2020/circuits/early-vision.

Pudil, P., Novovičová, J., and Kittler, J. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

Rafegas, I. and Vanrell, M. Color representation in CNNs: parallelisms with biological vision. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2697–2705, 2017.

Rafegas, I. and Vanrell, M. Color encoding in biologically-inspired convolutional neural networks. *Vision Research*, 151:7–17, 2018.

Taylor, J. and Xu, Y. Joint representation of color and form in convolutional neural networks: A stimulus-rich network perspective. *PLoS One*, 16(6):e0253442, 2021.

Ukita, J. Causal importance of low-level feature selectivity for generalization in image recognition. *Neural Networks*, 125:185–193, 2020.

Wurtz, R. H. Visual Receptive Fields of Striate Cortex Neurons in Awake Monkeys. *Journal of Neurophysiology*, 32(5):727–742, 1969.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

Zimmermann, R. S., Borowski, J., Geirhos, R., Bethge, M., Wallis, T. S. A., and Brendel, W. How Well do Feature Visualizations Support Causal Understanding of CNN Activations? In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in*

Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pp. 11730–11744, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/618faa1728eb2ef6e3733645273ab145-Abstract.html.

Zimmermann, R. S., Klein, T., and Brendel, W. Scale Alone Does not Improve Mechanistic Interpretability in Vision Models. *Advances in Neural Information Processing Systems*, 36, 2024.

Zini, S., Gomez-Villa, A., Buzzelli, M., Twardowski, B., Bagdanov, A. D., Van de Weijer, J., et al. Planckian Jitter: countering the color-crippling effects of color jitter on self-supervised training. *arXiv preprint arXiv:2202.07993*, 2022.

A. Methodological details

A.1. Dataset construction

Probing dataset construction details The five initial HSV hue ranges used for categorization are: red $[0°, 30°] \cup [330°, 360°]$, yellow $[60°, 90°]$, green $[90°, 150°]$, blue $[150°, 270°]$, and purple $[270°, 330°]$. An image is included only if the dominant hue range accounts for at least 60% of its pixels and the mean saturation across all pixels exceeds 0.5. After filtering the 35,000 training images, the purple category contained substantially fewer samples than the other four and was excluded. The remaining four classes were truncated to the size of the smallest class, yielding 139 images per class for training and 60 per class for evaluation (combining the 5,000-image validation split and the 10,000-image test split).

Spectral band definitions and natural image filtering Table 2 lists the ten spectral bands used throughout this work, spanning the visible spectrum from 380 to 750 nm. Synthetic stimuli are generated at the center of each band for the tuning curve experiments. Natural images are assigned to a band according to the procedure described in Subsection 4.4.

Table 2. Spectral bands used for synthetic and natural stimulus categorization.

Band	Range (nm)	Center (nm)
Violet	380–417	398.5
Indigo	417–454	435.5
Blue	454–491	472.5
Cyan	491–528	509.5
Green	528–565	546.5
Chartreuse	565–602	583.5
Yellow	602–639	620.5
Orange	639–676	657.5
Red	676–713	694.5
Deep red	713–750	731.5

Chromatic dataset composition Table 3 reports the number of images per spectral band in each split of the chromatic dataset. The strong skew toward the orange band reflects the well-documented bias of the ImageNet dataset toward orangish and bluish hues (Rafegas & Vanrell, 2018; Flachot & Gegenfurtner, 2018). Several bands at the extremes of the visible spectrum (violet, indigo, deep red) contain very few samples, as these wavelengths rarely appear as dominant colors in natural photographs. We retain these bands in the dataset for completeness of the spectral coverage, but results involving them should be interpreted with caution due to limited sample size.

Table 3. Number of images per spectral band in each split of the chromatic dataset.

Band	Train (35K)	Val (5K)	Test (10K)
Violet	2	1	1
Indigo	6	0	5
Blue	669	106	183
Cyan	751	110	167
Green	585	73	173
Chartreuse	495	73	145
Yellow	620	88	178
Orange	5,541	764	1,551
Red	625	106	209
Deep red	11	1	3
Total	9,305	1,322	2,615

A.2. Relevance criterion details

Sequential Floating Forward Selection In our experiments, we set the maximum number of neurons $M = 50$, backward interval $B = 5$, and backward threshold $\tau = 0.01$. The network itself is not retrained during ablation; only the probe is trained once on the train-test split of the activations and evaluated without retraining for each ablation configuration. This

ensures that performance changes reflect only the removal of the ablated units rather than compensatory adaptation. Ablation is performed by setting the output of the selected units to zero, as discussed in Section 4. The search terminates when M neurons have been selected, when accuracy falls to zero, or when no candidates remain.

Probing model The probing model is a single-layer linear classifier that maps the concatenated activations of all units in a layer to one of $n = 4$ color classes. Training uses the Adam optimizer (Kingma & Ba, 2014) with learning rate 10^{-3}, cross-entropy loss, and a batch size of 32 with shuffling for 20 epochs. Evaluation during SFFS uses a batch size of 64 without shuffling.

A.3. Completeness criterion details

Wavelength-to-RGB conversion Wavelengths are converted to RGB values using the Bruton algorithm (Bruton, 1996), which approximates the visible spectrum by fitting piecewise linear functions to the chromaticity coordinates of monochromatic light.

Statistical testing for spectral selectivity For each candidate unit, we test whether activation differs significantly across the ten spectral bands using a Kruskal–Wallis test. If significant at $\alpha = 0.05$, pairwise Mann–Whitney U-tests compare the preferred band (identified by the maximum CCMAS value) against each of the other nine bands. The preferred band is considered significantly selective only if all nine pairwise comparisons yield $p < 0.05$ after Bonferroni correction. This conservative criterion ensures that the unit responds more strongly to its preferred wavelength range than to any individual alternative.

As complementary measures of selectivity, we report the Class-Conditional Mean Activation Selectivity (CCMAS) (Morcos et al., 2018) and the sensitivity index d' from signal detection theory for each band. CCMAS is computed as $(\mu_A - \mu_{\neg A})/(\mu_A + \mu_{\neg A})$, where μ_A is the mean activation for images in the target band and $\mu_{\neg A}$ is the mean activation for all other spectral images. The d' index is computed as $(\mu_A - \mu_{\neg A})/\sqrt{(\sigma_A^2 + \sigma_{\neg A}^2)/2}$.

For the strong detector evaluation, following Gale *et al.* (2020), we apply a threshold of 0.75 to both recall and precision. Images from the preferred band are treated as positives and all other spectral images as negatives.

A.4. Falsifiability criterion details

A.4.1. Causal intervention via color shift

For a candidate unit whose preferred band has representative wavelength λ and target hue h_{target} (derived from λ via the Bruton algorithm (Bruton, 1996)), we randomly sample $m = 100$ images from the chromatic test set. Each image is transformed in $n = 10$ steps toward the target color using the following procedure. The image is converted to HSV space and, for each pixel, the hue component is linearly interpolated toward h_{target} with circular hue correction. The saturation and value components are left unchanged to preserve spatial structure. The shift intensity parameter $\alpha \in [0, 0.8]$ controls the degree of interpolation, with $\alpha = 0$ yielding the original image and $\alpha = 0.8$ a strongly shifted version.

The unit's activation is recorded for each α, producing an activation curve against shift intensity for each image. We compute the mean activation across images at each α and test for a monotonic relationship using Spearman's rank correlation. A significant positive correlation ($p < 0.05$, $r > 0$) constitutes evidence that the unit's activation tracks the proximity of the input to its preferred color.

As a control, the same procedure is applied shifting toward a distant band (opposite side of the hue circle) and toward a randomly selected non-adjacent band. The interpretation is considered falsified if the unit's activation increases significantly under any control shift.

All parameters use the default values: $n = 10$ steps, $\alpha \in [0, 0.8]$, $m = 100$ images, saturation fixed to original values.

A.4.2. Color shift target: wavelength vs. band center

The color shift experiment (Section 4.5) supports two modes for determining the target hue toward which images are shifted, controlled by the parameter `use_band_center`:

- **Band center mode** (`use_band_center=True`, default): Given a wavelength λ, the function identifies the spectral

band containing λ and uses the weighted center of its HSV hue interval as the shift target. This ensures consistency with the HSV-based categorization of natural images, where each band is defined by a fixed hue range rather than a single wavelength. This mode is appropriate when λ is the center of a spectral band (e.g., `wl2` in the main text).

- **Bruton mode** (`use_band_center=False`): The target hue is derived directly from λ via Bruton's wavelength-to-RGB conversion followed by RGB-to-HSV. This preserves the exact peak of the synthetic tuning curve without imposing band boundaries. This mode is appropriate when λ is the precise wavelength of maximum activation from the synthetic tuning curve.

In practice, we use band center mode for the main falsifiability evaluation, as the hypothesis under test is formulated in terms of the preferred spectral band. Bruton mode is available for exploratory analysis when the synthetic peak does not align cleanly with a single band. The shift procedure is otherwise identical in both modes: only the target hue changes; saturation and value are held constant, and the Spearman correlation test is applied to the mean activation curve.

A.5. Justification of the approach

A.5.1. Ablation procedure

Unlike Ukita (2020), who studied orientation selectivity–a feature ubiquitous across units in early layers–and could therefore evaluate relevance by ablating populations of units grouped by selectivity percentiles, our focus on unimodal color neurons requires a more targeted approach. Because such neurons are relatively sparse, we identify them individually through Sequential Floating Forward Selection, which directly measures the causal impact of each unit on color classification rather than relying on a precomputed selectivity metric. Future work could extend Ukita's population-level ablation approach to color selectivity using established selectivity indices (Rafegas & Vanrell, 2017; 2018), complementing the individual-unit SFFS validation presented here, but for now lies outside the objectives of the current research.

A.5.2. Color space

Unlike previous work that operated in model-dependent color spaces such as RGB_{PCA} (Flachot & Gegenfurtner, 2018) or OPP (Rafegas & Vanrell, 2017), we parametrize synthetic stimuli by wavelength (380–750 nm), a physical dimension independent of the network's training distribution. Wavelengths are converted to RGB via the standard visible-spectrum approximation of Bruton (1996). While this mapping is not colorimetrically exact, it provides a principled, externally grounded coordinate for characterizing chromatic tuning, and is sufficient to reveal relative selectivity across the spectrum. Expressing chromatic preference in physical units (nm) also facilitates comparison with the neurophysiological literature, where wavelength has been the standard independent variable since the foundational studies of De Valois et al. (1966).

For natural images, however, the Bruton approximation is not invertible: it maps only monochromatic stimuli, whereas natural colors are inherently broadband and do not correspond to a single wavelength. We therefore categorize natural images by their dominant hue in HSV space, a perceptually grounded and numerically stable representation. The ten HSV hue ranges used for categorization correspond to the ten spectral bands, and we report results using the band name and its approximate wavelength range. The synthetic tuning curve, generated from monochromatic stimuli, is reported in nanometers; natural image selectivity is reported by spectral band. The two scales are strongly correlated for saturated colors but are not identical, and this discrepancy is acknowledged throughout.

B. Supplementary results

B.1. ResNet50 - `layer1`

B.1.1. Unit 38

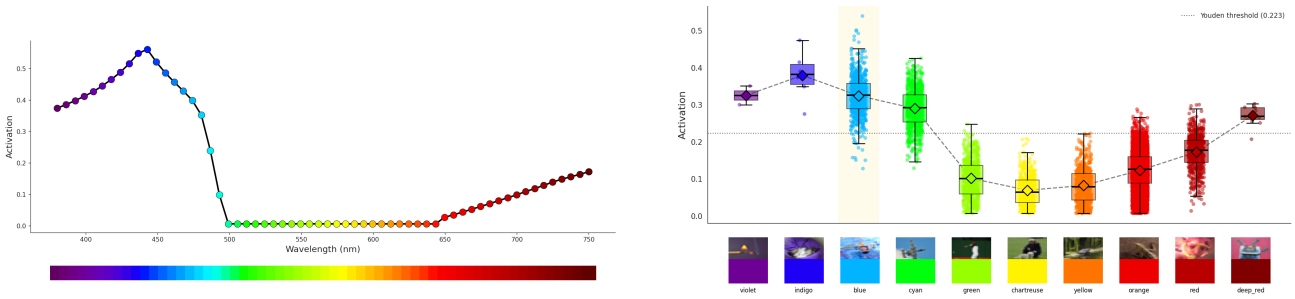
Unit 38 of ResNet50 (`layer1`) showed clear selectivity for cyan. The Kruskal–Wallis test was significant ($H = 1616.46$, $p < 0.001$). Cyan achieved the highest CCMAS (0.639) and d' (2.485). Pairwise Mann–Whitney tests confirmed significantly higher activation than eight of nine alternative bands. The only exception was violet ($p = 0.006$ vs. Bonferroni threshold ≈ 0.0056), likely due to the single violet image in the test set. Blue and green, the perceptually adjacent bands, also showed elevated CCMAS and d', consistent with a graded tuning curve centered on cyan. The Youden threshold ($\theta = 0.274$, TPR $= 0.973$, FPR $= 0.134$) yielded moderate global precision (0.308) and high recall (0.892), confirming partial selectivity localized to the blue–green region. See Table 4 for further details.

Table 4. Selectivity metrics for unit 38 (ResNet50, `layer1`). Preferred band: cyan. p-values are from one-sided Mann–Whitney U-tests against the cyan band. $\star$ denotes the preferred band; ✓ indicates significantly lower activation than cyan; ✗ indicates no significant difference.

Band	CCMAS	d'	MW p	Sig.
Violet	-0.769	-0.966	0.006	✗
Indigo	-0.370	-0.592	<0.001	✓
Blue	0.436	1.294	<0.001	✓
Cyan	0.639	2.485	—	$\star$
Green	0.553	2.029	<0.001	✓
Chartreuse	0.410	1.300	<0.001	✓
Yellow	-0.112	-0.210	<0.001	✓
Orange	-0.757	-1.724	<0.001	✓
Red	-0.579	-0.812	<0.001	✓
Deep red	-0.517	-0.740	<0.001	✓

B.1.2. Unit 39

Unit 39, identified by SFFS in the same layer as unit 38 (`layer1` of ResNet50), exhibited a bimodal selectivity pattern spanning both blue and red regions of the spectrum. The preferred band according to CCMAS was blue (CCMAS $= 0.416$, $d' = 2.970$), but violet (CCMAS $= 0.376$, $d' = 2.772$), indigo (0.440, 3.094), and deep red (0.295, 1.924) showed comparably high selectivity. Pairwise Mann–Whitney tests confirmed that violet, indigo, and deep red did not differ significantly from blue after Bonferroni correction (all $p > 0.0056$), while green through orange were strongly rejected (all $p < 0.001$). This pattern (depicted in Figure 4) suggests a double-color neuron jointly selective for short (blue–violet) and long (red) wavelengths, with suppression in the medium-wavelength region (green–yellow).



(a) Synthetic tuning curve. *(b)* Chromatic tuning curve (natural images).

Figure 4. Tuning curves for unit 39 (`layer1` of ResNet50).

The shift experiment (Table 5, Figure 5) confirmed the bimodal pattern. Shifts toward blue–violet (380–448 nm) produced strong positive correlations (all $r = 1.000$, $p < 0.001$), as did shifts toward red (693 nm, $r = 1.000$). In contrast, shifts toward the medium-wavelength region produced significant negative correlations (green, 505 nm: $r = -1.000$; yellow, 568 nm: $r = -1.000$). An intermediate wavelength (blue-cyan, 474 nm) showed a weaker but still positive effect ($r = 0.782$, $p = 0.008$), consistent with a transition zone between excitation and suppression.

B.2. ViT-B/16 - `encoder.block.0.mlp[1]`

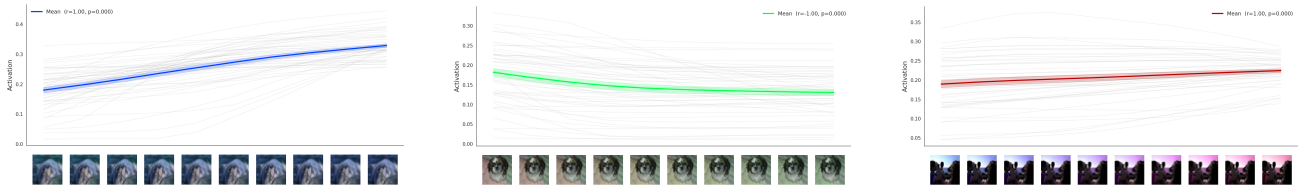
Figure 6 shows the ablation curve produced by the SFFS.

B.2.1. Unit 2437

Unit 2437 was the most causally relevant unit identified by SFFS: its ablation alone produced a drop of 0.1564 in probe accuracy. Despite this, it failed to exhibit interpretable chromatic tuning. The synthetic and natural tuning curves (Fig. 7) showed no clear spectral preference, with a maximum CCMAS of only 0.012 (orange band). The F1-score reached 0.739 for color images but dropped to 0.485 on the full test set, confirming weak absolute selectivity. Neither the synthetic peak

Table 5. Shift experiment for unit 39 (ResNet50, `layer1`). Preferred bands: blue ($\sim$ 442 nm) and red ($\sim$ 693 nm).

Target λ (nm)	r	p
380 (violet)	1.000	< 0.001
411 (violet)	1.000	< 0.001
436 (indigo)	1.000	< 0.001
442 (blue) (pref.)	1.000	< 0.001
Blue (HSV)	1.000	< 0.001
448 (blue)	1.000	< 0.001
474 (blue)	0.782	0.008
505 (green)	-1.000	< 0.001
568 (yellow)	-1.000	< 0.001
693 (red) (pref.)	1.000	< 0.001
Red (HSV)	1.000	< 0.001
Image subsets (shift to blue, 442 nm)		
Top-20 max blue	1.000	< 0.001
Top-20 min blue	1.000	< 0.001



(a) Shift toward preferred wavelength ($\lambda =$ 448 nm, blue).

(b) Shift toward unpreferred wavelength ($\lambda = 505$ nm, green).

(c) Shift toward preferred wavelength ($\lambda =$ 693 nm, red).

Figure 5. Color shift intervention for unit 39 (`layer1`, ResNet50). Gray lines show individual images; the colored line shows the mean activation across 50 randomly sampled images.

($\lambda = 750$ nm) nor the natural peak ($\lambda = 657$ nm) produced an increase in activation under the shift intervention (Spearman $r = -0.964$, $p < 0.005$ and $r = -0.600$, $p = 0.067$, respectively), failing the Falsifiability criterion (Fig. 8).

Unit 1762, selected by SFFS immediately after the ablation of unit 2437, showed nearly identical tuning properties and shift results, suggesting that it serves as a redundant contributor to the same non-chromatic aspects of the representation. Both units likely encode luminance or texture cues that correlate with color categories in the probing dataset without being genuinely color-selective.

B.2.2. Unit 283

This supplementary material for unit 283, the most extensively studied unit, contains the complete selectivity metrics (Tables 6 and 7), top and bottom activating images for the preferred class (Figure 9), full shift experiment results (Table 8), and the complete shift curve catalog (Figure 10). This provides a detailed view of the full analysis for a single unit; the metrics reported for the remaining units summarize the results of the corresponding analysis.

Table 7 reports the full pairwise precision, recall, and F1 scores for unit 283 at the Youden threshold ($\theta = 0.2234$). Each row treats the corresponding band as the positive class and all other bands as negatives. High precision against a given band indicates that the unit rarely activates above threshold for images of that band; low precision indicates frequent false positives. The pattern shows strong discrimination of green from the spectral extremes (violet–blue, red–deep red) and weaker discrimination from perceptually adjacent bands (cyan, chartreuse, yellow, orange).

Note that the green band (the unit's preferred band) shows the lowest precision (0.166), which is expected: the unit activates above threshold for 93.1% of green images, but many non-green images also exceed the threshold, reflecting the unit's graded response across the green–yellow region. Conversely, bands on the opposite side of the spectrum (violet, indigo, blue, red, deep red) achieve unit precision—the unit never activates above threshold for images outside these bands unless they contain green or yellow hues.

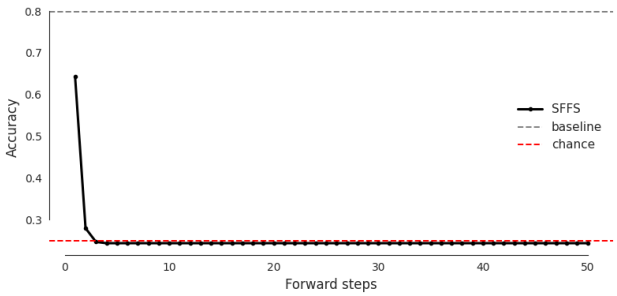
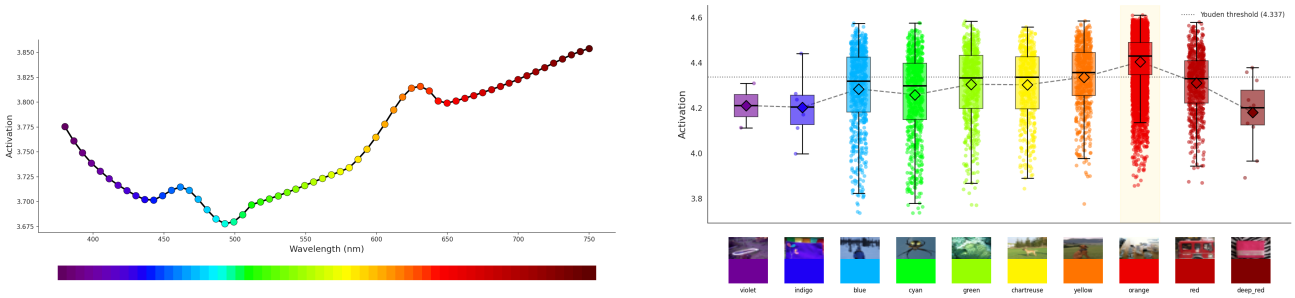


Figure 6. SFFS ablation curve for layer `encoder.block.0.mlp[1]` of ViT-B/16. The y-axis shows probe accuracy as units are sequentially ablated. The red dashed line marks chance level (25%).



(a) Synthetic tuning curve. *(b)* Natural tuning curve.

Figure 7. Tuning curves for unit 2437 (`encoder.block.0.mlp[1]`, ViT-B/16).

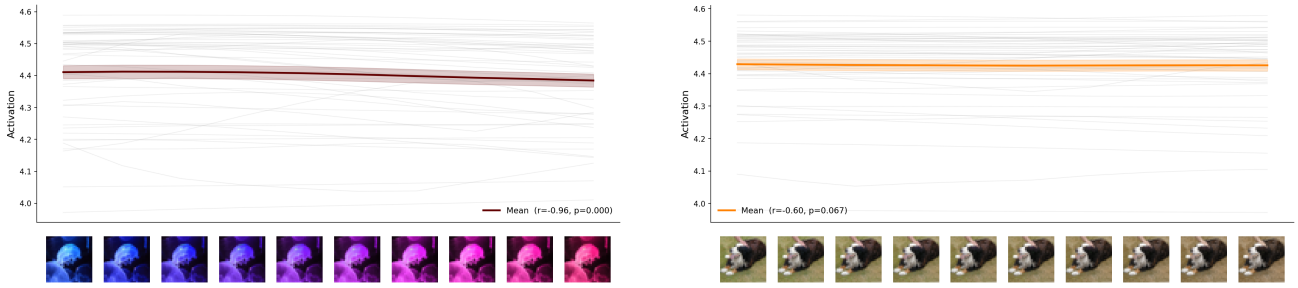
Table 6. Selectivity metrics for unit 283 (preferred band: green). p-values are from one-sided Mann–Whitney U-tests against the green band. $\star$ denotes the preferred band; ✓ indicates significantly lower activation than green. ✗ indicates no significant difference.

Band	CCMAS	d'	MW p	Sig.
Violet	-1.466	-2.598	0.006	✓
Indigo	-0.807	-1.459	<0.001	✓
Blue	-0.186	-0.613	<0.001	✓
Cyan	0.196	0.798	<0.001	✓
Green	0.408	2.451	—	$\star$
Chartreuse	0.384	2.322	0.088	✗
Yellow	0.241	1.174	<0.001	✓
Orange	-0.251	-0.829	<0.001	✓
Red	-0.707	-1.597	<0.001	✓
Deep red	-2.404	-2.908	<0.001	✓

Inspection of the top-20 maximally activating images (Fig. 9a) provides complementary qualitative evidence: all top-activating inputs are dominated by green or cyan-green regions. Conversely, the top-20 minimally activating images *among those categorized as green* (Fig. 9b) reveal substantial red, indigo, or violet regions alongside green, suggesting that these distant hues suppress the unit's response even when green is present. This pattern is consistent with a cell that is excited by medium wavelengths and inhibited by spectral extremes.

B.2.3. Unit 2299

Unit 2299 was the fourth most relevant unit identified by SFFS in the same layer (`encoder.block.0.mlp[1]` of ViT-B/16). It showed selectivity for the orange band, with a preferred wavelength of approximately 658 nm. The Kruskal–

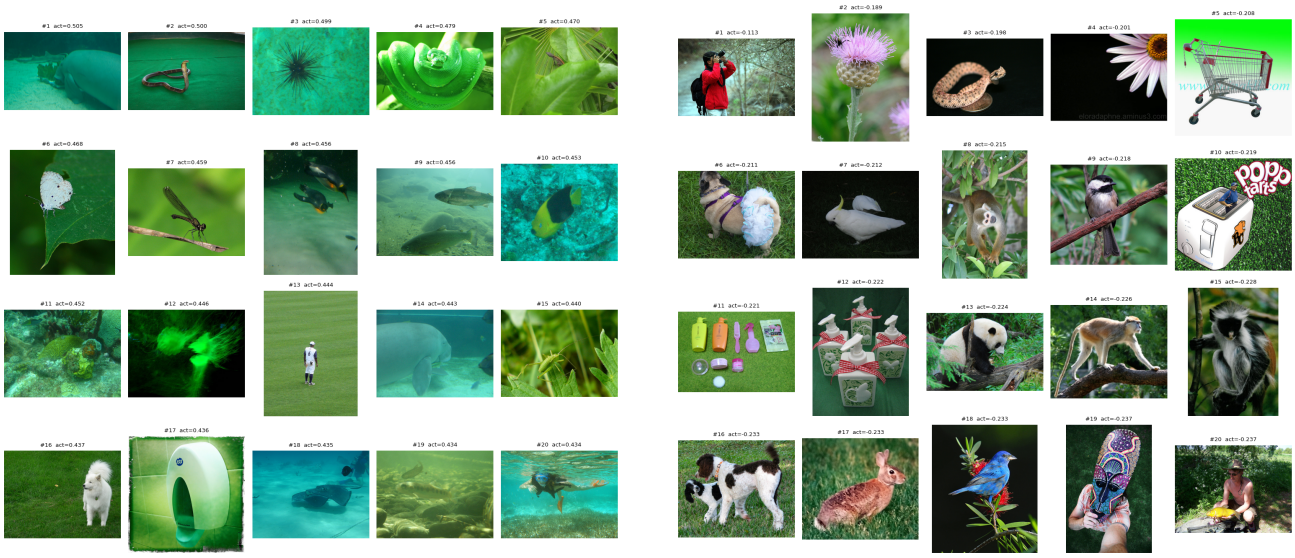


(a) Shift toward $\lambda = 750$ nm. *(b)* Shift toward $\lambda = 657$ nm.

Figure 8. Color shift intervention for unit 2437. Neither target produced a positive trend.

Table 7. Unit 283: pairwise precision, recall, and F1 by spectral band (threshold $\theta = 0.2234$).

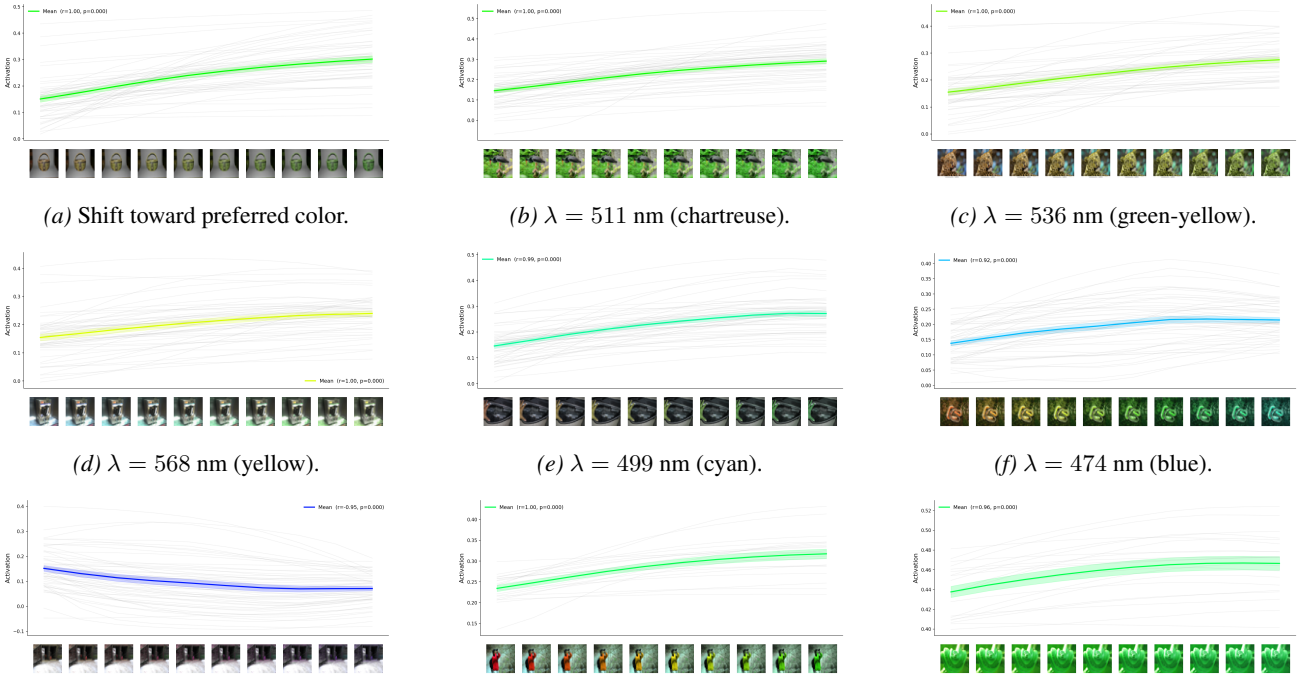
Band	Precision	Recall	F1
Violet	1.000	0.931	0.964
Indigo	1.000	0.931	0.964
Blue	0.994	0.931	0.961
Cyan	0.722	0.931	0.813
Green	0.166	0.931	0.281
Chartreuse	0.533	0.931	0.678
Yellow	0.622	0.931	0.745
Orange	0.910	0.931	0.920
Red	1.000	0.931	0.964
Deep red	1.000	0.931	0.964
Global	0.336	0.931	0.494



(a) Top-20 maximally activating inputs. *(b)* Top-20 minimally activating green inputs.

Figure 9. Top-activating and bottom-activating images for unit 283 in the selected class. Taken from the testing set.

Wallis test confirmed significant variation across bands ($H = 1403.80$, $p < 0.001$). The orange band achieved the highest CCMAS (0.309), although nearby bands (chartreuse, yellow, green) did not differ significantly from orange in pairwise



(a) Shift toward preferred color. *(b)* $\lambda = 511$ nm (chartreuse). *(c)* $\lambda = 536$ nm (green-yellow).

(d) $\lambda = 568$ nm (yellow). *(e)* $\lambda = 499$ nm (cyan). *(f)* $\lambda = 474$ nm (blue).

(g) $\lambda = 442$ nm (blue-violet). *(h)* Top-20 min, shift to 505 nm. *(i)* Top-20 max, shift to 505 nm.

Figure 10. Complete color shift results for unit 283 (`encoder.block.0.mlp[1]`, ViT-B/16). The top row (panels a–c) shows shifts toward the preferred green–yellow region. The middle row (d–f) shows the transition toward blue–cyan. The bottom row (g–i) shows the distant control (blue-violet) and the top-20 subsets shifted toward 505 nm. All subfigures follow the same format: gray lines = individual images, colored line = mean activation.

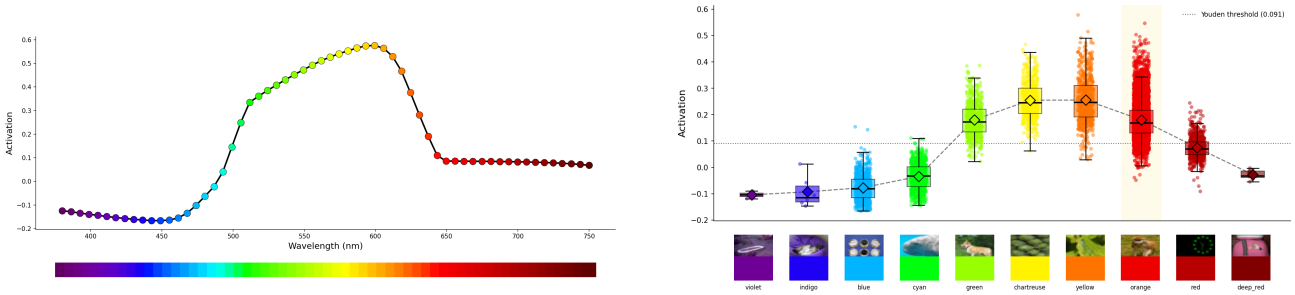
Table 8. Shift experiment for unit 283 (ViT-B/16, `encoder.block.0.mlp[1]`). Preferred band: green ($\sim$ 505 nm).

Target λ (nm)	r	p
442 (blue-violet)	-0.952	< 0.001
474 (blue)	0.915	< 0.001
499 (cyan)	0.988	< 0.001
505 (green) (pref.)	1.000	< 0.001
Green (HSV)	1.000	< 0.001
511 (chartreuse)	1.000	< 0.001
536 (green-yellow)	1.000	< 0.001
568 (yellow)	1.000	< 0.001
630 (orange-red)	-1.000	< 0.001
Image subsets (shift to green, 505 nm)		
Top-20 max green	0.964	< 0.001
Top-20 min green	1.000	< 0.001

Mann–Whitney tests (all $p > 0.05$), indicating broadly tuned selectivity across the yellow–orange–red region. Bands on the opposite side of the spectrum (violet through blue) were strongly rejected (all $p < 0.001$). At the Youden threshold ($\theta = 0.0914$), unit 2299 achieved a precision of 0.722 and recall of 0.934 in color images, performing substantially better as an orange detector than unit 283 did for green. Figures 11a and 11b depict the tuning curves for this unit. Table 9 contains more detailed information for the selectivity metrics.

Table 9. Selectivity metrics for unit 2299 (ViT-B/16, `encoder.block.0.mlp[1]`). Preferred band: orange. p-values are from one-sided Mann–Whitney U-tests against the orange band. $\star$ denotes the preferred band; ✓ indicates significantly lower activation than orange; ✗ indicates no significant difference.

Band	CCMAS	d'	MW p	Sig.
Violet	-6.077	-3.030	< 0.001	✓
Indigo	-4.521	-2.634	< 0.001	✓
Blue	-2.803	-3.027	< 0.001	✓
Cyan	-1.541	-2.357	< 0.001	✓
Green	0.114	0.389	0.690	✗
Chartreuse	0.292	1.214	1.000	✗
Yellow	0.299	1.153	1.000	✗
Orange	0.309	0.738	—	$\star$
Red	-0.339	-0.861	< 0.001	✓
Deep red	-1.468	-2.099	< 0.001	✓



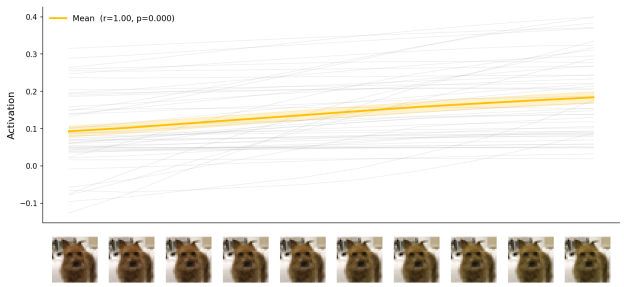
(a) Synthetic tuning curve. *(b)* Chromatic tuning curve (natural images).

Figure 11. Tuning curves for unit 2299 (`encoder.block.0.mlp[1]` of ViT-B/16).

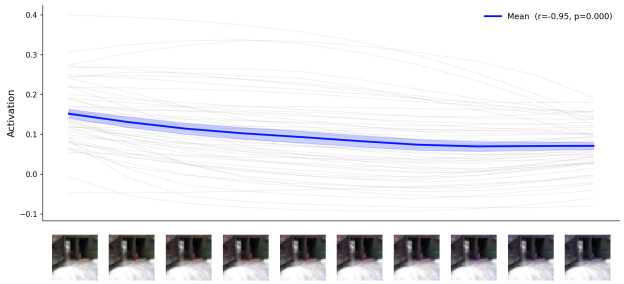
The color shift intervention confirmed the orange preference (Table 10). Shifting images toward the orange region (593–605 nm) produced significant positive correlations (all $r = 1.000$, $p < 0.001$), with the maximum effect at the preferred wavelengths. Shifts toward distant bands—blue (474 nm) and deep red (662–724 nm)—produced significant *negative*

Table 10. Shift experiment for unit 2299 (ViT-B/16, `encoder.block.0.mlp[1]`). Preferred band: orange ($\sim$ 599 nm).

Target λ (nm)	r	p
474 (blue)	-0.988	< 0.001
536 (green-yellow)	1.000	< 0.001
568 (yellow)	1.000	< 0.001
599 (orange) (pref.)	1.000	< 0.001
Orange (HSV)	1.000	< 0.001
630 (red)	-0.188	0.603
662 (deep red)	-1.000	< 0.001
724 (deep red)	-1.000	< 0.001
Image subsets (shift to orange, 599 nm)		
Top-20 max orange	1.000	< 0.001
Top-20 min orange	1.000	< 0.001



(a) Shift toward preferred wavelength ($\lambda = 599$ nm, chartreuse).



(b) Shift toward unpreferred wavelength ($\lambda = 474$ nm, blue; shown for unit 2299).

Figure 12. Color shift intervention for unit 2299 (`encoder.block.0.mlp[1]`, ViT-B/16).

correlations, indicating that the unit's activation decreases as the input moves away from orange. An intermediate wavelength (630 nm, red) produced no significant change ($r = -0.188$, $p = 0.603$), consistent with a transition zone between excitation and inhibition. The effect held for both top-20 maximally and minimally activating orange images (both $r = 1.000$, $p < 0.001$). See figures 12a and 12b for complementary examples.

The pattern mirrors that of unit 283: monotonic excitation toward the preferred spectral region, monotonic suppression toward the opposite side of the spectrum, and a graded transition zone in between. Together, units 283 and 2299 demonstrate that our protocol can identify and validate multiple color-selective units within the same layer, responding to different regions of the visible spectrum.

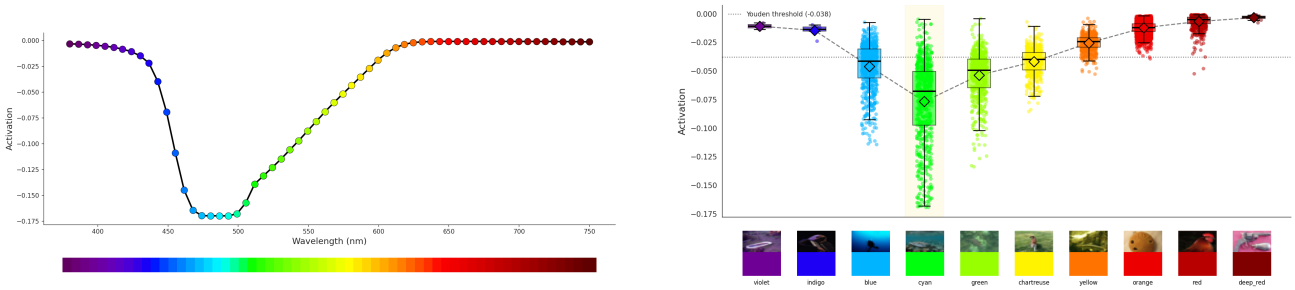
B.3. ViT-B/16 - `encoder.ln`

We extended the analysis to `encoder.ln`, the final LayerNorm layer before the classification head of ViT-B/16. SFFS identified 17 relevant units, but probe accuracy moved only from 0.767 to 0.658, indicating that color information is distributed across many units rather than concentrated in a few. None of the 17 units satisfied the three criteria. We group them by failure mode:

CCMAS artifact. Five units (61, 384, 49, 106, 370) showed superficially high CCMAS values but negligible d' (≤ 0.1 in absolute value). For example, unit 61 reached CCMAS $= 4.818$ with $d' = 0.083$ and discriminated only red and deep red from its preferred band.

Insufficient pairwise discrimination. Seven units (493, 212, 347, 238, 491, 674, 189) showed significant differences against only one or two bands, insufficient for a spectral interpretation. Unit 493 ($d' = -0.283$, inhibitory) discriminated just two bands; unit 189 achieved CCMAS $= 1.272$ and $d' = -1.469$ yet obtained F1 $= 0$.

Excitatory candidates with limited scope. The remaining four units (438, 239, 104, 553) showed excitatory preferences



(a) Synthetic tuning curve.

(b) Chromatic tuning curve (natural images).

Figure 13. Tuning curves for unit 1793 (`encoder.block.0.mlp[1]`, ViT-L/32).

with moderate d' but discriminated fewer than three bands, falling short of the completeness threshold.

In all cases, the shift experiment failed to produce a significant positive trend. This negative result is consistent with the view that deeper layers encode color entangled with semantic categories (Rafegas & Vanrell, 2017), making isolated chromatic selectivity difficult to validate at the unit level.

B.4. ViT-L/32 - `encoder.block.0.mlp[1]`

B.4.1. Unit 1793

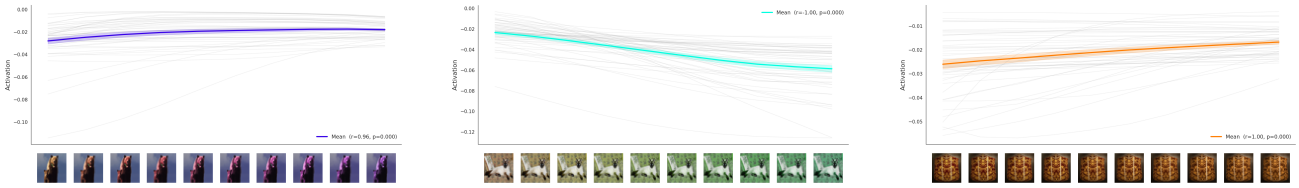
Unit 1793, identified by SFFS in the first MLP layer of ViT-L/32 (`encoder.block.0.mlp[1]`), exhibited a novel inhibitory selectivity pattern. The preferred band according to CCMAS was cyan (CCMAS $= 0.587$), but the d' values revealed a crucial distinction: cyan and neighboring bands (blue through yellow) showed *negative* d', while distant bands (orange through violet) showed *positive* d'. This means the unit is suppressed by cyan–yellow and excited by spectral extremes—the inverse of the pattern observed in units 283 and 38.

The shift experiment (Table 11) confirmed the inhibitory pattern. Shifts toward blue–cyan–green wavelengths produced significant *negative* correlations (all $r = -1.000$, $p < 0.001$), while shifts toward the complementary orange–red region produced positive correlations ($r = 1.000$, $p < 0.001$). The effect was robust across image subsets: top-20 maximally activating cyan images ($r = -1.000$) and minimally activating cyan images ($r = -0.988$) both showed the same inhibitory trend (Figure 14).

Table 11. Shift experiment for unit 1793 (ViT-L/32, `encoder.block.0.mlp[1]`). Preferred band: cyan ($\sim$ 492 nm), inhibitory.

Target λ (nm)	r	p
430 (indigo)	0.964	< 0.001
461 (blue)	-1.000	< 0.001
486 (blue)	-1.000	< 0.001
492 (cyan) (unpref.)	$\mathbf{-1.000}$	$\mathbf{< 0.001}$
Cyan (HSV)	$\mathbf{-1.000}$	$\mathbf{< 0.001}$
499 (cyan)	-1.000	< 0.001
524 (green)	-1.000	< 0.001
555 (green)	-1.000	< 0.001
618 (orange)	1.000	< 0.001
743 (deep red)	1.000	< 0.001
Image subsets (shift to cyan, 492 nm)		
Top-20 max cyan	-1.000	< 0.001
Top-20 min cyan	-0.988	< 0.001

This unit exhibits a functionally distinct profile from those found in ViT-B/16 and ResNet50: rather than being excited by a preferred wavelength, it is suppressed by a specific color. Figure 13 illustrates the tuning curves for this unit.



(a) $\lambda = 430$ nm (indigo). *(b)* $\lambda = 492$ nm (cyan, inhibitory). *(c)* $\lambda = 618$ nm (orange).

Figure 14. Color shift intervention for unit 1793 (`encoder.block.0.mlp[1]`, ViT-L/32).

Algorithm 1 Sequential Floating Forward Selection (SFFS) for Ablation

Input: probe P, candidate pool $\mathcal{C}$, max neurons M, backward interval B, threshold τ
Output: selected set $\mathcal{S}$
$\mathcal{S} \leftarrow \emptyset$
$acc_{\text{base}} \leftarrow$ accuracy of P with no ablation
$step \leftarrow 0$
while $|\mathcal{S}| < M$ **and** $\mathcal{C} \neq \emptyset$ **do**
  // Forward step: add the most damaging neuron
  $best \leftarrow$ None, $best_acc \leftarrow \infty$
  for each $n \in \mathcal{C}$ **do**
    $acc \leftarrow$ accuracy of P with $\mathcal{S} \cup \{n\}$ ablated
    if $acc < best_acc$ **then**
      $best_acc \leftarrow acc,\ best \leftarrow n$
    end if
  end for
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{best\},\ \mathcal{C} \leftarrow \mathcal{C} \setminus \{best\}$
  $step \leftarrow step + 1$
  if $best_acc = 0$ **then**
    break
  end if
  if $step \bmod B = 0$ **and** $|\mathcal{S}| > 1$ **then**
    // Backward pruning: remove redundant neurons
    $improved \leftarrow$ True
    while $improved$ **and** $|\mathcal{S}| > 1$ **do**
      $improved \leftarrow$ False, $worst \leftarrow$ None, $worst_\Delta \leftarrow -\infty$
      $acc_{\mathcal{S}} \leftarrow$ accuracy of P with $\mathcal{S}$ ablated
      for each $n \in \mathcal{S}$ **do**
        $acc_{\setminus n} \leftarrow$ accuracy of P with $\mathcal{S} \setminus \{n\}$ ablated
        $\Delta \leftarrow acc_{\setminus n} - acc_{\mathcal{S}}$
        if $\Delta > worst_\Delta$ **then**
          $worst_\Delta \leftarrow \Delta,\ worst \leftarrow n$
        end if
      end for
      if $worst_\Delta \leq \tau$ **then**
        $\mathcal{S} \leftarrow \mathcal{S} \setminus \{worst\},\ \mathcal{C} \leftarrow \mathcal{C} \cup \{worst\}$
        $improved \leftarrow$ True
      end if
    end while
  end if
end while
return $\mathcal{S}$