Computational evidence for an inverse relationship between retinal and brain complexity

Mitchell B. Slapik

Department of Neurobiology and Anatomy,
McGovern Medical School,
University of Texas–Houston,
Houston, TX, USA

Visual neuroscientists have long observed an inverse relationship between brain and retinal complexity: As brain complexity increases across species, retinas adapt to simpler visual processing. Lindsey et al. previously provided a computational explanation for this pattern, showing that shallow networks encode complex features in their first stage of processing, whereas deep networks encode simpler features. Here, these findings are extended to a suite of representational analyses and show that shallow networks generate high-dimensional representations with linear decision boundaries and specific visual features that can feed directly into behavioral responses. In contrast, deep networks generate low-dimensional representations with nonlinear decision boundaries and general visual features. These representations require further processing before they can produce the appropriate behavioral response. In summary, the findings extend a longstanding principle linking simpler retinal features to complex brains and offer a computational framework for understanding neural network behavior more generally.

Introduction

Visual neuroscience has long noted an intriguing trend: As the brain has evolved greater complexity, the retina has evolved to prioritize simpler visual features (Johnston & Lagnado, 2012). Species with smaller brains, such as frogs and locusts, often process remarkably complex features within the retina. For example, frogs famously have "fly-detector" cells that detect small, dark objects making jerky movements (Lettvin, Maturana, Maturana, Mcculloch, & Pitts, 1959). Locusts have specialized detectors of expanding shadows (Gabbiani, Krapp, & Laurent, 1999). Mice have highly specific detectors for the movement of a hawk in the sky (Zhang, Kim, Sanes, & Meister, 2012).

However, as brain complexity increased in primates, the retina adapted to process simpler visual features. Although complex detectors persisted in the retinal periphery, the midget ganglion cell—a simpler detector—emerged in the fovea. These cells detect relatively simple visual features, such as small spots of light. The brain later combines these spots into lines, which then coalesce into shapes, and so on, eventually forming full objects (Felleman & Van Essen, 1991), but this unfolds across many stages of processing in an extensive neural network. Thus, whereas a frog detects a fly directly in the retina, a primate uses an extensive neural network spanning multiple brain regions.

However, it remains uncertain whether this trend reflects causation or mere correlation. A vast range of variables change when we make comparisons across species. Different species may occupy diverse ecological niches with distinct environments, predators, and prey. They may rely on other senses such as touch and hearing to compensate for vision. Thus, an observed inverse correlation between brain and retinal complexity may reflect causation or merely track an external variable. In fact, any observational study of different species may struggle to control for confounding factors and disentangle correlation from causation.

Artificial neural networks offer a unique framework for addressing this question. Convolutional neural networks provide an established model of the visual cortex. These networks emulate biological vision systems through hierarchical processing, retinotopic organization, and robust object recognition (Schrimpf et al., 2020). Numerous studies have used these networks to model brain responses to natural images (Cadena et al., 2019; Yamins et al., 2014). In fact, they have even been used to decode visual stimuli from neural activations or, conversely, generate specialized images that evoke precise patterns of activation across the visual cortex (Bashivan, Kar, & DiCarlo, 2019; Cadena et al., 2019). However, unlike animals that live in vastly different environments and ecological niches, we

Citation: Slapik, M. B. (2025). Computational evidence for an inverse relationship between retinal and brain complexity. *Journal of Vision*, *25*(8):9, 1–12, https://doi.org/10.1167/jov.25.8.9.

https://doi.org/10.1167/jov.25.8.9 Received December 2, 2024; published July 7, 2025 ISSN 1534-7362

can avoid these differences in artificial networks by exposing them to the exact same dataset and computational task. Here, this study took advantage of this property by creating models of different depths while controlling for architecture, optimizer, loss function, and dataset. This approach isolates the effect of network depth on the first layer of processing, an experiment not feasible in biological systems.

This study follows a landmark paper by Lindsey, Ocko, Ganguli, and Deny (2019) which showed that shallow networks encode complex features for separating object classes in their first stage of processing, whereas deep networks encode simple features for transferring maximal information. Interestingly, they also added a bottleneck simulating the optic nerve. This bottleneck generated center–surround receptive fields in the first stage of processing, reminiscent of lateral geniculate nucleus (LGN), and edge detectors in their second stage of processing, reminiscent of V1 cells. This work expands on these findings, offering additional insights into feature selectivity and representational geometry. Specifically, rather than examining the complexity of first-layer features, their selectivity across input images was examined. This study also shows how selectivity is strongly influenced by other variables, such as the number of first-layer filters and output classes, which have direct biological relevance. It also examined both linear and nonlinear decoding accuracy at the population level, as well as the dimensionality of neural representations, which provides a key insight into how deep and shallow networks use different strategies for object recognition.

Overall, this study confirmed an inverse relationship between brain and retina complexity. Specifically, shallow networks were found to encode more selective features in their first layer of processing that directly relate to behavioral responses. Meanwhile, deep networks encoded more general features that require additional stages of processing before they can be turned into the appropriate behavioral response. Other relevant factors include the number of filters in the first layer, which are associated with increased selectivity, and the number of classes, which is associated with decreased selectivity. Similar results were observed with regard to these neural representations at a population level, showing that shallow networks have optimized the geometry and dimensionality in early layers for immediate behavior response, whereas deep networks have optimized them for further processing (Bernardi et al., 2020; Boyle, Posani, Irfan, Siegelbaum, & Fusi, 2024). Ultimately, these findings may extend beyond vision, providing a new principle for how neural networks function across both biological and artificial contexts.

Methods

Architecture

The visual cortex was modeled using LeNet5, a convolutional neural network designed by LeCun et al. (1989) in 1998 (Figure 1B). This network consists of two convolutional layers, followed by three fully-connected layers. The original sigmoid nonlinearities were replaced with rectified linear units (ReLUs) to enhance performance. The full architecture is as follows:

- Input: 32 × 32 × 3 (32 × 32 pixels with three color channels)
- Convolutional layer 1 (Conv1): six features of size 5 × 5 → 28 × 28 × 6
- Pooling layer 1 (Pool1): 2 × 2 features with a stride of 2 → 14 × 14 × 6
- Convolutional layer 2 (Conv2): 16 features of size 5 × 5 → 10 × 10 × 16
- Pooling layer 2 (Pool2): 2 × 2 features with a stride of 2 → 5 × 5 × 16
- Fully connected layer 1 (Fc1): 120 neurons → 120
- Fully connected layer 2 (Fc2): 84 neurons → 84
- Fully connected layer 3 (Fc3): 10 neurons → 10

Although this network is simple, it enables cleanly addressing our question and avoiding the potential confounds of more complex architectures such as skip connections or batch normalization. To assess the influence of network depth on the first stage of processing, the depth of LeNet was systematically altered while maintaining a consistent final layer to predict image classes (Figure 1A). The model architectures were as follows.

Depth	Convolutional layers	Fully connected layers
2	1	1
3	2	1
4	2	2
5	2	3

Dataset

The networks were trained on the CIFAR10 dataset, developed by the Hinton lab in 2009 (Krizhevsky, 2009). This dataset contains 32 × 32 color pictures of 10 different objects, including animals such as birds, cats, and dogs, and vehicles such as airplanes, trucks, and ships. In total, there are 60,000 images, consisting of 50,000 training images and 10,000 test images. The dataset captures a representative sample

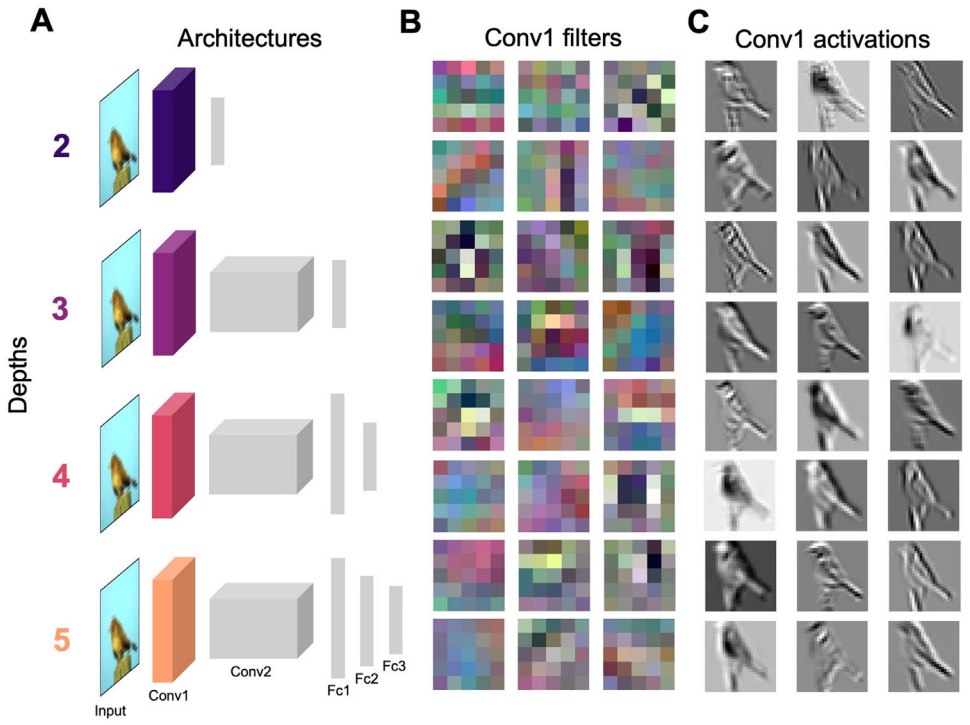

Figure 1. Experiment design. (**A**) Depths of different networks used in the experiment. Colors are used throughout to indicate the depth of different networks. Three-dimensional boxes represent convolutional layers, and two-dimensional strips represent fully connected layers. (**B**) Example features from the first layer of networks with different depths. (**C**) Example activations from the first layer of networks with different depths.

of the visual world and is small enough for quick training.

Training

Ten networks of each depth were trained for 10 epochs using a learning rate of 0.001, the Adam optimizer, and cross-entropy loss. Consistent with expectations, deeper networks demonstrated higher accuracy compared with shallower ones. The shallow networks plateaued at approximately 60% accuracy on the training data and 55% accuracy on the test data. Meanwhile, the deeper networks plateaued at approximately 65% accuracy of the training data and 60% on the test data. This suggests that around 5% of the training accuracy came from overfitting (see the full training and testing curves in Supplementary Figure S1). This experiment was conducted using Python 3.9.13, scikit-learn version 1.0.2, and PyTorch version 2.2.0.

Measures of feature selectivity

Feature selectivity in the first processing stage was evaluated using supervised and unsupervised metrics. The supervised measure of selectivity quantified the preference of a feature for its preferred class relative to other classes, according to the following equation:

$$Class\ selectivity = \frac{\mu_{max} - \mu_{-max}}{\mu_{max} + \mu_{-max} + \varepsilon}$$

Here, μ_{max} is the average activation for the preferred class, μ_{-max} is the average activation for the remaining classes, and ε is a small constant used to prevent dividing by zero. The selectivity of each filter was calculated across the entire CIFAR10 testing dataset for each location in the image, and the maximum selectivity value of that filter was then used, which came from a single location in the image. This value was considered to be the most important for downstream image classification.

Meanwhile, the unsupervised measure of selectivity looked at the preference of a feature for its preferred subset of images relative to the remaining images, regardless of their labeled class. This was measured using kurtosis, the fourth moment or "tailedness" of the activation distribution, which is defined as follows:

$$Kurtosis = \frac{E(x - \mu)^4}{\sigma^4}$$

Here, again, the kurtosis for each filter was calculated across the entire CIFAR10 test set at every location in the image and then selected the maximum kurtosis across all locations as the most informative value for downstream decoding.

Linear discriminant analysis

These changes in individual neurons were also examined to see if they affected how information is encoded at a population level. This structure or "geometry" determines what information can be directly accessed by a downstream neuron and what information cannot be directly accessed, and it requires additional stages of processing (Fusi, Miller, & Rigotti, 2016). Dimensionality reduction techniques were employed to visualize the representational geometry in the first stage of processing. Specifically, linear discriminant analysis (LDA) was used to determine the dimensions that maximally separate object classes.

Logistic regression

When looking at a neural representation, we can distinguish between information that is directly accessible to downstream neurons and information that is relatively "encrypted" or not directly accessible. Specifically, a simple linear decoder can measure accessible information, but nonlinear decoders measure relatively "encrypted" information (Fusi et al., 2016). For the linear decoder, a logistic regression was used with a liblinear solver. This decoder was applied to the Conv1 activations on the CIFAR10 test set in models of different depths, using a train–test split of 80/20.

k-nearest neighbors

For the nonlinear decoder, a k-nearest neighbor classifier was used, where $k = 10$, to match the number of classes (Fusi et al., 2016). Again, this decoder was applied to the Conv1 activations on the CIFAR10 test set in models of different depths, using a train–test split of 80/20.

Participation ratio

Finally, the mechanism behind pattern separation in these different networks was investigated. Prior work has shown that high dimensionality helps separate different patterns of inputs (Bernardi et al., 2020; Boyle et al., 2024). Here, dimensionality was computed by applying the participation ratio to the eigenvalues of the Conv1 activations on the CIFAR10 test set in networks for different depths (Gao et al., 2017):

$$Participation\ ratio = \frac{(\sum \lambda)^2}{\sum \lambda^2}$$

Results

This study examined the idea that brain complexity and retina complexity are inversely related. Therefore, a series of neural networks of different depths was created while maintaining the dataset of natural images, loss function, optimizer, and architecture; Figure 1A shows which layers were included and excluded from each model. This allowed isolating the effect of neural network depth, the proxy for brain complexity, on the first processing stage, Conv1, the proxy for the retina (Figure 1C).

In early processing, shallow networks extract specific feature and deep networks extract more general features

How depth modulates feature selectivity in the first stage of processing was examined first. Shallow networks were found to have more selective features in the sense that they activated strongly for a small portion of images compared with deep networks. In other words, shallow networks displayed narrower tails in their Conv1 activation distributions compared with fatter tails in Conv1 of deep networks (Figure 2D). To quantify this difference in selectivity, the kurtosis or "tailedness" of activations was measured in the first stage of processing (Figure 2E). This measure could be considered an "unsupervised" version of selectivity, because it refers to high selectivity for certain images regardless of their class label. Specifically, the kurtosis for each filter was calculated at every location in the image across the CIFAR10 test set, and then the maximum kurtosis was selected as the value that mattered most for downstream decoding. Here, two-layer networks had higher kurtosis than three-layer networks: with one-way analysis of variance (ANOVA), $F(3, 236) = 21.34$, $p < 0.001$; for post hoc two-sample t-tests with Bonferroni correction, $t(118) = 5.1314$, $p < 0.001$; four-layer networks, $t(118) = 5.93$, $p < 0.001$); five-layer networks, $t(118) = 6.34$, $p < 0.001$.

Thus, it was observed that shallow networks had more selective features in their initial layer, generating a distribution with more outliers and thicker tails. In contrast, deep networks had more general features in their first stage of processing, generating a more

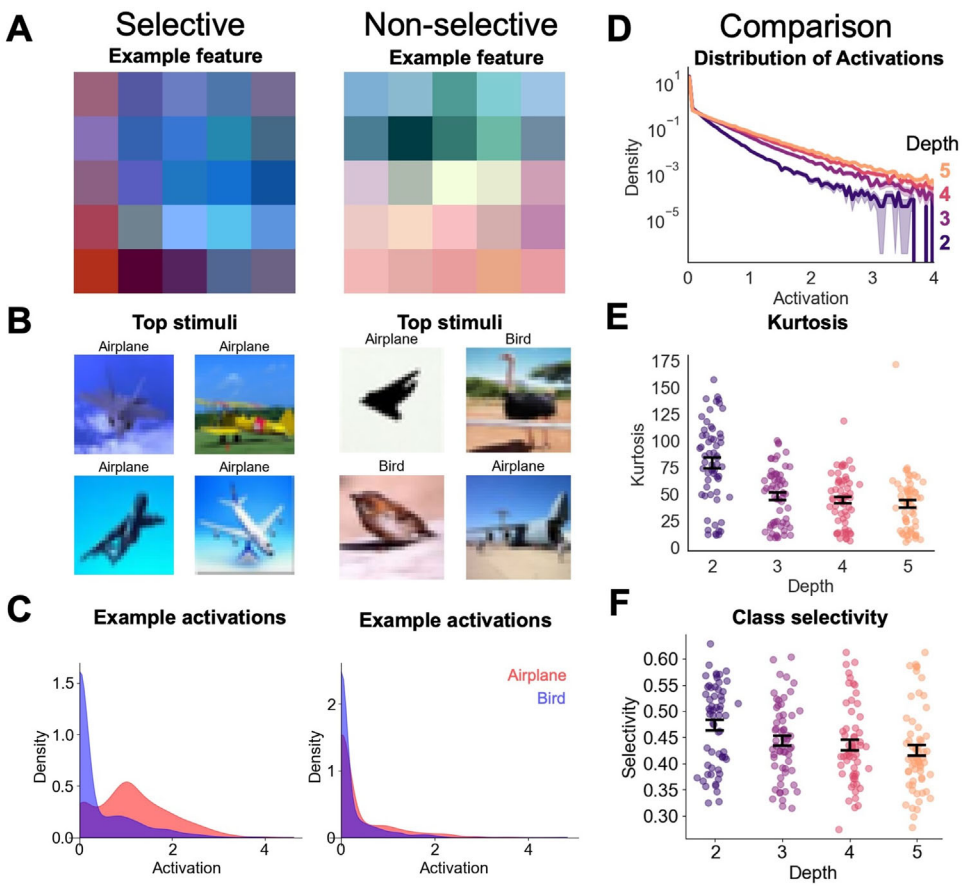

Figure 2. Shallow networks encode specific features, whereas deep networks encode more general features. (**A**) Example selective and non-selective visual features. (**B**) Top four stimuli for example selective and nonselective features across two classes (airplanes and birds). (**C**) Activations for example selective and nonselective features across two classes (airplanes and birds). (**D**) Histograms of activations in Conv1 for networks different depths. Here, activation histograms are shown for each filter at the location with the highest kurtosis. (**E**) Kurtosis or "tailedness" in Conv1 for models of different depths. Here, kurtosis was calculated separately for each filter at the location that maximized the kurtosis measure. Dots represent kurtosis for each of the 60 filters; black error bars show mean ± *SEM*. (**F**) Shallow networks had greater feature selectivity in their first layer than deep networks. Here, selectivity was calculated separately for each filter at the location that maximized the selectivity measure. Dots represent kurtosis for each of the 60 filters; error bars show mean ± *SEM*.

Gaussian distribution. Further analysis suggested that additional filters in Conv1 increased kurtosis but additional classes had a negligible impact (Supplementary Figure S2).

Additionally, a second, "supervised" measure of selectivity was defined based on the selectivity of a feature for its preferred class over the remaining classes. For illustration purposes, two example filters and their activations across only two classes are shown: birds and airplanes. An example selective feature consists of a red square against a blue background, and it activated more strongly for airplane images compared with bird images. In contrast, an example non-selective feature consisted of a green line over a pink line. This feature responded similarly to pictures of airplanes and birds and would require further processing before making a correct classification. Again, the selectivity for each filter was calculated at every location in the image across the CIFAR10 test set, and then the maximum selectivity was selected across all locations as the most informative value for downstream decoding. Using this measure, it was again found that shallow networks encoded more selective features than deep networks in their first stage of processing, in the sense that they more strongly preferred one class of images over the remaining classes. Two-layer networks had higher selectivity than three-layer networks: one-way ANOVA, $F(3, 236) = 4.46$, $p < 0.01$; post hoc two-sample t-tests with Bonferroni correction, $t(118) = 2.19$, $p = 0.091$; four-layer networks, $t(118) = 2.72$, $p < 0.05$; five-layer networks, $t(118) = 3.35$, $p < 0.01$. Additional analysis suggested that more Conv1 filters increased selectivity but more classes decreased it (Supplementary Figure S2).

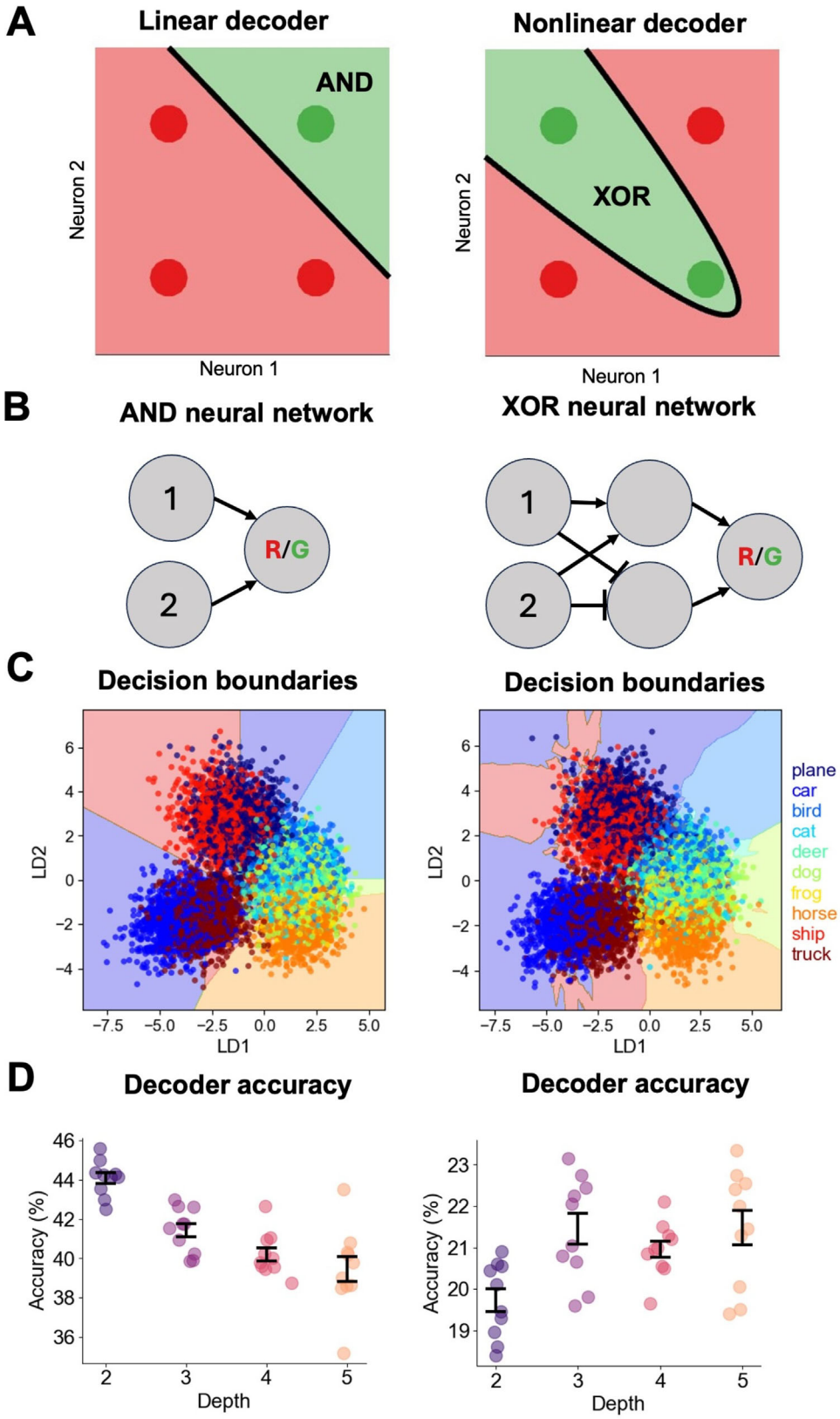

Figure 3. Shallow networks display linear encodings of class identity, but deep networks display nonlinear encodings, requiring further processing. (**A**, top) Linear decision boundary for logical AND. (**A**, bottom) Nonlinear decision boundary for logical XOR. (**B**, top) Minimal neural network required to implement logical AND. (**B**, bottom) Minimal neural network required to implement logical XOR. (**C**, top) LDA visualization of Conv1 activations for the CIFAR10 test set with linear decision boundaries. (**C**, bottom) LDA visualization of Conv1 activations for the CIFAR10 test set with nonlinear decision boundaries. (**D**, top) Linear decoding accuracy for Conv1

→

←
activations in neural networks of different depths. Here, linear decoders were trained on Conv1 activations for 80% of the CIFAR10 test set and then tested on the remaining 20%. Dots represent accuracy for each of the 10 models; error bars show mean ± *SEM*. (**D**, bottom) Nonlinear decoding accuracy for Conv1 activations in neural networks for different depths. Here, nonlinear decoders were trained on Conv1 activations for 80% of the CIFAR10 test set and then tested on the remaining 20%. Dots represent accuracy for each of the 10 models; error bars show the mean ± *SEM*.

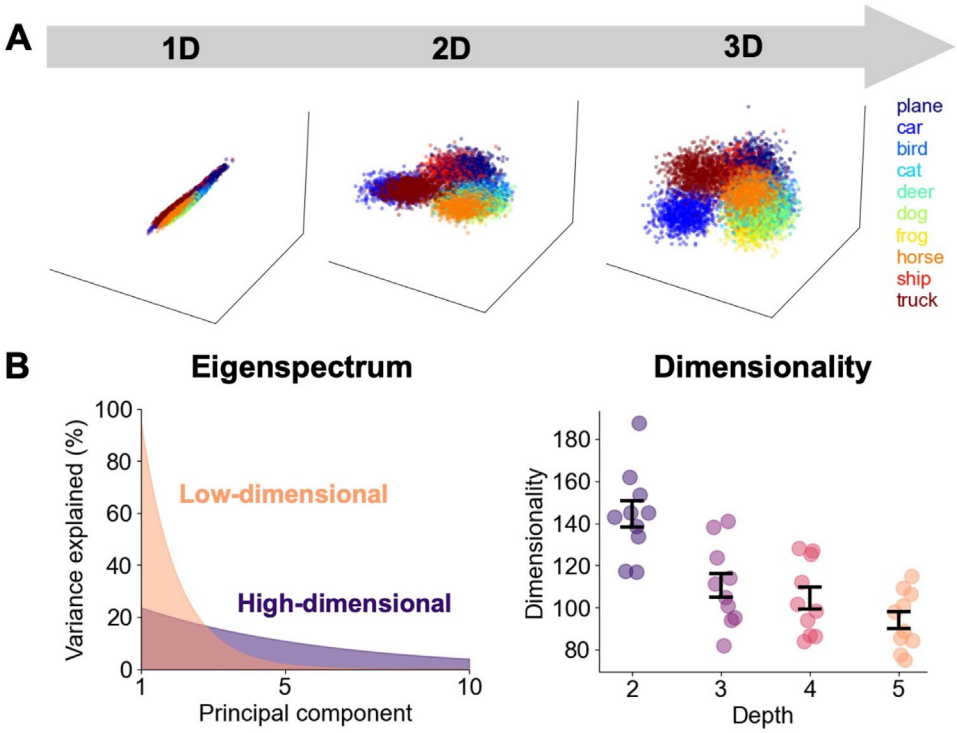

Figure 4. Shallow networks have high-dimensional representations in their first stage of processing, but deep networks have low-dimensional representations. (**A**) Schematic depicting increasing dimensionality of a neural representation. (**B**) Schematic showing the eigenspectra of low-dimensional and high-dimensional neural representations. (**C**) Dimensionality of Conv1 activations on the CIFAR10 test set in models of different depths. Dots show dimensionality for each of the 10 models; error bars show the mean ± *SEM*. Maximum dimensionality would be 28 height × 28 width × 6 filters = 4704 dimensions, if all locations and filters were uncorrelated.

Shallow networks have representations with linear decision boundaries and deep networks have representations with nonlinear decision boundaries

Next, how these changes at the individual neuron level affect encoding at the population level was examined. Within the population code, we can distinguish between information that is directly accessible to downstream neurons and information that is relatively "encrypted" and not directly accessible (Figure 3A). Specifically, linear decoders measure accessible information, whereas nonlinear decoders measure relatively "encrypted" information. (Figure 3B). Based on this idea, accessible information was assessed using a simple linear model to decode class identity from the activations in early processing (Figure 3C). Specifically, this linear decoder was applied to the Conv1 activations in the CIFAR10 test set using a train–test split of 80/20. Shallow networks were found to have better linear representations of class identity (Figure 3D). Specifically, two-layer networks had significantly higher linear decoding accuracy than three-layer networks: one-way ANOVA, $F(3, 36) = 21.10$, $p < 0.001$; post hoc two-sample t-tests with Bonferroni correction, $t(18) = 5.75$, $p < 0.001$; four-layer networks, $t(18) = 8.63$, $p < 0.001$); five-layer networks, $t(18) = 6.38$, $p < 0.001$.

Meanwhile, the opposite trend was found when using a nonlinear classifier to measure information that is relatively "encrypted" or not directly accessible in the population representation (Figure 3C). A k-nearest neighbor classifier where $k = 10$ was applied this to the Conv1 activations on the CIFAR10 test

set using a train–test split of 80/20. Deep networks were found to have more nonlinear representations of class identity compared with shallow networks (Figure 3D). Specifically, two-layer networks had significantly lower nonlinear decoding accuracy than three-layer networks: one-way ANOVA, $F(3, 36) = 5.63$, $p < 0.01$; post hoc two-sample t-tests with Bonferroni correction, $t(18) = -3.54$, $p < 0.01$; four-layer networks, $t(18) = -3.48$, $p < 0.01$; five-layer networks, $t(18) = -3.31$, $p < 0.05$. This observation highlights how neural network depth shapes the representational geometry in the initial processing stage. Visual stimuli naturally have highly nonlinear decision boundaries between object classes. As shown by Lindsey et al. (2019), deep networks maintain more of this information into the first layer of processing. This preserves the nonlinear decision boundaries of the original data but also requires additional processing to generate the appropriate behavioral responses. Meanwhile, shallow networks destroy some of this information in order to generate linear decision boundaries that can directly feed into a behavioral response.

Shallow networks use high-dimensional representations and deep networks use low-dimensional representations

Finally, the mechanism behind pattern separation was examined in these different models (Figure 4A). Prior work has shown that high-dimensional representations help to make different classes more linearly separable (Bernardi et al., 2020; Boyle et al., 2024). Using the participation ratio measure of dimensionality on Conv1 activations in the CIFAR10 test set, shallow networks had high-dimensional representations in their first layer, but deep networks had low-dimensional representations (Figure 4B). Specifically, two-layer networks had significantly higher dimensionality than three-layer networks: one-way ANOVA, $F(3, 36) = 14.54$, $p < 0.001$; post hoc two-sample t-tests with Bonferroni correction, $t(18) = 3.77$, $p < 0.01$; four-layer networks, $t(18) = 4.63$, $p < 0.001$; five-layer networks, $t(18) = 6.37$, $p < 0.001$. This suggests that shallow networks achieve rapid pattern separation using high-dimensional representations in the first stage of processing. Meanwhile, deep networks use low-dimensional representations to perform more gradual pattern separation, which they continue through dimensionality expansion in later stages of processing. In other words, shallow networks perform pattern separation all in one step, whereas deep networks do it through an iterative process of dimensionality expansion and reduction (Supplementary Figure S3).

Discussion

This experiment tested a seemingly paradoxical evolutionary trend: As brains have evolved greater complexity, retinas have specialized in simpler visual features. Although biologists have observed this inverse relationship between brain and retinal complexity (Johnston & Lagnado, 2012), distinguishing causation from correlation remains a significant challenge. Comparisons between species are plagued by numerous confounding factors, including different environments, niches, predators, prey, and reliance on other sensory modalities. To address this, this study used artificial networks to eliminate these external confounds. Networks of different depths were trained on the same dataset with the same architecture, loss function, and optimizer. The analysis indicates that shallow networks encode highly selective features in their early processing stages, enabling immediate behavioral responses. Meanwhile, deep networks encode more general features that require additional stages of processing before they can generate the appropriate behavioral response (Figure 5A). Feature selectivity was found to be strongly influenced by other variables such as the number of first-layer filters, which increase selectivity, and the number of output classes, which decrease selectivity. These variables may also be highly relevant to the differences in selectivity found across species.

Also examined was how these changes at the individual neuron level affect representations at the population level. Specifically, we can distinguish between accessible information in a representation, which can be read out by a linear decoder, and relatively inaccessible information, which requires a nonlinear decoder. Here, the results indicate that shallow networks encode object classes with more linear decision boundaries, facilitating direct decoding by downstream neurons. Meanwhile, deep networks encode object classes with more nonlinear decision boundaries that require additional stages of processing before they can be translated into the appropriate behavioral response (Figure 5B). This observation mirrors the results from the individual neuron analysis in this study but at a population level. In fact, a much stronger effect at the population level compared with the individual neuron level was found, highlighting the importance of looking at population measures.

Finally, the mechanism behind pattern separation in these different networks was analyzed. According to previous research, high-dimensional representations help separate different patterns of inputs (Bernardi et al., 2020; Boyle et al., 2024). This study found that shallow networks had high-dimensional representations in their first stage of processing, indicating rapid pattern separation. Meanwhile, deep networks had lower dimensional representations, leaving additional pattern

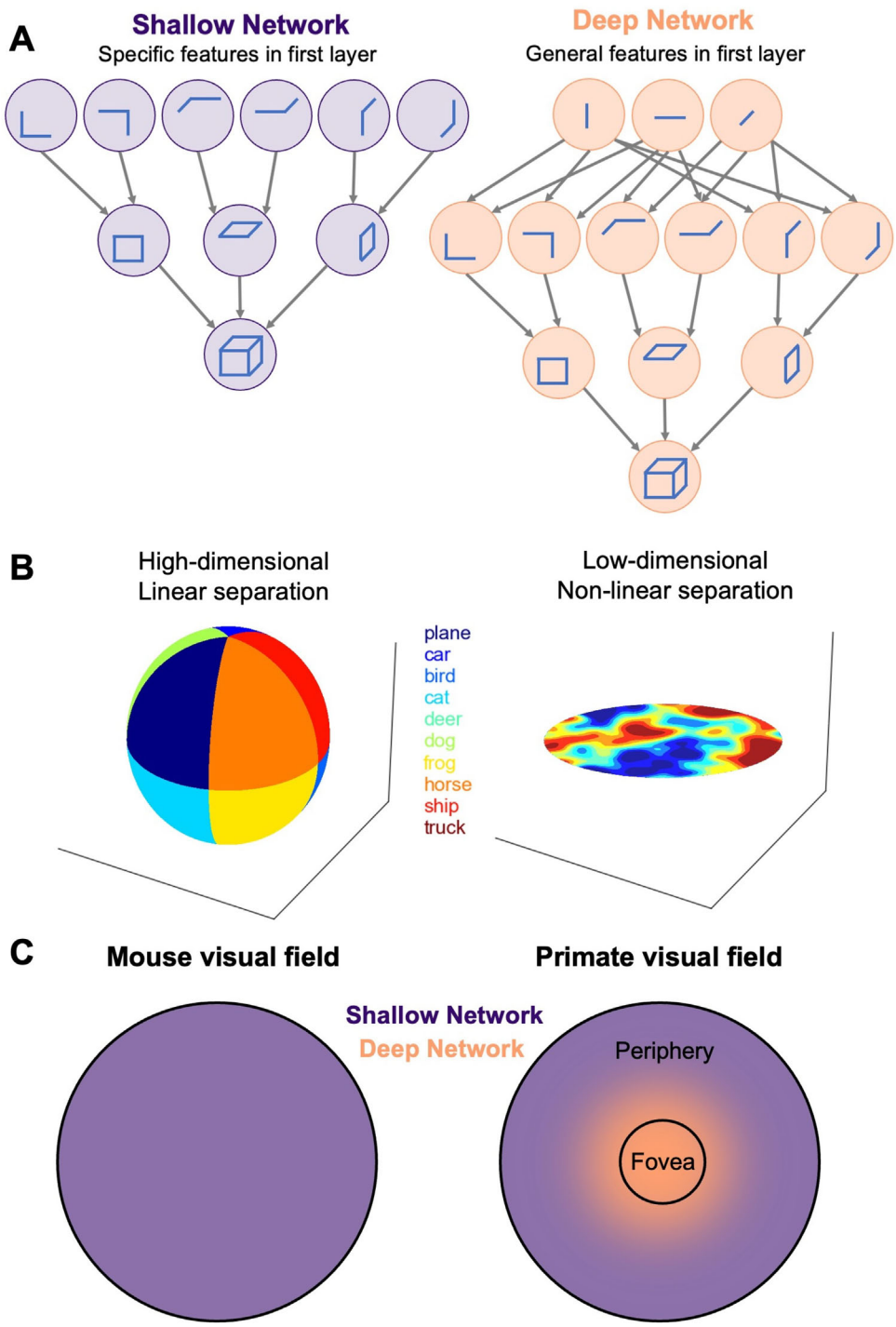

Figure 5. Shallow and deep networks have different selectivity and geometry in their first stage of processing. (**A**) Schematic depicting how shallow networks have specific features in their first layer and deep networks have more general features. (**B**) Schematic showing how shallow networks have high-dimensional and linear encoding of class identity, whereas deep networks have low-dimensional and nonlinear encoding. (**C**) Depth of processing across the visual field in different species: mouse (left) and primate (right).

separation to later stages of processing (Figure 5B). Thus, shallow networks take a faster approach to pattern separation, and deep networks take a more gradual and patient approach. The literature also contains examples where low-dimensional representations have improved linear classification through generalization across different patterns (Bernardi et al., 2020; Boyle et al., 2024). However, this mechanism is usually found in higher brain areas, after the initial pattern separation described in this paper.

Taken together, these findings indicate that shallow networks employ a quick and efficient strategy for image classification, leveraging highly selective features, linear decision boundaries, and high-dimensional representations to rapidly generate behavioral responses. Conversely, deep networks adopt a patient approach, using more general features, nonlinear decision boundaries, and low-dimensional representations that require more extensive processing but ultimately generate superior responses. Thus, this study offers computational validation of a longstanding biological observation that species with complex brains tend to detect simpler features in the retina, and vice versa.

The primate visual system offers an interesting case study because it integrates both deep and shallow networks (Figure 5C). The primate fovea consists of midget ganglion cells that detect relatively simple features such as small spots of light. Consistent with the findings of this study, these simple visual features preferentially connect to a deep network, the ventral stream of visual processing, which specializes in object recognition (Felleman & Van Essen, 1991; Jusuf, Martin, & Grünert, 2006; Kolb, 1995; Wool, Packer, Qasim Zaidi, & Dacey, 2019). Meanwhile, the peripheral retina consists of complex detectors such as direction-selective ganglion cells (Liu et al., 2022). These connect to a relatively shallow network, superior colliculus, which specializes in quick reactions to looming objects (Hafed et al., 2023; Marrocco & Li, 1977; Shi et al., 2017; Tailby, Cheong, Pietersen, Solomon, & Martin, 2012). Interestingly, they also connect to the LGN, which may be related to direction selectivity V1 and the dorsal stream. Thus, within the primate visual system, we find both sides of our principle, with simple visual features feeding into a deep network and complex visual features feeding into a shallow network. In fact, this hybrid network may represent the ideal solution because it combines the benefits of both systems: the detailed processing of a deep network and the fast reaction times of a shallow network. Therefore, a deep network is not simply a better version of a shallow network; in fact, deep networks may only confer a survival benefit when paired with a shallow network that quickly and automatically responds to threats.

Although convolutional neural networks are an established model for biological vision, they also have some important differences from their biological counterparts, including their learning algorithms, (backpropagation vs. Hebbian plasticity) (Hebb, 1949; Rumelhart, Hinton, & Williams, 1986), architectures (entirely feedforward vs. recurrent and feedback) (Sinz, Pitkow, Reimer, Bethge, & Tolias, 2019), and tasks (a single, isolated task vs. many, multi-sensory tasks) (Zador et al., 2023). Future studies could explore how these additional variables influence the findings of this study.

In addition, future work can explore how this principle may apply to other forms of hierarchical processing beyond vision. The auditory cortex also performs hierarchical processing, combining basic features such as sound frequency and intensity into more complex features like patterns of sound (Sharpee, Atencio, & Schreiner, 2011). Language networks combine phonemes into words, which are then combined into sentences and mapped to semantic meaning (Frank & Christiansen, 2018). Thus, the principle about depth and complexity may generalize to these other domains. For example, a deep auditory network might encode simple sounds in its first stage of processing, whereas a shallow auditory network encodes more complex sounds. Likewise, a deep language network may encode simpler structures in its first layer, whereas a shallow language network encodes more complex structures. Thus, these findings may extend beyond vision, offering a more general principle about how neural networks function.

Keywords: evolution, retina, computational modeling

Acknowledgments

Research reported in this publication was supported by the National Eye Institute of the National Institutes of Health under award number 5F30EY035603-02 and the UTHealth Houston Center for Clinical and Translational Sciences under award number TL1 TR003169. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data and code availability: https://github.com/mitchellslapik/retina.

Commercial relationships: none.
Corresponding author: Mitchell B. Slapik.
Email: mitchell.slapik@uth.tmc.edu.
Address: Department of Neurobiology and Anatomy, McGovern Medical School, University of Texas–Houston, Houston, TX 77030, USA.

References

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science, 364*(6439), eaav9436.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell, 183*(4), 954–967.e21.

Boyle, L. M., Posani, L., Irfan, S., Siegelbaum, S. A., & Fusi, S. (2024). Tuned geometries of hippocampal representations meet the computational demands of social memory. *Neuron, 112*(8), 1358–1371.e9.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., . . . Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology, 15*(4), e1006897.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1–47.

Frank, S. L., & Christiansen, M. H. (2018). Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience, 33*(9), 1213–1218.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology, 37*, 66–74.

Gabbiani, F., Krapp, H. G., & Laurent, G. (1999). Computation of object approach by a wide-field, motion-sensitive neuron. *The Journal of Neuroscience, 19*(3), 1122–1141.

Gao, P., Trautmann, E., Yu, B., Santhanam, G., Ryu, S., Shenoy, K., . . . Ganguli, S. (2017). A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv,* https://doi.org/10.1101/214262.

Hafed, Z. M., Hoffmann, K. P., Chen, C. Y., & Bogadhi, A. R. (2023). Visual functions of the primate superior colliculus. *Annual Review of Vision Science, 9*, 361–383.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Johnston, J., & Lagnado, L. (2012). What the fish's eye tells the fish's brain. *Neuron, 76*(2), 257–259.

Jusuf, P. R., Martin, P. R., & Grünert, U. (2006). Random wiring in the midget pathway of primate retina. *The Journal of Neuroscience, 26*(15), 3908–3917.

Kolb, H. (1995). Midget pathways of the primate retina underlie resolution and red green color opponency. In: H. Kolb, E. Fernandez, B. Jones & R. Nelson (Eds.), *Webvision: The organization of the retina and visual system* [Internet]. Salt Lake City, UT: University of Utah Health Science Center.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Retrieved from https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., . . . Jackel, L. D. (1989). Backpropagation applied to handwritten Zip Code recognition. *Neural Computation, 1*(4), 541–551.

Lettvin, J. Y., Maturana, H. R., Maturana, H. R., Mcculloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE, 47*(11), 1940–1951.

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. Retrieved from https://ganguli-gang.stanford.edu/pdf/19.retinal.resource.pdf.

Liu, X., Huang, H., Snutch, T. P., Cao, P., Wang, L., & Wang, F. (2022). The superior colliculus: cell types, connectivity, and behavior. *Neuroscience Bulletin, 38*(12), 1519–1540.

Marrocco, R. T., & Li, R. H. (1977). Monkey superior colliculus: properties of single cells and their afferent inputs. *Journal of Neurophysiology, 40*(4), 844–860.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron, 108*(3), 413–423.

Sharpee, T. O., Atencio, C. A., & Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex. *Current Opinion in Neurobiology, 21*(5), 761–767.

Shi, X., Barchini, J., Ledesma, H. A., Koren, D., Jin, Y., Liu, X., . . . Cang, J. (2017). Retinal origin of direction selectivity in the superior colliculus. *Nature Neuroscience, 20*(4), 550–558.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron, 103*(6), 967–979.

Tailby, C., Cheong, S. K., Pietersen, A. N., Solomon, S. G., & Martin, P. R. (2012). Colour and pattern selectivity of receptive fields in superior colliculus of marmoset monkeys. *Journal of Physiology, 590*(16), 4061–4077.

Wool, L. E., Packer, O. S., Qasim Zaidi, X., & Dacey, D. M. (2019). Connectomic identification and three-dimensional color tuning of S-OFF midget ganglion cells in the primate retina. *The Journal of Neuroscience, 39*(40), 7893–7909.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, USA, 111*(23), 8619–8624.

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., … Tsao, D. (2023). Catalyzing next-generation artificial intelligence through NeuroAI. *Nature Communications, 14*(1), 1597.

Zhang, Y., Kim, I. J., Sanes, J. R., & Meister, M. (2012). The most numerous ganglion cell type of the mouse retina is a selective feature detector. *Proceedings of the National Academy of Sciences, USA, 109*(36), E2391–E2398.