JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 4, 2022

Pp. 370–381
DOI: 10.1089/cmb.2021.0368

Emergence of Direction-Selective Retinal Cell Types in Task-Optimized Deep Learning Models

KEITH T. MURRAY,[1,i] MIEN BRABEEBA WANG,[2] and NANCY LYNCH[2]

ABSTRACT

Convolutional neural networks (CNNs), a class of deep learning models, have experienced recent success in modeling sensory cortices and retinal circuits through optimizing performance on machine learning tasks, otherwise known as task optimization. Previous research has shown task-optimized CNNs to be capable of providing explanations as to why the retina efficiently encodes natural stimuli and how certain retinal cell types are involved in efficient encoding. In our work, we sought to use task-optimized CNNs as a means of explaining computational mechanisms responsible for motion-selective retinal circuits. We designed a biologically constrained CNN and optimized its performance on a motion-classification task. We drew inspiration from psychophysics, deep learning, and systems neuroscience literature to develop a toolbox of methods to reverse engineer the computational mechanisms learned in our model. Through reverse engineering our model, we proposed a computational mechanism in which direction-selective ganglion cells and starburst amacrine cells, both experimentally observed retinal cell types, emerge in our model to discriminate among moving stimuli. This emergence suggests that direction-selective circuits in the retina are ecologically designed to robustly discriminate among moving stimuli. Our results and methods also provide a framework for how to build more interpretable deep learning models and how to understand them.

Keywords: biological constraints, convolutional neural network, direction-selectivity and interpretable deep learning, task optimization.

1. INTRODUCTION

THE RETINA SERVES AS THE FIRST STEP in visual processing for the brain in nearly all animals (Baden et al., 2020). While it may serve only as a first step, the retina processes visual information using complex neural circuits that are yet to be fully understood or modeled (Gollisch and Meister, 2010). As an

Departments of [1]Brain and Cognitive Sciences and [2]Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
[i]ORCID ID (https://orcid.org/0000-0002-5509-9744).

understanding of the function of these complex retinal circuits has evolved, so too have the models used to explain them. Linear/nonlinear Poisson cascade models of the retina that can reproduce retinal responses to white noise (Chichilnisky, 2001) have now been replaced by deep convolutional neural network (CNN) models that are able to reproduce retinal responses to natural sight (McIntosh et al., 2016; Maheswaranathan et al., 2019), extract previously unknown computational mechanisms (Tanaka et al., 2019), and investigate the theoretical reasons behind efficient encoding (Ocko et al., 2018; Lindsey et al., 2019).

Previous work modeling visual cortices (Yamins et al., 2014) with CNNs optimized for performance on computer vision tasks, otherwise referred to as task optimization, initially inspired the use of CNNs to model retinal circuits (McIntosh et al., 2016). Task-optimized CNNs have been shown to also model the auditory cortex (Kell et al., 2018), produce synthetic images that maximally drive activity in the visual cortex (Bashivan et al., 2019), and explain the constraints that shape neuronal organization (Yamins and DiCarlo, 2016; Kell and McDermott, 2019).

For retinal circuits, task-optimized CNNs have been previously used to show that certain retinal cell types critically contribute to efficient encodings of natural stimuli (Ocko et al., 2018) and that retinal size informs how retinal circuits should balance feature extraction and efficient encoding (Lindsey et al., 2019). In summation, the previous works on task-optimized CNNs have shown them to be capable of answering a variety of theoretical and computational neuroscientific questions.

To expand on the existing work of using task-optimized CNNs to model retinal circuits, we sought to use task-optimized CNNs as a means of explaining computational mechanisms responsible for motion-selective retinal circuits. Motion-selective retinal circuits have been well studied experimentally (Barlow and Levick, 1965; Takemura et al., 2013; Kim et al., 2014) and theoretically (Baccus et al., 2008; Gollisch and Meister, 2010), making these circuits ripe for an explanation from task optimization. By creating a biologically constrained CNN to optimally perform motion-related tasks, we sought to provide more theoretical evidence as to how motion-selective circuits in the retina organize and process moving stimuli.

Our CNN model contained biologically constrained connectivity, in line with Dale's principle (Eccles et al., 1954), between convolution layers as a means of eliciting biologically plausible computational mechanisms (Song et al., 2016). We trained the model on a simple binary motion-classification task via gradient descent to invoke motion-selective computational mechanisms in our model.

It has been well documented that deep learning models are difficult to interpret (Yosinski et al., 2015) and some may argue that this lack of interpretability prevents deep learning models from being useful models of neural circuits. We sought to overcome this difficulty through biologically constraining our model and analyzing it via a toolbox of analysis methods inspired by methods in psychophysics, deep learning, and systems neuroscience. We designed five separate computational methods that enabled us to infer and reverse engineer the learned computational mechanisms in our model. These techniques are most effective on biologically constrained deep learning models and highlight the effectiveness of including these constraints.

We found that motion-selective retinal cells emerged from our task-optimized CNN. In particular, direction-selective ganglion cells (DSGCs) and starburst amacrine cells (SACs), both experimentally observed retinal cells (Wei, 2018), were functionally realized by layers in our CNN model. Given the nature of our motion-classification task, our results indicate that DSGCs, SACs, and other motion-selective cell types are realized in retinal circuits to optimally perform motion discrimination. Furthermore, the interpretability of our trained model suggests that biological constraints could drive deep learning models toward interpretability.

2. METHODS

Our approach can be partitioned into three separate steps: **Task Design**, **Model Design**, and **Analysis Methods**. For **Task Design**, we drew inspiration from previous motion-perception, systems neuroscience experiments (Newsome and Pare, 1988) to design a simple binary motion-classification task. For **Model Design**, we used design principles from Richards et al. (2019) to design a biologically constrained CNN model. For **Analysis Methods**, we incorporated a mix of psychophysics (Klein, 2001), deep learning (He et al., 2016; Meyes et al., 2019), and systems neuroscience (Baccus et al., 2008; Bashivan et al., 2019) methods to create a toolbox of methods we could draw from to computationally understand our model.

2.1. Task design

To design our task, we drew inspiration from Newsome and Pare (1988), in which the authors investigated the neural coding of the middle temporal (MT) visual area . The MT area is thought to be responsible for motion perception and the authors elicited MT activity in monkeys through engagement in a motion-discrimination task. Their task involved a population of dots moving coherently against other noncoherently moving populations of dots and required monkeys to select the direction of the coherently moving population of dots. For our task, we borrowed the authors' concept of eliciting motion-perception activity in a neural circuit via a motion-discrimination, decision-making task.

Our task consisted of dots moving either left or right (Fig. 1A) and to solve the task, our model would decide which direction of moving dots was fastest. Each population of dots moved with some constant speed and the ratio of speeds between populations of dots was determined by a variable α. We hypothesized that for any CNN to achieve a high level of performance on our task, some computational mechanism would have to be learned that could robustly discriminate between moving stimuli of different speeds.

The data set consisted of 1000 stimuli with each stimulus being an instance of the task. Each stimulus was a 255 by 255 dimensional video consisting of 51 frames in which the two populations of dots continuously moved. For each stimulus, the placement of the dots was randomly initialized with an average of 16.67 dots being in one frame at any given frame. Each dot in a stimulus was assigned a direction, right or left, with uniform probability. For each stimulus, the speed for the right- and left-directional was determined via the following method:

1. The slow direction (S) is chosen uniformly between left and right.
2. S is assigned a speed by $s \sim U(0.5,\ 1.5)$.
3. The fast direction (F) is assigned a speed by $f = \alpha * s$, where α is the speed ratio variable.

The environment of the data set was chosen to have an α of 2 for training and testing the model. The choice of 2 for α was a critical detail because it created overlaping s and f speed distributions (Fig. 1B) that forced our model to learn nontrivial, biologically plausible computations through task optimization. The s distribution ultimately determined the difficulty of any given stimulus. An S near the lower end of s is more difficult because the speed difference between S and F is smaller and less visually perceptible (Fig. 1C).

2.2. Model design

Determining the design of a task-optimized deep learning model to be used in modeling a neural circuit is not yet an established science; however, there have been attempts to standardize a list of criteria. Richards et al. (2019) proposed that the three guiding criteria should be (1) *architecture*, the layers of a

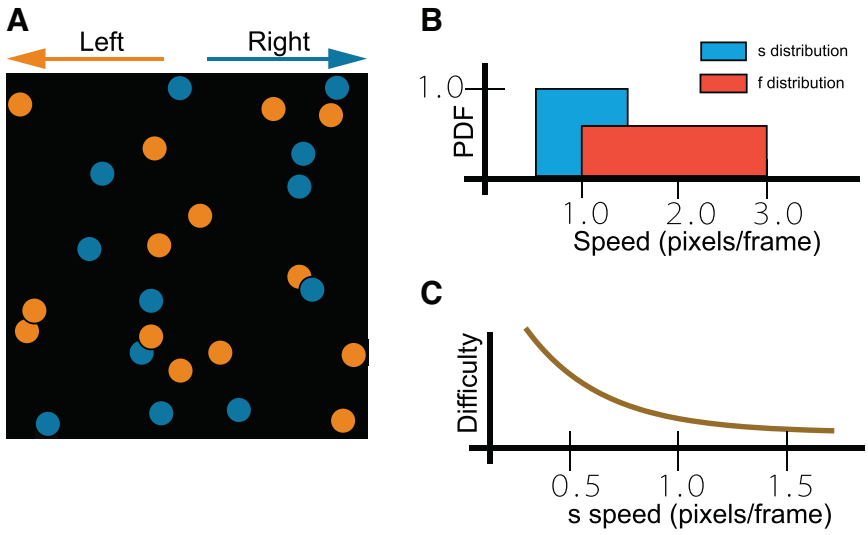

FIG. 1. Illustrations of task design. **(A)** An example stimulus frame. Note, there are no colors in the actual stimulus. **(B)** The s and f distributions and the overlap between them. This overlap causes substantial difficultly in the task. **(C)** The estimated difficulty of the task decreases as the S speed increases and creates a larger separation from F.

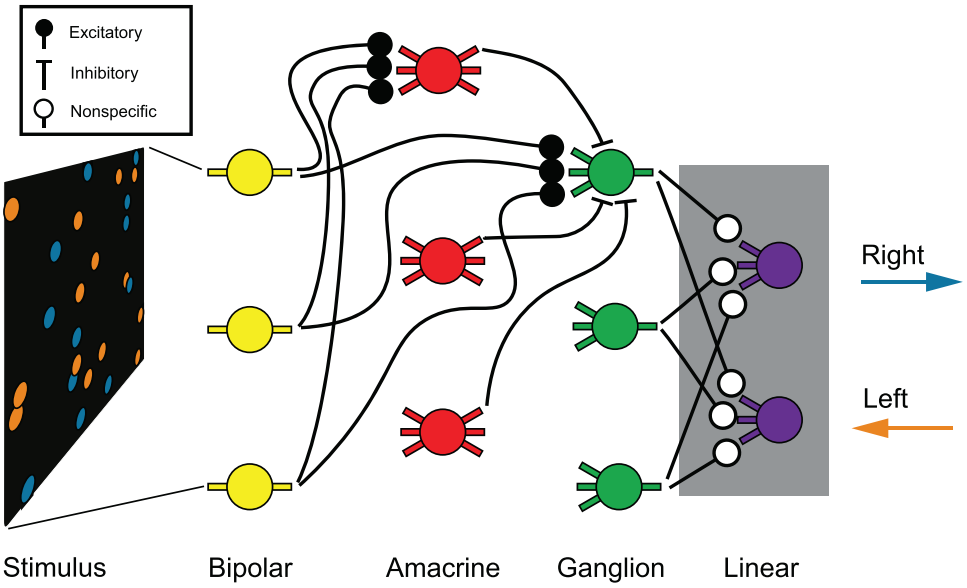

FIG. 2. Diagram of the convolutional neural network and its connections. Only three cell types per layer are depicted here, but eight per layer were used. Each cell type consists of a spatial convolution denoted by the connections, an internal temporal convolution, and a rectified linear unit activation function. This diagram highlights the *all-to-all* connectivity between the cell types in the bipolar/amacrine, amacrine/ganglion, and bipolar/ganglion layers. It is also important to note that there the linear layer is highlighted in *gray* because it serves as part of the objective function, rather than part of our model of the retina. Each neuron in the linear layer corresponds to a *direction* in the task.

model and their associated connections are organized, (2) *objective function*, the mathematical function that compares the computed output of the model with the desired output, and (3) *learning rule*, the algorithm used to train the model.

2.2.1. Architecture. The *architecture* of our CNN model consisted of three convolutional layers, with each layer being analogous to a class of neurons in the retina. The first layer of our model represented bipolar cells, the second layer represented amacrine cells, and the third layer represented ganglion cells (Fig. 2). The connections between each layer were also constrained to resemble the structure of the retina. There were two types of connections: (1) weight-constrained connections between consecutive layers and (2) weight-constrained *residual connections** between the bipolar and ganglion layers (Fig. 2). While *residual* connections have been shown to increase the performance of CNNs in object-recognition tasks (He et al., 2016), our motivation for inserting a weight-constrained residual connection in our model came from an experimentally observed connection between the bipolar and ganglion layers in the retina (Barlow and Levick, 1965; Baccus et al., 2008).

By including this residual connection, input/output constraints between cell types could be enforced that previously did not exist in other CNN models of the retina (McIntosh et al., 2016; Ocko et al., 2018) but have been experimentally observed (Gollisch and Meister, 2010). It is these input/output constraints that ensured that the different layers in our model would learn different computations and mimic the different computations experimentally observed in the retina. Input/output constraints were realized through weight constraints in the model and followed the wiring scheme displayed in Figure 2. Through including and enforcing connectivity constraints, Dale's principle, which states that neurons only use one type of neurotransmitter (Eccles et al., 1954), could be realized in our model to elicit biologically plausible computations through task optimization (Song et al., 2016).

In addition to the structure of the retina containing bipolar, amacrine, and ganglion cells, it has also been observed that there exist multiple different types of bipolar, amacrine, and ganglion cells (Gollisch and

**Residual connections* differ from other connections because they are between layers. that are not consecutive (He et al. 2016).

Meister, 2010). For example, there are many types of bipolar cells, each responding to different profiles of light. We sought to include this variety of *cell types* in our model by having eight copies of each layer. Each copy[†] of one layer received the same inputs as every other cell type in the layer but would have independent weights. This independence of weights allows for a diversity of cell types to emerge through training. Each cell type in our model was assigned a number 0–7 that allowed for identification and analysis in a structured manner. We assumed an *all-to-all* connectivity between the cell types in the bipolar/amacrine connection, amacrine/ganglion connection, and bipolar/ganglion connection (Fig. 2).

Each cell type consisted of a spatial convolution, a temporal convolution, and a rectified linear unit activation function. Spatial convolutions are analogous to synapses between consecutive layers and are analogous to photoreceptor bipolar synapses in the bipolar layer. Temporal convolutions are analogous to feedback introduced in cell types via feedback connections to other cell types. In this cell-type framework, the residual connection took shape in the form of an additional spatial convolution between the bipolar and ganglion layers. Weight constraints in the amacrine and ganglion layers were enforced by constraining the learned spatial and temporal convolutional filters to be strictly positive. After every weight update during training, negative weights in these filters would be clipped to 0. The inhibitory output of amacrine cell types (Fig. 2) was realized in our model by multiplying the output of each amacrine cell type by -1.[‡]

The *architecture* of our CNN model can be summarized as having three distinct layers, each consisting of eight different cell types. Layer 1 represented bipolar cells and receives pixel information from the task. There are no weight constraints in the bipolar layer. Layer 2 represented amacrine cells and receives strictly excitatory activity from the bipolar layer. There are non-negative weight constraints on all convolutional kernels in this layer and the output is nonpositive.

Layer 3 represented ganglion cells and receives excitatory activity from the bipolar layer and negative activity from the amacrine layer. There are non-negative weight constraints on all convolutional kernels in this layer and the outputs are non-negative. These specifications and constraints allow for the model to learn 24 different cell types that contribute to achieving a high performance on our task with a guarantee that the learned computations will be biologically plausible.

2.2.2. Objective function. The *objective function* of a deep learning model is a differentiable function that numerically describes the performance error of some tasks and is crucial for defining how the weights of a deep learning model should be updated to decrease performance error (Richards et al., 2019). Our model's *objective function* consisted of a two-neuron linear layer, with each neuron representing a *left* or *right* direction response, connected to the ganglion layer (Fig. 2) and a cross-entropy loss function that compared the output of the linear layer with the expected output during task performance. Each stimulus consisted of 51 frames and created 51 corresponding outputs in the linear layer, but the cross-entropy loss function only evaluated the output of the linear layer on the last frame of the stimulus. This loss value was then used to begin updating the weights of the model through the *learning rule*.

2.2.3. Learning rule. The *learning rule* for our model was simply the Adam optimization algorithm (Kingma and Ba, 2014), an adjustable learning rate gradient descent algorithm that propagated error derivatives throughout each layer. An important note is that the non-negative weight constraints in our model introduced many local minima in our model's optimization space. To counteract these local minima, we implemented a dropout rate of 40% on all weights.[§] Training took place over 500 epochs of 40 batches per epoch and 25 training stimuli per batch.

2.3. Analysis methods

To develop our toolbox of analysis methods, we drew inspiration from the psychophysics, deep learning, and systems neuroscience literature. We applied these analysis tools to our trained model and its cell types to *reverse engineer* its learned computational mechanisms and compare those mechanisms with experimentally observed retinal mechanisms. The analysis methods we created were a deep learning

[†]From here on referred to as cell type.
[‡]The output would otherwise be positive because of the non-negative weight constraints on the filters.
[§]The inclusion of this high dropout rate is critical for the model to achieve performance on the task above chance.

implementation of **psychometric functions**, a cell-type-specific **ablation study**, a visualization of cell-specific spatiotemporal filters via **deep dreams**, a task performance description via **modified stimuli**, and a **connectivity visualization** to indicate motifs in the model's learn structure.

2.3.1. Psychometric functions. Psychometric functions are a class of models in psychophysics that predict the relationship between a single environmental parameter and the behavioral response by some test subject in that environment (Klein, 2001). The usual shape of a psychometric function is a sigmoidal curve because the two extremes of an environment will elicit gradually attenuating responses that follow the shape of a sigmoid function. For our task, the numerical parameter of the environment is α, the test subject is our model, and the subject's response is our model's accuracy on the task at a particular α. Creating a psychometric function for our trained model will indicate our model's ability to generalize outside its training environment.

2.3.2. Ablation studies. Artificial neural networks (ANNs) are often considered a black box because of the many parameters involved and a lack of understanding as to how those parameters contribute to task performance. Ablation studies provide a window into ANNs by fixing a parameter's value to 0 and evaluating the effect on task performance (Meyes et al., 2019). However, given that ANN models can have millions of parameters, the ablation of one parameter typically does not affect a model's task performance. Ablation studies can implement a variety of techniques to intelligently ablate a group of parameters and test how this group influences task performance (Meyes et al., 2019).

In our model, we can simply ablate all the parameters in one cell type to understand that cell type's effect on task performance. By structuring our model into cell types, our model not only gains biological plausibility but also becomes easier to understand through ablation studies. Our specific implementation of ablation involved fixing all parameters in one cell type to zero. This allows us to draw strong conclusions about the functional role of each cell type instead of some randomly related set of parameters.

To decrease the number of overall parameters of our model, we can ablate multiple cell types at once to drastically reduce parameter complexity. We used combinatorial techniques to explore the space of possible ablations to reduce our model to its simplest yet accurate form. While there are 24 cell types in total, the *architecture* of the model necessitated that there must be at least one nonablated cell type in each layer of the model to allow for information to propagate to the linear layer. This consequence decreased the search space considerably.

2.3.3. Deep dreams. A unique method for understanding the function of certain neurons in an ANN is to utilize a method called deep dreaming.[**] This method visualizes the spatiotemporal receptive field of an ANN neuron by freezing the ANN's parameters and instead parameterizing the pixels of some input so that gradient descent methods can transform the input into an input that maximally activates the neuron (Yosinski et al., 2015). This method has been applied to ANN models of the visual cortex (Bashivan et al., 2019), but involved optimizing a single image to uncover only the spatial filter of some neuron. Our implementation of deep dreaming involved parameterizing a white-noise video and optimizing that video to uncover the entire spatiotemporal filter of a specific cell type.

2.3.4. Modified stimuli. Given that our stimuli of moving dots are laden with parameters dictating distributions of speed, we can readily change these parameters to understand how aspects of our stimuli drive the computational mechanisms in our model. One example is that we created a data set of stimuli that contained an F population of dots moving right and no S population. Each dot in any stimulus was assigned an F speed drawn uniformly from $U(0.5,\ 3.0)$ and a randomly initialized location. Such a data set allowed us to understand how bipolar cells respond to rightward moving dots.

2.3.5. Connectivity visualizations. The connections between various neurons in any neural circuit influence the computations in that circuit. In systems neuroscience, data sets that map these connections are becoming more valuable as scientists use these connections to inform their models (Oh et al., 2014).

[**]Deep dreaming is sometimes referred to as deep visualization.

We can analyze and visualize the learned spatial convolutional kernels in our model to suggest computational mechanisms. Our analysis of learned spatial convolutional kernels took shape in the form of averaging kernels across cell types to identify connectivity motifs.

2.4. Code availability

All methods, models, figures, and code in this article are available at (https://github.com/ktmurray1999/emergence-retina-cell-types).

3. RESULTS

We found that by using our analysis methods, we were able to reverse engineer the computational mechanisms learned in our model. We found that DSGCs and SACs, both experimentally observed direction-selective retinal cell types (Wei, 2018), were present in our model. Furthermore, our results are evidence that biological constraints can increase the interpretability of deep learning models via our proposed analysis methods.

3.1. Full-model analysis

We created a **psychometric function** for our trained model that showed how it can generalize to environments outside its own training environment (Fig. 3A). This generalization suggests that the computational mechanisms in our model may be biologically plausible given the retina's own experimentally observed ability to adapt to new environments (Ozuysal and Baccus, 2012). An **ablation study** was performed on each of the 24 cell types in the model and it was revealed that few cell types are crucial for task performance (Fig. 3B). An ablation of ganglion cell types 0 and 2 had the most drastic impact on task performance, suggesting that those cell types may play an important computational role.

A combinatorial ablation study was performed on the full nonablated model to find the highest performing configuration of 4 cell types. The identified cell types were bipolar cell type 4, amacrine cell type 2, and ganglion cell types 0 and 2. This *maximally ablated model* achieved an accuracy of 81.00% on the test set and had a psychometric function similar to that of the full model (Fig. 3A). For the rest of the analysis, we utilize the simplicity of the *maximally ablated model*, referred as the ablated model, to build an understanding as to how the full model operates.

3.2. Ganglion cell-type analysis

To understand the spatiotemporal filters and the functional outputs of the ganglion cells in our ablated model, we ran the **Deep Dreams** procedure on both ganglion cell types 0 and 2. For ganglion cell type 0, the deep dream movie revealed that activity in the rightward direction persisted as the movie evolved (Fig. 4A). The deep dream video was then fed back into the ablated model and the linear layer showed strong activation in the *right cell* indicating that the model perceived the deep dream to be a rightward

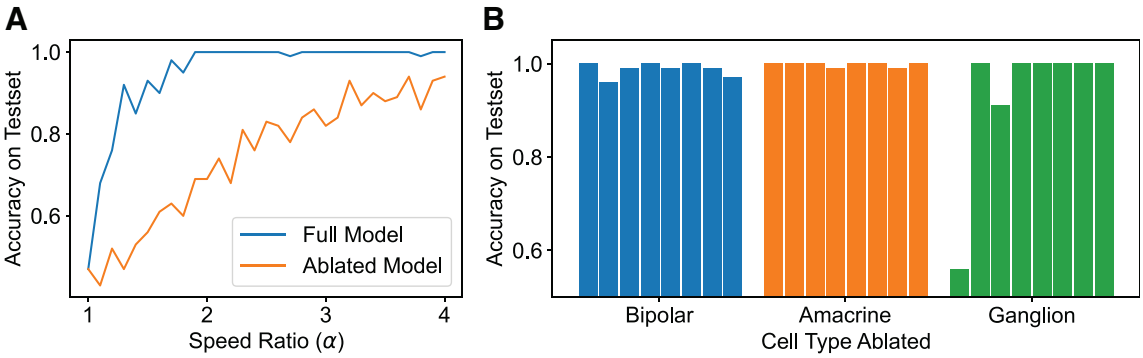

FIG. 3. Full- and ablated model analysis. **(A)** Psychometric curves for the full and ablated model. The full-model curve shows that the trained model demonstrates robust performance and the ablated model curve shows that even with four cell types, much of the robust performance is maintained. **(B)** Ablation study for each cell type in the full model. Most cell types in the full model are not critical to performance except for ganglion cell types 0 and 2.

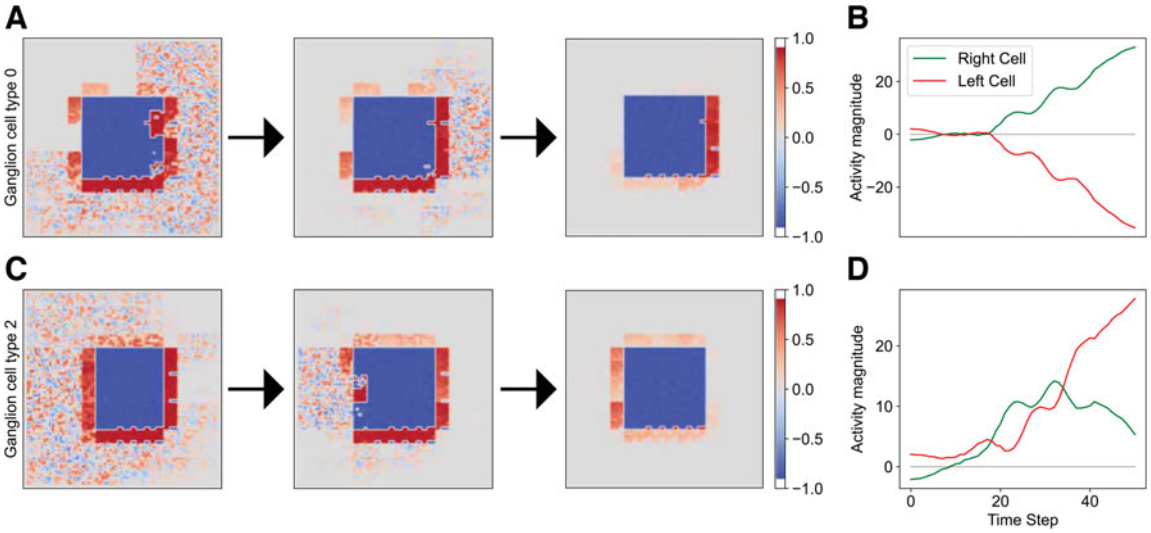

FIG. 4. Deep dreams for both ganglion cell types 0 and 2. **(A)** Deep dream video for ganglion cell type 0 in the ablated model. Activity on the right side of the video lingers. **(B)** Linear output of deep dream video for ganglion cell type 0 in the ablated model. The deep dream video is fed into the model and the right and left linear decoder neuron's outputs are shown here. Ganglion cell type 0 is proven to be direction selective for the right direction because the activation of the right cell in the linear decoding layer is strongest. **(C)** Deep dream video for ganglion cell type 2. Activity on the left side of the video lingers. **(D)** Linear output of deep dream video for ganglion cell type 2. Ganglion cell type 2 is proven to be direction selective for the left direction because the activation of the left cell in the linear decoding layer is strongest.

moving stimulus (Fig. 4B). This indicates that ganglion cell type 0 was direction selective toward the right direction. The same procedure was applied for ganglion cell type 2 and it was revealed that ganglion cell type 2 was direction selective toward the left direction (Fig. 4C, D).

The emergence of direction selectivity in ganglion cell types 0 and 2 is reminiscent of DSGCs experimentally found in the retina (Wei, 2018). This convergence to biology in our model indicates that our CNN architecture and motion-classification task are sufficient to purpose theoretical reasons as to why the retina has motion-computing circuits. However, it is still unclear whether other computational mechanisms in our model are also biologically grounded. An understanding of how the DSGCs computationally emerge will require understanding how bipolar and amacrine cell types function in our model.

3.3. Bipolar cell-type analysis

We utilized the **Modified Stimuli** method to probe the functions of bipolar cells. The modified stimulus created was similar to the task stimulus except for an omitted *S* population. This modified stimulus had only a population of dots right with a modified *f* distribution (Fig. 5A). This modified stimulus could allow for the probability of activation to be ascertained from the bipolar cell types. In general, the probability of activation for a bipolar cell type was 0 in low ranges of *f* but rose to 1 after some speed threshold (Fig. 5B). Each bipolar cell type has a different speed filter that indicated roughly how fast the stimuli. This speed threshold phenomenon allows for our model to partition the distribution of speeds, a useful ability given that the distributions of *s* and *f* overlap in our task (Fig. 1B).

While this speed threshold phenomenon allows for the model to discriminate between overlapping distributions of speed, this behavior in bipolar cell types has, to the best of our knowledge, not been experimentally observed. The bipolar cell types in our model act as high-pass speed filters; we would expect biological bipolar cell types to act as band-pass speed filters to more precisely indicate stimuli speed.

3.4. Amacrine cell-type analysis

To understand what function the amacrine cells perform in our model, we examined our model and various other ablation performances on specific *F* directions in a test set of stimuli with an α of 2. For the full and maximally ablated models, performance was reasonable among the left and right directions (Fig. 6A). When ablating either the bipolar cell or amacrine cell in the already maximally ablated model,

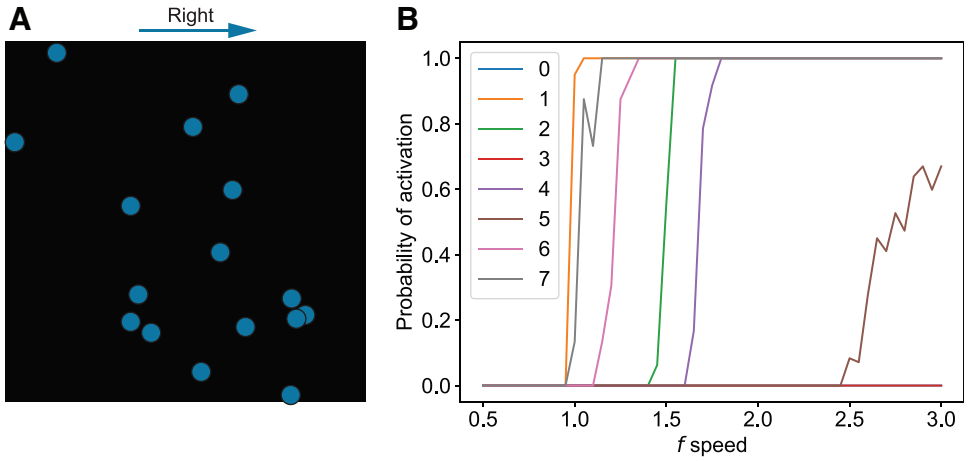

FIG. 5. Modified stimuli to uncover bipolar function. **(A)** Diagram of the modified stimuli. All dots move in the right direction. **(B)** Probability of a bipolar cell activating as stimuli speed increases. The bipolar cells each have a unique speed threshold.

the resulting accuracies indicate that there exists some variability in the *s* distribution of accuracies (Fig. 6B). We created **Modified Stimuli** that included both *F* and *S* populations of dots but instituted a fixed *s* distribution where the speeds of stimuli were evenly distributed along *s*. As *s* increases, the ablated model with further amacrine ablation reveals that there is a failure to properly detect left *F* populations of dots (Fig. 6B). This result indicates that the amacrine cell is direction selective.

SACs are a commonly observed direction-selective amacrine cell that plays a crucial role in the function of DSGCs (Wei, 2018). A hallmark of SACs is that the direction of their dendrites indicates which direction of moving stimuli causes them to be inhibitory. We utilized our **connectivity visualization** method on the connections between all amacrine cell types and the five bipolar cell types with the most prominent speed threshold (Fig. 5B) as a means of visualizing the directions of amacrine dendrites. The resulting visualization indicated that the right side of our model's amacrine dendritic tree was responsible for inhibitory activity (Fig. 6C).

We can infer that when dots move in the left direction (from right to left) across the amacrine cell's receptive field, an inhibitory signal is sent to the right DSGC (ganglion cell type 0). If the amacrine cell in our model were to be ablated, then the right DSGC (ganglion cell type 0) would not be able to be properly inhibited and would dominate the left DSGC (ganglion cell type 2). This exact behavior is observed in Figure 6B and observed experimentally (Wei, 2018). Thus, the amacrine cell types in our model exhibit morphology and behavior like SACs, further proving that our model is biologically realistic.

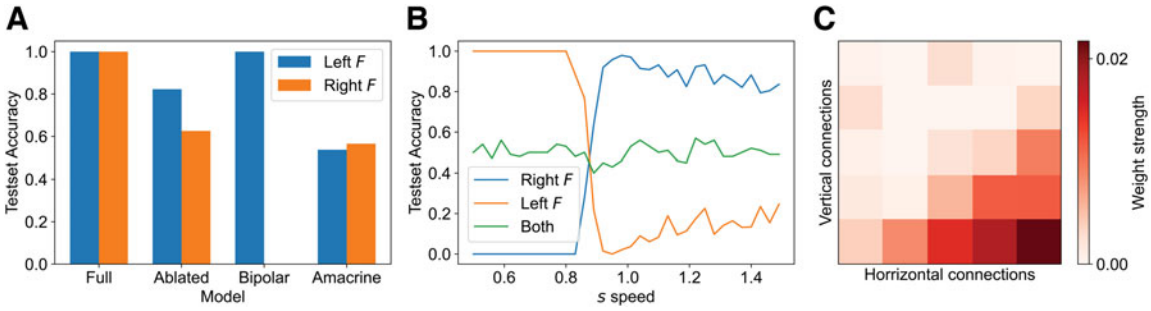

FIG. 6. Connectivity visualization to uncover amacrine function. **(A)** Comparison of left and right *F* accuracies for various models. The *Bipolar* and *Amacrine* models in the plot are both instances of the *Ablated* model, but with the bipolar or amacrine cell type ablated, respectively. Ablating the amacrine cell type reduces performance to chance. **(B)** Left and right *F* accuracies as the *S* speed increases for the ablated model with ablated amacrine cell. Total accuracy is always roughly 50%, but the underlying accuracies show that ablating the amacrine cell causes the model to lose left *F* accuracy. **(C)** Connectivity visualization for all amacrine to bipolar connections. Note, this is not all the bipolar cells. Those bipolar cells that did not have a speed threshold were excluded. Weights were restricted to a value of 0 and up, after the bias was factored in. A rightward bias is found in the connections.

4. DISCUSSION

Through task optimization on a simple motion-categorization task, our biologically constrained CNN emerged to have experimentally observed direction-selective retinal cell types, DSGCs and SACs. Through utilizing our analysis methods, we were able to reverse engineer our model's computational mechanisms. We conceived of the following computational process:

1. Bipolar cells selectively respond to moving stimulus of a certain speed and excite amacrine and ganglion cells.
2. If the stimuli represented by bipolar activation move in the leftward direction, then inhibition from the amacrine cell prohibits excitation from the bipolar cell in the right DSGC.
   - The left DSGC is activated.
3. If the stimuli represented by bipolar activation move in the rightward direction, then inhibition from the amacrine cell fails to prohibit excitation from the bipolar cell in the right DSGC.
   - The right DSGC is activated.

Our theory is that bipolar cells and amacrine cells pair together as subunits in our model to tile the *s* distribution (Fig. 7A, B). The width of the subunits' tile on the *s* distribution is proportional to the difficulty of the motion-classification task (Fig. 7C).

While this computational mechanism is not a new mechanism (excluding speed thresholds in bipolar cells) (Barlow and Levick, 1965; Wei, 2018), its emergence via task optimization alone is revealing and suggests that our task drives motion selectivity in retinal circuits. In line with previous task-optimized deep learning neural circuits (Yamins and DiCarlo, 2016; Kell and McDermott, 2019), it suggests that direction-selective circuits in the retina exist to optimally discriminate moving stimuli in a behaving animal's field of view. It should be noted that theories from task optimization are yet to achieve consensus in the

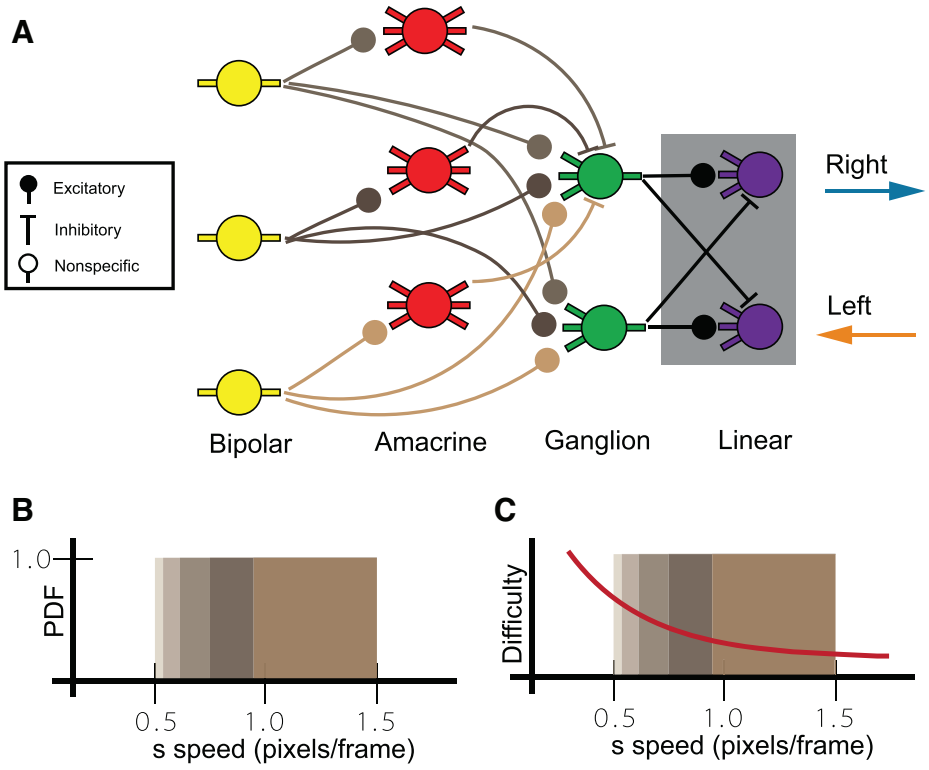

FIG. 7. Proposed computational mechanism for our model. **(A)** Bipolar/amacrine subunits transmit information to the direction-selective ganglion cell. **(B)** Subunits tile the *s* distribution. Each subunit is responsible for task performance in some range of the *s* distribution. **(C)** Subunits tile the *s* distribution with varying thicknesses proportional to the varying difficulty.

neuroscience community (Saxe et al., 2021), but it opens the door for more research into task-optimized retinal circuits. Task optimization could possibly provide insights into lingering computational questions in the retina where a computational model is not well established (Tanaka et al., 2019).

Another result of our work is an insight into how to build interpretable deep learning models. By organizing layers into cell types and enforcing Dale's principle, we were able to use simple methods inspired by other scientific fields to interpret how the deep learning model achieved high performance on the task. Deep learning models are often criticized for being *black boxes* (Saxe et al., 2021), but they allow for discovery of novel computational mechanisms (Tanaka et al., 2019) that can then be experimentally validated (Bashivan et al., 2019). A future direction for this work is to experimentally use deep dream movies from our model to maximally drive DSGC activities.

Another possible direction for our model is to include more biological constraints and mechanisms. Synaptic (Burnham et al., 2021) and molecular (Ozuysal and Baccus, 2012) dynamics have been linked to computational benefits and by including these into our model, more experimentally observed phenomena may emerge. It is also possible that speed thresholds emerge in bipolar cells due to a lack of adaptation mechanisms and by including more biologically relevant dynamics, speed thresholding bipolar cell types may be replaced by bipolar cell types that have been experimentally observed.

5. CONCLUSION

In conclusion, task optimization of biologically constrained CNN to motion-classification tasks led to the emergence of experimentally observed direction-selective retinal cells. This emergence suggests that optimization for ecologically relevant abilities is a driving force behind the development of retinal circuits. Biologically constrained CNNs also confer a high degree of interpretability through relatively simple methods. Further research into task-optimized deep learning models of retinal circuits would not only explain the purpose of retinal circuits but also purpose new computational mechanisms in the retina and create more interpretable deep learning models.

ACKNOWLEDGMENT

We acknowledge Sabrina Drammis for her helpful comments.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

FUNDING INFORMATION

Funding was provided by NSF Award Nos. CCF-0939370, CCF-1461559, and CCF-2003830. Additional funding was provided by the MIT Department of Electrical Engineering and Computer Science's SuperUROP program.

REFERENCES

Baccus, S.A., Ölveczky, B.P., Manu, M., et al. 2008. A retinal circuit that computes object motion. *J. Neurosci.* 28, 6807–6817.

Baden, T., Euler, T., and Berens, P. 2020. Understanding the retinal basis of vision across species. *Nat. Rev. Neurosci.* 21, 5–20.

Barlow, H., and Levick, W.R. 1965. The mechanism of directionally selective units in rabbit's retina. *J. Physiol.* 178, 477–504.

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

Burnham, D., Shea-Brown, E., and Mihalas, S. 2021. Learning to predict in networks with heterogeneous and dynamic synapses. *bioRxiv* bioRxiv:2021.05.18.444107.

Chichilnisky, E. 2001. A simple white noise analysis of neuronal light responses. *Network: Comput. Neural Syst.* 12, 199.
Eccles, J.C., Fatt, P., and Koketsu, K. 1954. Cholinergic and inhibitory synapses in a pathway from motor-axon collaterals to motoneurones. *J. Physiol.* 126, 524–562.
Gollisch, T., and Meister, M., 2010. Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron* 65, 150–164.
He, K., Zhang, X., Ren, S., et al. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
Kell, A.J., and McDermott, J.H. 2019. Deep neural network models of sensory systems: Windows onto the role of task constraints. *Curr. Opin. Neurobiol.* 55, 121–132.
Kell, A.J., Yamins, D.L., Shook, E.N., et al. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98, 630–644.
Kim, J.S., Greene, M.J., Zlateski, A., et al. 2014. Space–time wiring specificity supports direction selectivity in the retina. *Nature* 509, 331–336.
Kingma, D.P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv* arXiv:1412.6980.
Klein, S.A. 2001. Measuring, estimating, and understanding the psychometric function: A commentary. *Percept. Psychophys.* 63, 1421–1455.
Lindsey, J., Ocko, S.A., Ganguli, S., et al. 2019. A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnns. *arXiv* arXiv:1901.00945.
Maheswaranathan, N., McIntosh, L.T., Tanaka, H., et al. 2019. The dynamic neural code of the retina for natural scenes. *BioRxiv* bioRxiv:10.1101/340943.
McIntosh, L.T., Maheswaranathan, N., Nayebi, A., et al. 2016. Deep learning models of the retinal response to natural scenes. *Adv. Neural Inf. Process. Syst.* 29, 1369.
Meyes, R., Lu, M., de Puiseau, C.W., et al. 2019. Ablation studies in artificial neural networks. *arXiv* arXiv: 1901.08644.
Newsome, W.T., and Pare, E.B. 1988. A selective impairment of motion perception following lesions of the middle temporal visual area (mt). *J. Neurosci.* 8, 2201–2211.
Ocko, S.A., Lindsey, J., Ganguli, S., et al. 2018. The emergence of multiple retinal cell types through efficient coding of natural movies. *bioRxiv* bioRxiv:10.1101/458737.
Oh, S.W., Harris, J.A., Ng, L., et al. 2014. A mesoscale connectome of the mouse brain. *Nature* 508, 207–214.
Ozuysal, Y., and Baccus, S.A. 2012. Linking the computational structure of variance adaptation to biophysical mechanisms. *Neuron* 73, 1002–1015.
Richards, B.A., Lillicrap, T.P., Beaudoin, P., et al. 2019. A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770.
Saxe, A., Nelli, S., and Summerfield, C. 2021. If deep learning is the answer, what is the question?. *Nat. Rev. Neurosci.* 22, 55–67.
Song, H.F., Yang, G.R., and Wang, X.-J. 2016. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLoS Comput. Biol.* 12, e1004792.
Takemura, S.-Y., Bharioke, A., Lu, Z., et al. 2013, A visual motion detection circuit suggested by drosophila connectomics. *Nature* 500, 175–181.
Tanaka, H., Nayebi, A., Maheswaranathan, N., et al. 2019. From deep learning to mechanistic understanding in neuroscience: The structure of retinal prediction. *arXiv* arXiv:1912.06207.
Wei, W., 2018. Neural mechanisms of motion processing in the mammalian retina. *Annu. Rev. Vision Sci.* 4, 165–192.
Yamins, D.L., and DiCarlo, J.J. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365.
Yamins, D.L., Hong, H., Cadieu, C.F., et al. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. U. S. A.* 111, 8619–8624.
Yosinski, J., Clune, J., Nguyen, A., et al. 2015. Understanding neural networks through deep visualization. *arXiv* arXiv:1506.06579.

Address correspondence to:
Prof. Nancy Lynch
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
32 Vassar St
Cambridge, MA 02139
USA

E-mail: lynch@csail.mit.edu