Predicting individual neural responses with anatomically constrained task optimization

Authors

Omer Mano[1,2], Matthew S. Creamer[3], Bara A. Badwan[4], Damon A. Clark[1,2,3,5,6]

Affiliations

1 – Department of Molecular Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA
2 – Department of Neuroscience, Yale University, New Haven, CT 06511, USA
3 – Interdepartmental Neuroscience Program, Yale University, New Haven, CT 06511, USA
4 – School of Engineering and Applied Science, Yale University, New Haven, CT 06511, USA
5 – Department of Physics, Yale University, New Haven, CT 06511, USA
6 – Lead contact: damon.clark@yale.edu

Abstract

Artificial neural networks trained to solve sensory tasks can develop statistical representations that match those in biological circuits. However, it remains unclear whether they can reproduce properties of individual neurons. Here, we investigated how artificial networks predict individual neuron properties in fruit fly motion circuits. We trained anatomically-constrained networks to predict movement in natural scenes, solving the same inference problem as fly motion detectors. Units in the artificial networks adopted many properties of analogous individual neurons, even though the networks were not explicitly trained to match these properties. Among these properties was the split into ON and OFF motion detectors, which is not predicted by classical models for motion detection. The match between model and neurons was closest when the model was trained to be robust to noise. Our results demonstrate how anatomical, task, and noise constraints can explain response properties of individual neurons in a small neural network.

Introduction

Biological neural networks (BNNs) have evolved through natural selection to perform tasks that promote survival, but it is often unclear how their properties relate to the tasks they perform. Recent work in sensory systems has shown that artificial neural networks (ANNs) optimized to perform ethologically-relevant tasks often develop stimulus representations similar to those in BNNs. For instance, ANNs trained to categorize visual objects possess intermediate representations similar to those in the hierarchical processing steps in primate visual cortex [1]. Similarly, representations of temperature in zebrafish are similar to those in artificial neural networks trained to navigate thermal gradients [2]. These comparisons between ANNs and BNNs test a hypothesis about the goal of the biological circuit: is optimizing a network to perform a specific task under specific constraints sufficient to account for a set of the biological network's properties?[3-5] These prior studies have drawn connections between clusters of nodes or layers in an ANN and the heterogeneous response properties of groups neurons in regions of the brain. In this study, we show that specific nodes in a trained ANN can have properties that correspond to individual neurons in a biological circuit. To do this, we apply connectomic constraints to small ANNs to create an *a priori* correspondence between specific ANN nodes and individual neurons in the biological network. In this framework, we compare the task-optimized ANN to the

evolved BNN to show how optimization and constraints—especially noise—account for the properties of individual neurons in a biological circuit.

We focus on the fruit fly *Drosophila*'s motion detection circuits (**Fig. 1A**), which are critical to the fly's visual navigation behaviors [6-9]. These circuits are well-studied, so that anatomical connectivity has been measured [10-13], along with many functional properties of neurons in the circuit [7, 9, 14-31]. These motion circuits have evolved two types of motion detectors: T4 neurons, which are selective for moving light edges, and T5 neurons, which are selective for moving dark edges. T4 and T5 neurons are arranged retinotopically so that for each location in visual space there are two T4 neurons and two T5 neurons sensitive to motion along the horizontal axis, one T4 and one T5 neuron sensitive to leftward motion and one of each sensitive to rightward motion. Each individual neuron in these classes receives excitatory and inhibitory input from neurons that signal visual intensity in 3 spatially-separated locations (**Fig. 1A**) [10, 11, 18].

Textbook models for motion estimation, including the Hassenstein-Reichardt correlator model and the motion energy model [32, 33], may be largely derived from first principles [34, 35] and suggest that temporal delays, spatially separated inputs, and nonlinear processing are critical to the task of motion detection [36]. These models specify a minimum set of conditions and plausible computations to arrive at direction-selective signals, but they fail to account for many of the features measured in the fly's motion circuits. In particular: (1) The three spatially-separated inputs to T4 and T5 are organized such that the central signal is fast, while the two flanking signals are slow [18, 24, 37] (**Fig. 1B**). Moreover, when local luminance increases, one flanking signal has the opposite influence on the downstream motion detector compared to the other two spatial locations [10]. For an individual T4 cell, activity is effectively inhibited by light at one location in visual space while excited by the other two, and T5 cells are effectively excited by light at one location while inhibited at the other two [16, 27]. None of these properties of the inputs to the fly motion detectors follows clearly from classical models. (2) Horizontal motion detection is organized into four parallel pathways, consisting of light and dark moving edge detectors in both horizontal directions [7] (**Fig. 1C**), a split not present in or explained by classical models. (3) Although T4 and T5 neurons are direction-selective, they also respond to specific stationary light or dark edges [30] (**Fig. 1D**), an unexpected result for cells that detect visual motion. (4) The motion detectors T4 and T5 show opponent suppression: they respond less to the sum of null and preferred direction motion stimuli than to preferred direction motion stimuli alone. This phenomenon runs counter to predictions of common, classical motion detection models [29] (**Fig. 1E**). (5) Last, the four neurons that encode horizontal motion have signals that tend to be non-coactive when presented with moving natural scenes, so that their signals are decorrelated [28] (**Fig. 1F**). This decorrelation is not addressed by classical models of motion detection. Since classical models do not account for this suite of qualitative properties of identified neurons in the fly's motion circuits, we asked whether they could be explained by optimizing a network to detect motion under the anatomical constraints of the fly's motion circuits.

In this study, we developed a set of three shallow, convolutional ANN models ranging from abstract to more biophysically realistic. The models created a direct correspondence between analogous ANN units and BNN neurons. We trained these models to predict the velocity or direction of moving natural scenes, and then examined their solutions and response properties. The trained models could account for many response properties of individual neuron types

measured in the fly's motion circuits, including the five listed above (**Fig. 1**). Therefore, the task of predicting natural scene velocities, combined with anatomical constraints from the circuit connectivity, were sufficient to account for circuit response properties that are not accounted for by classical models. Moreover, by investigating different model constraints, we found that robustness to noise was the primary factor that generated artificial units with properties like the neurons in the fly's circuits. In sum, these results show that many unexplained properties of individual neurons in this small neural network are consistent with and predicted by a system optimized for motion detection in the presence of noise.

Results

Detecting motion in natural scenes

Our goal is to relate optimized ANNs to the evolved circuits in the fly. To make this comparison, we began by setting up a problem for the artificial networks to solve that is similar to problem solved by the fly (**Fig. 2A-C**). In the fly eye, the direction-selective neurons T4 and T5 use luminance information over time from different points in space to infer the direction and speed of visual motion [34, 35, 38]. Individual T4 and T5 neurons perform this task in small regions of the visual field, receiving input from approximately three neighboring columns [10, 11, 18]. They perform these operations while the fly navigates natural environments.

To approximate the naturalistic inputs to fly motion detectors, we rigidly translated panoramic natural scenes [39] using stochastic velocities in yaw rotation (**Fig. 2A**). The rigid translation of panoramic scenes ignores the occlusions and the different angular velocities that arise from an animal translating through the world, but it mimics closely the type of scenes generated by an animal purely rotating in the world. Flies use motion detection circuitry to stabilize their orientation in the face of angular perturbations [40, 41], so this is a reasonable starting point. Rigid translation of natural scenes has been used with some success in other studies aimed at understanding processing properties of motion detectors [38, 42-47].

The stochastic velocities for scene motion were drawn from a zero-mean Gaussian distribution with standard deviation of 100º/s and a correlation half-life of 200 ms (**Fig. 2B**, see **Methods**). This amplitude of turning is typical of walking flies [48, 49]. In this study, it is critical that stimuli do not have a constant velocity over time, since constant velocities would allow stimuli arbitrarily far in the past to inform current velocity estimates [35]. The correlation time of 200 ms roughly matches correlation times in fly turning during walking [48, 49], and ensures that only recent information can be used to infer current visual velocities.

Last, we approximated the optical filtering of scenes by the discrete fly ommatidia. These filtering properties largely persist into downstream medulla neurons [24, 50]. To do this, we created discrete signals with separation of 5º, roughly matching the separation of neighboring *Drosophila* ommatidia [51]. For each ommatidial signal, we spatially filtered the scene at each time point with a two-dimensional Gaussian that roughly matched the acceptance angles of ommatidial optics [51], creating sets of 72 ommatidial signals from 360º horizontal strips across scenes (**Fig. 2C**).

Overall, these procedures generated a dataset with naturalistic neighboring ommatidial signals over time associated with a specific random velocity trace. The signals were obtained from

random elevation and azimuthal positions on randomly chosen panoramic images. The task for the ANN (and for the fly eye) is to infer the velocity or direction of motion (latent variables) from this suite of ommatidial luminance signals.

Shallow neural networks for motion detection

We developed shallow model architectures that incorporated varying degrees of biophysical detail. To do this, we defined three basic unit types for motion detection (**Fig. 2D**). All unit types received inputs over time from three neighboring ommatidia, constraining the units to match the three spatial inputs measured to T4 and T5 [10]. The units linearly filter these inputs in time with three distinct kernels that are learned through training. The three different types of units are distinguished by the nature of the nonlinearity used to combine the three spatially-offset inputs (see **Methods**).

The first unit type employs a linear-nonlinear (LN) processing step, so that the temporally filtered signals are simply added together and the sum acted upon by a threshold-linear rectifier (**Fig. 2D**, *left*). A nonlinearity is required to generate direction-selective signals [33, 36]. This unit type is closely related to the motion energy model and is similar to models describing directional neural signals in mouse retina [52, 53] and directional and other signals in mammalian cortex [54-57]. It is also similar to models suggested to describe directional signals in T4 and T5 [16, 17]. We call this the linear-nonlinear (LN) unit.

The second unit type employs an additional threshold-linear rectifier after each ommatidial signal is filtered in time, but before the three signals are linearly combined and thresholded again (**Fig. 2D**, *middle*). This rectification of the signals from each spatial location mimics rectification observed in the calcium and voltage signals of medulla interneurons upstream of motion detectors in the fly [15, 50, 58]. Because this model involves two sequential stages of linear-nonlinear processing, we call this the LNLN unit.

The third unit type also rectifies the filtered ommatidial signals, but it uses a synaptic nonlinearity to combine the three filtered, rectified signals (**Fig. 2D**, *right*). This synaptic nonlinearity considers each of the three inputs to be synaptic conductances with associated reversal potentials, which are learned though training. The nonlinearity is a weighted sum of the conductances divided by an unweighted sum of the conductances (see **Methods**). This model is similar to other biophysically realistic models for T4 and T5 direction-selectivity [19, 21, 29, 59, 60]; models of downstream, wide-field neurons have also productively taken conductances into account [61]. The biophysical model in this case assumes a pseudo-steady-state response, which is justified by the fast cellular time constants measured in T4 [19]. We call this the synaptic nonlinearity unit.

These three unit classes are nested within one another. That is, the LN unit is a special case of the LNLN unit, and the LNLN unit is a special case of the synaptic nonlinearity unit. Thus, progressing from LN to LNLN to synaptic nonlinearity adds more parameters, and in principle, the more complex units can only perform better, since each could still obtain the solution of the simpler units.

These three unit types were each placed into models with architecture that matched the circuitry in the fly eye (**Fig. 2E**). The three model classes consisted of multiple units of the same type, and the weights in each unit were optimized through training on the naturalistic dataset we defined. In this architecture, each unit (+) was paired with a unit (–) constrained to be mirror symmetric in space, and the two resulting signals were subtracted from one another. This differencing reflects the opponent subtraction of oppositely tuned motion signals that occurs in the fly eye downstream of T4 and T5 [62, 63]. In each case, two pairs of symmetric units were trained (A_+/A_- and B_+/B_-), unless otherwise noted. In all three model classes, the temporal filters were free parameters, as were weights in linear combinations and biases before the rectifications.

We scaled the natural scene training images so that each set of 72 ommatidial signals had zero mean and unit variance. This is because early visual processing computes deviations from average, rather than absolute luminance levels [26, 64, 65]. The signals arriving at motion detectors in flies also undergo processing to normalize signal amplitude [66, 67].

Last, we added two forms of noise to our models (**Fig. 2E**, see **Methods**). First, we imposed additive noise at the input signals, after contrast computation. This front-end noise reflects noise noted in photoreceptor and lamina cell signaling [68, 69]. Second, we included multiplicative noise at the output of each unit of the model before they were subtracted to generate the overall signal. This back-end noise represents intrinsic noise in the circuit [70], which could arise from variability in the signals and signal transmission of directional units [19, 21]. We varied both the front-end and back-end noise to investigate how noise affects the types of solutions found by fitting our models.

Training models

We used TensorFlow [71] to train multiple instances of these three model classes using gradient descent with different initializations (**Fig. 2F**, see **Methods**). The models were trained to use the preceding 300 ms of visual data, reflecting plausible filtering properties of biological neurons, to predict the current scene velocity. During training, the models were optimized by adjusting the temporal filter weights, as well as biases that were applied before each nonlinearity, and additional weighting parameters in the LNLN and synaptic nonlinearity model. The different models all converged on solutions, but the more complex LNLN and synaptic nonlinearity models converged more slowly, and the converged solutions had larger variability in their performance (**Fig. 2F**). We evaluated model performance on a hold-out dataset, which was independent of the training data. Model output depended on the particular scene, but gave reasonable velocity estimates over many scenes (**Fig. 2G, H**).

Trained models possess the neuronal features of fly motion detectors

When we trained the three model classes to predict image velocity in the presence of noise, the trained models showed many of the non-canonical properties possessed by the fly's motion detectors (**Fig. 3**). Most importantly, the paired units in all three models could be classified as 'T4-like' or 'T5-like', based on whether they responded most to light or dark flashes. We evaluated the properties of trained models in a noise-free regime, corresponding to a bright visual stimulus and responses averaged over many trials (see **Methods**).

All three trained models had units possessing temporal filters with similar shapes, and with relative dynamics and polarities similar to those measured in cells upstream of T4 and T5 (**Fig. 3(i)**). The measured filters are slower than those in the trained ANNs, potentially because they represented calcium measured using optical indicators [72]. However, like T4 and T5 cells (**Fig. 1B**), all trained units had high-pass filters on the center input, and slower, more lowpass filters on the flanking inputs. The central input of the T4-like units were sensitive to positive derivatives, while the center input of the T5-like unit were sensitive to negative derivatives, just as in T4 and T5 cells. Both T4- and T5-like units had a positively-signed filter on one side and a negatively-signed filter on the other, in the pattern of T4 and T5 cells.

In all trained units, the third filter (f_3) in the trained T4- and T5-like units had a small initial response of the opposite sign to its prolonged, delayed response. This feature was not observed in measurements of calcium in the cells proposed to correspond to input 3 (**Fig. 1A**) [24, 37], or in measurements of voltage responses in T4 or T5 [19, 21]. In the learned filters for the LN, LNLN, and synaptic nonlinearity T4- and T5-like units, the prolonged, second lobe had a larger integral than the initial lobe by factors of 10 to 15. Thus, the second lobe tended to dominate the initial transient.

This pattern of temporal filtering in the trained models led to strong direction- and edge polarity-selectivity (**Fig. 3(ii)**). Each unit responded much more strongly to a single direction and a single edge type (ON-edges or light edges vs. OFF-edges or dark edges) than to any other combination. The ON- vs. OFF-edge selectivity of each unit corresponded to the sign of the central derivative filter, just as in the fly's circuitry. The direction-selectivity corresponded to the signs and shapes of the two flanking filters. The LNLN model was more selective than the LN model, responding exclusively to one edge type, while the synaptic nonlinearity showed intermediate selectivity. Critically, all three models generated ON- and OFF-edge direction-selective units, even though no such constraint was imposed on them.

Several other features of the fly motion circuits were also reproduced. All three models showed stationary edge responses that matched the empirical response patterns in T4 and T5 (**Fig. 3(iii)**). In these cases, the trained units responded to edges of the same polarity as the analogous fly neuron (T4 or T5). The neurons T4 and T5 respond less to a sum of preferred and null direction sinusoids than to preferred direction sinusoids alone (**Fig. 1E**) [29], a form of opponent suppression. When the models were trained, the LNLN and synaptic nonlinearity models also showed this sort of opponency in the responses of their individual direction-selective units (**Fig. 3(iv)**). (The LN units are mathematically incapable of generating this opponency [29].) We observed low coactivation between units presented with natural scenes, with coactivation decreasing from LN to LNLN to synaptic nonlinearity models (**Fig. 3(v)**). Last, when presented with sinusoidal stimuli, these trained models respond to signal strength and to temporal frequencies with tuning that roughly matches physiological and behavioral measurements downstream of T4 and T5 [9, 73-75] (**Fig. S1**). This was particularly true of the LNLN and synaptic nonlinearity models.

Collectively, these data show that these model classes, when trained to predict natural scene velocities, adopt many properties of T4 and T5 circuits that are not explained by classical models of motion detection. Thus, this training regime is sufficient to account for a wide array of

specific response properties found in this circuit. Solutions found by ANNs depend not only on the loss function, but also on constraints imposed on the network. To understand how model and training constraints affected model solutions, we therefore set about investigating how different aspects of the model classes, loss functions, training data, and noise affected the trained solutions. Since the trained LN model is readily interpretable and can account qualitatively for most of the biological data, we focus on that model for the remainder of this study, except to probe opponent suppression.

ANN solutions do not depend strongly on the loss function or training data

First, we asked how the model solution depended on the loss function being optimized. We initially trained the models to estimate the true velocity, minimizing a loss function equal to the squared error between model output and the instantaneous image velocity (**Fig. 3**). However, while this objective for model motion detectors has been used previously with some success [38, 44], fly motion detectors might instead have evolved predict some other, nonlinear function of the true velocity. How much does the solution depend on the loss function? To answer this question, we trained models to predict an extreme function of the velocity: its direction only (**Fig. 4AB**). The LN model was trained to classify just the direction of the motion, without regard to its speed (see **Methods**). Interestingly, the units in the trained models looked largely identical in this case, becoming direction and edge polarity selective, sensitive to stationary edges, and showing little coactivity between units.

We wondered whether the mirror symmetry we had imposed on our model pairs would arise naturally through training. We trained a set of four units without the mirror symmetry pairing, using 12 independent temporal filters, 3 for each units. We found that the best performing solutions always included two mirror-symmetric pairs that were subtracted (**Fig. S2**). This likely reflects the mirror symmetry imposed in our training dataset, which matched the natural world's visual mirror symmetries.

Next, we asked whether the division into ON- and OFF-edge detector units that we observed (**Fig. 3**) depended on asymmetries in light and dark in natural scenes. These natural scene asymmetries have been hypothesized to account for a variety of asymmetries in fly behavior [38, 43, 76] and differences between T4 and T5 [44]. Could those asymmetries in the inputs also lead to these models splitting detector units into ON- and OFF-edge selective units? One may imagine alternate divisions between unit pairs, for instance one pair tuned to fast and one to slow stimuli. To address this question, we trained the models with the same velocity distribution, but instead of panoramic naturalistic photographs as the visual input we used sinusoidal gratings (**Fig. 4C**). Unlike the photographs, the sinusoidal gratings are light-dark symmetric. Interestingly, the two unit pairs in each model still became sensitive to ON- and OFF-edges. We wanted to test whether this split into ON- and OFF-edge selective channels depended on the precise nonlinearity we used. When we changed the LN-model's nonlinearity from a threshold-linear function to a saturating, sigmoid function while training on natural scenes, it had little effect on the model solution (**Fig. S2**). Our results indicate that the division into ON- and OFF-edge selective units is a natural outcome when estimating motion in scenes that contains both positive and negative contrasts and when there are two unit pairs available to optimize.

Largest marginal performance improvement comes from adding the second detector pair

We wanted to better understand why flies have two primary motion detectors types (i.e., T4 and T5 neurons), rather than 1 or 3. Our initial models had included two different unit pairs, which in trained models developed properties similar to T4 and T5. We therefore created and trained LN models with different numbers of unit pairs, ranging from 1 to 5 (**Fig. 4D**). Increasing the number of unit pairs increased model performance under low- and high-noise training conditions, but the largest marginal improvement in performance came from increasing from 1 unit pair to 2 unit pairs, where the performance metric more than doubled under the low-noise training conditions. After that, adding more unit pairs provided smaller performance improvements. If the cost of adding additional units in biological systems is high, this result may explain why flies have only two elementary motion detector types, tuned to light and dark edges respectively.

Training with high noise is more robust to changes in noise

We next asked how noise during training affected the structure of solutions. To investigate this, trained LN models under a range of front-end noise and back-end noise conditions. We then asked how well models performed when tested under conditions that were different from their training noise level. The best-performing models in a particular noise regime were the ones trained under that same noise regime (**Fig. 4E**). However, when models trained in high noise regimes were tested in low noise regimes, they still performed reasonably well, while models trained in low noise regimes performed very poorly in high noise regimes. Similarly, the high-noise trained model performed better over many noise regimes (**Fig. S2D**). This effect held for both front-end and back-end noise.

The high-noise trained models performed worse on the hold-out training data because they are solving a far more difficult task when there is substantial noise injected (**Fig. S3**). Importantly, the high-noise trained units were far more direction-selective to sinusoids than low-noise trained units (**Fig. S3E**), better matching the strong direction-selectivity to sinusoids of T4 and T5 cells [7, 9, 16, 17]. We therefore set out to compare properties of the high- and low-noise trained solutions.

Training noise strongly affects direction-selectivity and edge-polarity-selectivity

The noise amplitude at both the front- and back-end substantially changes the learned solutions (**Fig. 5**). First, the front-end noise amplitude dramatically changed the temporal extent of the learned filters (**Fig. 5(i)**). When more noise was added, the filters became more extended in time, averaging over time to minimize the influence of the noise. The correlation time scale of the velocity means that averaging over more than ~200 ms is not useful for computing the current velocity [35]. With less noise, there was less need to average, and using only the most recent measurements of intensity produced the best estimate of the current velocity.

Second, the back-end noise strongly influenced the degree of edge-polarity- and direction-selectivity in the individual LN units (**Fig. 5(ii)**). In the high noise cases, the T4- and T5-like units were more edge-selective and more direction-selective. In the low noise cases, the units responded strongly to light edges in one direction and dark edges in the other, with a slight imbalance that was direction-selective; this pattern is unlike T4 and T5 responses. In low noise, the trained units are responding strongly to spatial gradients and only slightly to direction. In the low noise case, the opponent subtraction step could cancel out large responses, leaving only the

small difference as an estimate of motion. When back-end noise was added, this computational strategy was no longer viable, since the subtraction of the paired unit could no longer reliably subtract the non-direction-selective components of the responses. As a result, the individual units within each pair converged on solutions that were robust to this noise by being more direction-selective even before the subtraction step.

Last, adding noise to the system made the units less selective for stationary edges (**Fig. 5(iii)**). In all cases, the spatial pattern of responses to stationary edges matched those in T4 and T5, but when more back-end noise was added, the units responded less to these stationary edges. This seems likely to be closely to related to the increase in edge-polarity and direction-selectivity with increasing noise. The back-end noise prevents precise cancellation of the signals from stationary scenes, making it advantageous for the model to respond less to such stimuli. In the fly, responses to these stationary edges are about one quarter of responses to preferred direction moving edges [30].

Increased noise increases opponency and sparsity

To evaluate the effects of noise on opponency and sparsity, we performed the same sweep of front- and back-end noise while training LN models (**Fig. 6(i)**). We measured unit opponent suppression as the degree to which the mean response was decreased when a null-direction sinusoid was added to a preferred-direction sinusoid. In the case of LN models, the response to the sum can never be less than the response to the preferred-direction sinusoid alone [29]. But increasing the noise in the system made the response to the sum closer and closer to the response to the preferred direction sinusoid alone (**Fig. 6CD(i)**).

When we trained the LNLN model with different noise levels, opponency increased with increasing noise levels (**Fig. 6(ii)**). That is, the units decreased their response to the sum of the sinusoids in the presence of high noise (**Fig. 6B-D(ii)**). Opponency in these primary directional cells in *Drosophila* has been hypothesized to cancel out 'common mode' correlations, leaving a larger dynamic range for motion signals [29]. When noise is added to the model, it may be more important to make the unit signals as direction-selective as possible, and the opponent properties measured could reflect that additional direction-selectivity.

Last, we examined how changing the noise characteristics affected the decorrelation of the signals among units in an LN model (**Fig 6(iii)**). As back-end noise was increased during training, the resulting LN units became less coactivated by the naturalistic stimuli. This seems linked to the increased direction- and edge-selectivity of the units under larger back-end noise (**Fig. 5(ii)**). When units are more edge- and direction-selective, a given stimulus activates only one of them, since there should be only one edge type moving in one direction through the model receptive field.

Discussion

This study has demonstrated the potential for fine-grained, neuron-level mapping between task optimized ANNs and real neural circuits. Results showed that an optimized model for visual motion detection could account for many measured neural properties in the *Drosophila* motion detection circuits that are not predicted by textbook models for motion detection, like the Hassenstein-Reichardt correlator and motion energy models [32, 33]. Anatomical constraints from

the real circuit were key to developing this correspondence. Our work also demonstrated that robustness to noise was critical to generating artificial networks that matched measured properties.

Importantly, these results were not built into the fitting routine or model architecture. The tests of similarity between ANN and BNNs were also distinct from the training data. One can imagine other solutions that might have performed well in the training. For instance, one could imagine splitting the A_+/A_- and B_+/B_- unit pairs to be tuned to fast and slow motion, respectively, thus covering a wide range of input velocities, rather than dividing them into ON- and OFF-edge detectors. Alternately, the ON- and OFF-edge segregation need not be complete, as happens in the low-noise optimizations, where units were not particularly edge or direction selective (**Fig. 5**). Last, the two flanking filters do not need to have opposite signs and be delayed with respect to the center filter: if all three filters had single lobes with the same sign and have delays of τ, 2τ, and 3τ, they could sum up above a nonlinear threshold only for stimuli consisting of motion in one direction. However, these counterfactual solutions did not occur when the models were optimized for performance. This leads us to interpret these features in fly motion detectors as having evolved to optimize performance in motion detection, and suggests that we have identified crucial constraints on the circuit.

Loss functions and optimization

In this study, we used loss functions that minimized error in predicting the velocity or direction of a moving natural scene. How realistic is this task? Motion detectors in flies generate graded responses that depend on direction and speed [36], so it's a reasonable place to start. But future studies could incorporate more realistic tasks, such as training a motion detector to act as the input for an agent-based model that attempts to move with a stable course through an environment. Such a task would require incorporating knowledge of the downstream circuitry and locomotor control [77]. One could also imagine that motion-sensing would arise from networks trained to detect and land on objects, which would be a highly ethological task. More simply, one could also incorporate known downstream circuitry, such as the shunting mechanisms that perform a kind of gain control in spatial integration of T4 and T5 units [61, 78]. Such studies might generate new hypotheses about the evolutionary origin of motion detectors. Here, the simplest loss functions we considered appear sufficient to generate many of the features in the fly's circuits.

Our study used gradient descent to optimize the models. We examined the best performing models from a suite of initializations, since models could become trapped in sub-optimal local optima. How might optimization occur in the fly's visual circuit? There is some experience dependent plasticity in flies dependent on light level [79], but it seems likely that optomotor circuit structure and function is genetically determined to a large degree, and optimized over generations of natural selection. The gradient descent procedure we used can become stuck in local optima because our networks are shallow and not over-parameterized [80]. Interestingly, optimization algorithms that are similar to natural selection can optimize models efficiently and may be able to avoid local optima [81]. Our results show that one can think productively about these visual circuits as solutions to an optimization problem, solved by evolution.

Influence of noise

The model features that matched biology did not arise from the task and network structure alone, but depended critically on noise in the system. The back-end noise we added made units more direction-selective and forced them to have larger differences between preferred and null direction responses. In that way, it penalized large, correlated responses from opposing units. Thus, adding noise had an effect similar to adding a sparsity constraint explicitly, for instance by adding a term to the loss function proportional to the absolute value of all unit responses. Sparsity is commonly observed in neural systems [82], and non-coactivity of parallel motion detectors has previously been hypothesized to organize their response properties [28]. Here, by adding noise during training, we can see one logic of the sparse solution, since the non-coactivation of the units makes the system more robust to noise. Interestingly, the common technique of dropout training, in which only a stochastic subset of weights are updated during each learning iteration, is equivalent in to injecting certain types of noise into the network [83]. This means that many artificial networks trained using dropout techniques are already implicitly trained to be robust to corruption by noise. By adding this noise explicitly, we control this constraint and can more easily relate it to biological sources of noise.

Sources of noise

Given the influence of noise on the model solutions (**Figs. 4, 5, 6**), it is important to ask whether the noise injected into the models is consistent with what is known about the fly's visual circuits. If so, then our modeling suggests that those noise sources impose strong constraints on the circuit that influence the solutions that have evolved in the fly.

Front-end noise could be attributed to fluctuations in photoreceptor signals or signals in downstream lamina and medulla cells. Photoreceptor signal-to-noise has been well-characterized and depends strongly on the absolute light intensity, as well as on temperature [69, 84]. When light intensity is high, the signal-to-noise ratio of photoreceptors in flies can be ~10, while under low-light conditions, the signal-to-noise ratio can decrease to ~0.1 (ratio of powers). This range extends beyond the range of noise in our sweeps (**Figs. 4-6**). Since the front-end noise is variable, it seems likely that the fly has evolved to deal with the full range, not just a single noise level, as in our numerical experiments.

Less is known about noise deeper in the visual system. Studies in locust have suggested that signal-to-noise actually decreases in feature detectors further from the sensory periphery [85]. Electrical recordings of T4 and T5 responses to strong driving stimuli show relatively little trial to trial variability (SNR of ~10, signal/std noise) but higher variability between cells (SNR of ~2 mean signal/std) [19, 21]. The larger variance between cells could reflect long timescale gain fluctuations within cells. Noise within a cell could also be amplified by expansive nonlinearities that transform voltage into calcium and calcium into synaptic release. Synaptic transmission might also decrease signal-to-noise due to synaptic vesicle release statistics, since it's metabolically expensive to transmit high SNR signals [86]. With more careful measurements of the noise characteristics of T4 and T5, one could add more accurate, spectrally-matched noise models to the fitting procedure performed here.

Although we trained models with noise at specific levels, the biological circuit is likely exposed to varying levels of noise, dependent on stimulus and internal state. Since the high-noise training

regime is most generalizable across noise regimes (**Fig. 4E, Supp. Fig. S2**), it may be that training in high-noise is most similar to optimization under a range of different noise levels.

Structure of delays in 3-input motion detectors

The optimized models in this study consistently showed a fast central input and delayed flanking inputs with opposite signs (**Fig. 3**). This configuration appears in the fly motion detectors (**Fig. 1A**) but has also been suggested to explain cortical direction-selective signals [87]. This functional organization emerged with all three unit types, and did not depend on whether the model was predicting the stimulus velocity or just its direction (**Fig. 4**). It has a clear orientation in space-time, suggestive of motion energy-like processing [16, 17]. Interestingly, this delayed-opposite-flanks weighting structure also appears in a completely different optimization task, in which a network is trained to preserve similarity under translation of images [88]. This flanking organization in the circuit has been postulated to improve opponency through synaptic nonlinearities [29], but it is present in the LN and LNLN trained models, so it appears to be helpful even without the synaptic nonlinearity. Thus, this spatiotemporal weighting structure in motion detection acts flexibly to solve many different constraints and optimization problems. This could also serve the fly's visual system well, since neurons downstream to T4 and T5 are specialized to detect both visual flow [6, 63] and looming stimuli [89], and likely other visual features [90].

ON- and OFF-edge detectors and natural scenes

In this study, ON- and OFF-edge selective motion detectors emerged naturally as solutions to the task of detecting motion. This did not depend strongly on the loss function, training data, or form of the nonlinearity (**Fig. 4, S2**). The units must remain near the nonlinear threshold in order to generate direction-selective signals, and it appears that there is a greater benefit to tuning units to ON- and OFF-edges, rather than choosing to respond to ON-edges only, for instance, but with unit pairs tuned to different speeds. If the system did not contain both ON- and OFF-edge selective pairs, it would not respond to roughly half of all inputs. This logic could explain parallels in motion computation among species[91, 92], including the split into ON- and OFF-edge motion detectors in flies and in mouse retina [93, 94] and the evidence for edge polarity-selective motion responses in primate cortex [30, 76, 95-98]. Flies, zebrafish, and humans all treat light and dark signals asymmetrically in computing motion in ways that seem tuned to improve naturalistic performance [38, 43, 44, 76, 96, 99]. This suggests there are additional benefits to pathway splitting not explored here. One powerful explanation for sensory spliting into ON and OFF pathways is based on preserving stimulus information under metabolic constraints [100, 101], but it is not clear whether that logic maps onto the optimization task here, in which the model infers a latent variable and there is no obvious analogue to a metabolic constraint.

Stationary edge responses

Prior experiments have shown responses in T4 and T5 neurons to flashes [17, 19, 21, 23, 25, 28] or sinusoids [102], which are consistent with classical models [33]. Our results shed light on T4 and T5 responses to stationary edges of specific polarities (Fig. 1D) [30]. Under low-noise conditions, models performed best when they were strongly sensitive to edges of a single polarity and only mildly direction-selective (**Fig. 5(ii)**). The non-directional responses were cancelled by other units. Under high-noise training, direction-selectivity increased, reducing their responses to stationary edges (**Fig. 5(iii)**). These results suggest that a motion detector unit predicts motion best when it responds to stationary edges and has partner units to cancel this signal. This

approach is limited by system noise, reducing the response amplitude to stationary edges. The biological responses to stationary edges may reflect the emphasis on spatial gradients in the low-noise solutions.

Circuit features neglected in these models

While our three models proceeded from more abstract to more biophysical, they all neglected many known features of the circuit, which could have important effects on the learned solutions. The network architecture was based on connectome reconstruction, but it focused on connections between T4 and T5 and their inputs and did not take into account many other circuit features. We summarize here some important simplifications made in this study. (1) We represented all the circuitry upstream of medulla interneurons as a simple spatial and temporal filter, when in fact there is complex gain control and changes in dynamics that take place upstream of medulla interneurons [69, 103-105]. Our calculation of contrast during training may relate to some of these early operations. Moreover, all these early visual neurons have nonlinear response properties, while we focused on only the rectification in the inputs to directional units and the nonlinearities within those units. (2) Visual interneurons upstream of T4 and T5 have different shapes of receptive fields, including center-surround antagonism [14, 24, 50, 104], which could influence their response properties in performance-based fitting procedures. (3) In training models, we assumed perfect contrast normalization, when in fact there are dynamics and spatial scales for this operation [66, 67]. (4) Our units have only one neuron at each of the three spatially separated inputs, but anatomy suggests there are multiple input neurons at some positions [10, 11, 13], and functional studies show they may interact nonlinearly [18]. (5) Our models are feedforward, when there are multiple instances of lateral interactions and feedback in the true circuit [10, 11, 13]. (6) Early temporal and spatial processing change to integrate signals differently under different levels of signal and noise [64, 65, 106], while our model did not include adaptation. In understanding constraints on the system, some of these features are likely to be important to determining the solution. However, the simplifications made in this study still allowed us to generate trained models with neuronal features shared by the biological circuits.

Performance optimization and model realism

Models of motion estimation in the fly range from the abstract to the biophysically detailed. The abstract ones are harder to relate to circuitry, but are easier to understand and can explain broad phenomenology [34, 35, 38, 107]. The biophysically detailed ones have the power to explain specific voltage signals [19, 21, 59, 60]. A range of models in between these two extremes connect motion detection to various aspects of the fly's specific circuitry [16, 17, 29, 44, 67, 108]. In the modeling here, we moved across this spectrum by optimizing three models that ranged progressively from more abstract — similar to motion energy models [33] — to more biologically realistic — similar to previously published models that could be hand-tuned to perform well [29, 59]. The more abstracted models provided results that were easier to interpret, but this work shows how abstracted models can be related to the biologically realistic ones in terms of the performance and properties of optimized solutions. It was not necessary to include synaptic biophysics to reproduce the circuit features we examined here, but it was helpful to include rectifications that occur upstream of the motion detectors (**Fig. 3**).

This work adds to a suite of models that have shown how constraints and optimization contribute to sensory processing. Some of these models have been fit directly to predict data [109, 110], while

ours and others [1, 2, 111] have been optimized to perform specific tasks. Our work is closest to two prior approaches. In fitting retinal responses to a convolutional neural network, other studies have found that the model that fits best has units that look similar to the responses of the progression of cell types in the retina [109, 110]. These studies included temporal processing, as ours did, but had weaker anatomical constraints, using 3 layers of units, without specifying a priori how units in each layer were connected. The artificial network was fit to recorded retinal outputs, so features of the artificial network reflect circuit components but do not provide information about the tasks performed by the biological circuit.

A different approach used detailed connectomic data to investigate the fly motion circuits by training a network to detect the position and displacement of a visual object in a movie [111]. That study employed a far more detailed set of connection constraints, encompassing 40+ neuron types arrayed over a large swath of visual space. It used a separate network to interpret the outputs of the fly eye. That study found that it could obtain direction-selective signals in T4 and T5 neurons in the model when using measured synaptic connectivity with inferred signs and manually imposed delays. In the present study, we focused on a small set of neurons upstream of T4 and T5 with well-defined spatial receptive fields and we fit both temporal processing and synaptic weighting. This allowed us to interpret how processing properties in a shallow, feedforward ANN compared to those measured in neurons in the biological circuit.

Mappings between artificial networks and biological circuits

Both anatomical constraints and functional optimization were essential to creating a mapping between this artificial network and the biological one. Our results were made more interpretable by making simplifying assumptions based on the connectome. There are many properties one could measure in a circuit, and comparisons with task-optimized models allow one to evaluate how such properties relate to a specific task or constraint. This study argues that when strong anatomical constraints are included in performance optimized models, there can be a close correspondence between the model units and the analogous individual neurons.

Contributions

OM, MSC, and DAC conceived of the framework and numerical experiments. OM, MSC, and BAB wrote code and ran numerical experiments. OM analyzed models and data. OM and DAC wrote the paper.

Acknowledgements

We thank J. Fitzgerald, N. Kadakia, J. Lafferty, J. Murray, and members of the Clark lab for feedback and illuminating conversations. We thank L. Romero and L. Khazan for their contributions to coding on related projects. The Yale Center for Research Computing provided helpful guidance and research computing infrastructure. DAC and this project were supported by NIH R01EY026555, NSF IOS1558103, a Searle Scholar Award, and a Sloan Fellowship in Neuroscience.

STAR Methods

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for code or data should be directed to and will be fulfilled by the lead contact, Damon A. Clark (damon.clark@yale.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Python and Matlab code to train all models in this paper and generate all figures in this paper is available at http://www.github.com/ClarkLabCode/T4T5TrainingCode. Code is in Matlab (Mathworks, Natick, MA), Python, and several Python libraries [71, 112-114]. The natural image database used in this study has DOI https://doi.org/10.4119/unibi/2689637 and is available at https://pub.uni-bielefeld.de/rc/2689637/2693616.

METHOD DETAILS

Training data

We wanted to train neural networks to predict velocity traces $v(t)$ from simulated visual input signals over space and time. To create velocity traces with the statistical properties similar to fly rotation, we first drew samples from a Gaussian distribution with mean of 0°/s and standard deviation of 100°/s. These samples were placed in a 1-dimensional vector with a sample rate of 100 Hz. To create autocorrelations in the trace, this vector was convolved with an exponential filter $h(t) = K \exp(-t/\tau)$ where $\tau = 0.2/\log 2$ s and K was chosen so that the variance remained unchanged under filtering. This resulted in a velocity trace with an autocorrelation half-life of 200 ms and a standard deviation of 100 °/s (as in the original trace). This trace corresponds to an auto-regressive Gaussian process of order 1, which is a discrete time approximation of an Ornstein-Uhlenbeck process. These scales are comparable to those in walking flies [48, 49]. The final traces contained 101 elements each, corresponding to 1.01 seconds of simulated time.

After creating the velocity traces, we constructed corresponding matrices of simulated photoreceptor activation values. Conceptually, for each 101-element velocity trace, we needed a 3x101 element photoreceptor matrix that corresponds to the activations of the three inputs to our models. In order to efficiently generate and use these 3x101 element matrices, we generated 72x101 element matrices corresponding to a full 360 degrees of photoreceptor activities, spaced 5 degrees apart [51]. These 72 photoreceptors observed natural scenes rotating at the speed specified by the 101-element velocity trace. These matrices can be used in convolution operations to quickly simulate the behavior of many model motion detectors.

To generate these 72x101 matrices, we took a dataset of natural scenes [39] and selected 241 images of natural environments, excluding indoor and architectural scenes. These scenes were panoramic captures of 360x97.5 degrees sampled at around 2.6 pixels per degree. For each velocity trace, we selected a natural scene image at random. We convolved these images with a 5

degree FWHM gaussian filter, approximating the acceptance angle of fly photoreceptors [51]. We converted the velocity trace into a position trace by integrating over time. These positions were used as offsets when converting the images in the spatially filtered dataset from 927x251 pixels to 72x20x101 elements representing the activations of an array of 72x20 photoreceptors at 101 points in time. The 20 rows of photoreceptors were spaced every 5º in elevation. Each row of photoreceptors had an associated set of signal traces $s_{n,t}$ where n represents the azimuthal location and t represents time. Each set of $s_{n,t}$ was treated independently in further processing by duplicating the corresponding velocity traces such that the responses of all rows of photoreceptors could be used to predict the same velocity trace. For each velocity trace and photoreceptor matrix generated in this manner, we also created a paired trace with the entire spatial structure reversed (and negated velocities), in order to ensure that the dataset was balanced with respect to the direction of motion. Finally, the input images were mean subtracted and scaled so that the set of spatially filtered signals $s_{n,t}$ had a mean of zero and a unit variance, computed over all signals in a row and over time. In total, we created 8664 velocity traces and corresponding 72x101 element photoreceptor matrices, divided into a 6346 trace training set and a 2318 trace test set.

To generate the sinusoidal training data (**Fig. 4**), we substituted the natural scenes with sinusoidal gratings with wavelengths chosen from a uniform distribution ranging from 20º to 90º. All other processing steps were identical.

Model definitions

Our models consisted of multiple units whose outputs were summed to generate the model predictions. We defined (+) and (–) versions of each unit type, corresponding to mirror symmetric units that were added and subtracted to generate the final model outputs. To obtain the unit outputs, we filtered signals, s_t, in time by convolving them with filters, f_t, with 30 elements, corresponding to 300 ms in time. We define this convolution as $(\boldsymbol{f} * \boldsymbol{s})_t = \sum_{\tau=0}^{29} f_\tau s_{t-\tau}$.

The mirror symmetric LN units were defined as:

$$u_{k+,t} = \phi\left(\left(\boldsymbol{f}_{k,1} * \boldsymbol{s}_1\right)_t + \left(\boldsymbol{f}_{k,2} * \boldsymbol{s}_2\right)_t + \left(\boldsymbol{f}_{k,3} * \boldsymbol{s}_3\right)_t + b_k\right)$$

$$u_{k-,t} = \phi\left(\left(\boldsymbol{f}_{k,1} * \boldsymbol{s}_3\right)_t + \left(\boldsymbol{f}_{k,2} * \boldsymbol{s}_2\right)_t + \left(\boldsymbol{f}_{k,3} * \boldsymbol{s}_1\right)_t + b_k\right)$$

Where s_i are the input signals, and all parameters ($\boldsymbol{f}_{k,i}$ and b_k) are identical for both units in the pair, and the pairs are indexed by k. The activation function ϕ is everywhere a rectified linear unit (ReLU):

$$\phi(x) = \begin{cases} x \text{ if } x > 0 \\ 0 \text{ if } x \leq 0 \end{cases}$$

The mirror symmetric LNLN units were defined as:

$$u_{k+,t} = \phi\left(w_{k,1}\phi\left(\left(\boldsymbol{f}_{k,1} * \boldsymbol{s}_1\right)_t + b_{k,1}\right) + w_{k,2}\phi\left(\left(\boldsymbol{f}_{k,2} * \boldsymbol{s}_2\right)_t + b_{k,2}\right) + w_{k,3}\phi\left(\left(\boldsymbol{f}_{k,3} * \boldsymbol{s}_3\right)_t + b_{k,3}\right) + b_{k,4}\right)$$

$$u_{k-,t} = \phi\left(w_{k,1}\phi\left(\left(\boldsymbol{f}_{k,1} * \boldsymbol{s}_3\right)_t + b_{k,1}\right) + w_{k,2}\phi\left(\left(\boldsymbol{f}_{k,2} * \boldsymbol{s}_2\right)_t + b_{k,2}\right) + w_{k,3}\phi\left(\left(\boldsymbol{f}_{k,3} * \boldsymbol{s}_1\right)_t + b_{k,3}\right) + b_{k,4}\right)$$

As above, all the parameters are the same for both units in the pair. Each $w_{k,i}$ is a scalar free parameter.

For the synaptic nonlinearity, we followed previous work to create a nonlinearity that treats input LN lines as conductances in the membrane of a postsynaptic cell, and then computes the steady state voltage, given weighting parameters that are equivalent to reversal potentials in a real cell [19, 29, 59, 115]. This approximates the membrane time constants as being much smaller than typical variations in inputs [19]:

$$S_{k,t}(s_1, s_2, s_3) = \frac{w_{k,1}\phi\left(\left(\boldsymbol{f}_{k,1} * \boldsymbol{s}_1\right)_t + b_{k,1}\right) + w_{k,2}\phi\left(\left(\boldsymbol{f}_{k,2} * \boldsymbol{s}_2\right)_t + b_{k,2}\right) + w_{k,3}\phi\left(\left(\boldsymbol{f}_{k,3} * \boldsymbol{s}_3\right)_t + b_{k,3}\right)}{1 + \phi\left(\left(\boldsymbol{f}_{k,1} * \boldsymbol{s}_1\right)_t + b_{k,1}\right) + \phi\left(\left(\boldsymbol{f}_{k,2} * \boldsymbol{s}_2\right)_t + b_{k,2}\right) + \phi\left(\left(\boldsymbol{f}_{k,3} * \boldsymbol{s}_3\right)_t + b_{k,3}\right)}$$

We then defined our two units as:

$$u_{k+,t} = \phi(S_{k,t}(s_1, s_2, s_3) + b_{k,4})$$
$$u_{k-,t} = \phi(S_{k,t}(s_3, s_2, s_1) + b_{k,4})$$

Where the second activation function could correspond to a calcium nonlinearity acting on the membrane voltage [17, 29, 59].

Our model outputs, R_t, weighted two pairs of units by the scalars a_k as follows:

$$R_t = a_1(u_{1+,t} - u_{1-,t}) + a_2(u_{2+,t} - u_{2-,t})$$

This arrangement of units within models gave us three models: the LN model using two pairs of LN units, the LNLN model using two pairs of LNLN units, and the synaptic nonlinearity model using two pairs of synaptic nonlinearity units. When we examined additional units in Figure 4, we added the additional pairs with new weight parameters.

Noise in the models

We added noise to the models at two stages. First, we added front end noise by adding random samples from a zero-mean Gaussian distribution to each element in the matrices $s_{n,t}$. Since the standard deviation of these matrices was unity, the standard deviation of the added noise controlled the relative amplitude of signal and noise. Second, we multiplied the output of each model unit $u_{k\pm,t}$ by random draws from a lognormal distribution for each point in time. The lognormal distribution was chosen such that its mean was 1, and its standard deviation determined the relative size of the output noise. The output noise was chosen to be multiplicative rather than additive so that the models could not escape the noise by producing very large unit outputs and then rescale them with the model weights after the addition of noise. The standard deviations for both these sources varied according to the experiment.

Training protocols

All models were trained in Python using TensorFlow [71]. Due to the convolution operations employed by our neural network models, for each 72×101× [batch size] input to our model, the output was a set of 70×72×[batch size] velocities corresponding to a set of 72×[batch size] true velocities. We duplicated these true velocities to create tensors of 70×72×[batch size]. To train the models, we chose the loss function to be the mean squared error between the true input velocity and the individual model outputs, R_t (not averaged over space). In the case of models

trained to predict the direction of motion (**Fig. 4B**), we converted the velocity trace into a binary direction trace, and the loss function became the cross-entropy of the true direction with a sigmoid function acting on the model output, R_t. The primary analyzed models each had two unit types with three filters each (180 parameters). In the LN model, each of the two LN unit pairs had one additional bias term associated with the threshold nonlinearity. In the LNLN model, each of the LNLN units had four additional bias terms associated with the four threshold nonlinearities, and three additional weight parameters for the three rectified input arms. For the synaptic nonlinearity model, the additional parameters were the same as for the LNLN model.

To train our models, we used the Adam optimizer with an initial learning rate of 0.03 and learning rate decay such that the final learning rate was 0.0027. We trained for 1000 epochs with a batch size of 128. For each set of model hyperparameters (model type, direction prediction, input and output noise, etc.), we trained 50 instantiations of that model. Each instantiation had a different initial set of weights drawn from a "Glorot" distribution [116]. For analysis, we chose the 9 highest performing models for each set of hyperparameters as evaluated by the coefficient of determination in the training dataset. Multiple training runs from the same initialization tended to arrive at the same solution, suggesting that in our training regime, the stochasticity of initialization affects solutions more than stochasticity in training protocol.

Stimuli for comparison with biological data

To compare model responses to those measured in fly visual circuits, we created several visual stimuli to present to our models. First, to obtain the effective linear filters of the inputs to the synaptic model, we stimulated the model with independent, Gaussian noise to each input, with zero mean and unit variance, then extracted the kernels from the unit output, using standard methods [117].

To make comparisons with responses to edges (**Fig. 1C**), we created light and dark edges expanding over time so that the image, m, over space and time, was:

$$m(x,t) = \pm 2\left(H(x \pm vt) - \frac{1}{2}\right)$$

where we used all combinations of ± to make light and dark edges moving in both directions. The stimulus velocity v was 30º/s. These images were spatially filtered to create the input signals $s_{n,t}$.

To compare responses to different stationary edges (**Fig. 1D**), we created a light and dark square wave with an image over space of:

$$m(x) = \text{sign}\left(\sin\left(\frac{2\pi x}{\lambda}\right)\right)$$

where the wavelength λ was chosen to be 80º. These images were spatially filtered to create the input signals $s_{n,t}$.

To compare responses to sinusoids moving the preferred and null directions and to their sum (**Fig. 1E**), we created images as follows:

$$m_{PD}(x,t) = \frac{1}{2}\sin(kx - \omega t)$$

$$m_{ND}(x,t) = \frac{1}{2}\sin(kx + \omega t)$$

$$m_{PD+ND}(x,t) = \frac{1}{2}\sin(kx - \omega t) + \frac{1}{2}\sin(kx + \omega t)$$

The spatial frequency was chosen to be $k = 2\pi/60$ deg^{-1} and $\omega = 2\pi$ s^{-1}. In sweeps of spatial and temporal frequency (**Supp. Fig. 1**), the spatial and temporal frequencies were chosen as labeled. When signal strength was swept, the sinusoid amplitude was changed as labeled. As with the other stimuli, these images were spatially filtered.

To compare the degree of coactivation (**Fig. 1F**), we used the natural scenes test (holdout) dataset described above.

In all comparisons of the model with data, we set the noise values in the model to 0, regardless of training regime, unless otherwise noted. Setting the input noise to 0 is the equivalent of having a bright stimulus with high signal to noise, as is typical of experiments. Setting the output noise to 0 is the equivalent of averaging over many trials of the same stimulus (since the multiplicative noise has expected value of 1). Averaging over trials was typical in the comparison data (**Fig. 1**).

Metrics

We summarized properties of models with several metrics (Figs. 5 and 6). Fraction of variance explained was evaluated using the coefficient of determination in the holdout (test) dataset; it could be negative if the model performed worse than uniformly predicting the average velocity in the dataset. We evaluated the timescale of the learned filters by calculating the center of mass (or expected value) of the absolute value of the filters.

We also evaluated the edge selectivity indices (ESIs) and the direction selectivity indices (DSIs) of the models by simulating the responses to the moving edges. We simulated a light edge and a dark edge moving in the positive and negative direction, each as a separate trace. Then, for each unit in the model, we calculated the maximum of the absolute value of the response. For each unit, we averaged the PD and ND max responses across the dark and light edges, and separately averaged the light and dark max responses across the PD and ND edges. Then, for each unit we compute the selectivity index; $ESI = \frac{R_{light} - R_{dark}}{R_{light} + R_{dark}}$, where R_{light} is the average of the max response to light edges in both preferred and null directions and R_{dark} is the average of the max response to dark edges in both preferred and null directions. Similarly, $DSI = \frac{R_{PD} - R_{ND}}{R_{PD} + R_{ND}}$ where R_{PD} is the average of the max response to light and dark edges in the preferred direction while R_{ND} is the average of the max response to light and dark edges in the null direction. Finally, we computed selectivity index for the model as a whole by taking the mean of the absolute values of the selectivity indices of the individual units.

To summarize the static edge activation as a scalar value for each model, we stimulated the model units with static edges of both polarities centered on the central receptor and found the steady state response. We report the model response as the average of all unit responses to both edges.

In order to measure opponent suppression, we generated a moving sinusoidal grating dataset with PD, ND, and PD+ND stimuli, as described above. We then calculated the space- and time-averaged responses of the individual units our models to these three stimuli. We defined an opponency index of these units as $OI = \frac{R_{PD} - R_{CP}}{R_{PD} + R_{CP}}$ where R_{PD} and R_{CP} are the time-averaged unit response to the preferred direction sinusoid grating and the response to the counterphase grating respectively. We then defined the model's opponency index as the average of the opponency indices of its units.

Finally, we evaluated the sparsity of the coactivation of the model units in response to the test set, naturalistic stimuli, with no noise added. Coactivation between units m and n was defined as $C_{nm} = \frac{1}{T}\sum_{t=1}^{T} \frac{u_{n,t}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} {u_{n,t}}^2}} \frac{u_{m,t}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} {u_{m,t}}^2}}$, where $u_{n,t}$ is the response trace of unit n at time t and $u_{m,t}$ is defined similarly; T is the length of the trace in time. Averages were taken over the entire test dataset. We defined a sparsity index as the root mean square difference between the coactivation matrix of the model units and the identity matrix and then rescaled it so that a sparsity index of 1 corresponds to the identity matrix and a sparsity of 0 corresponds to all units being 100% coactive.

Figures

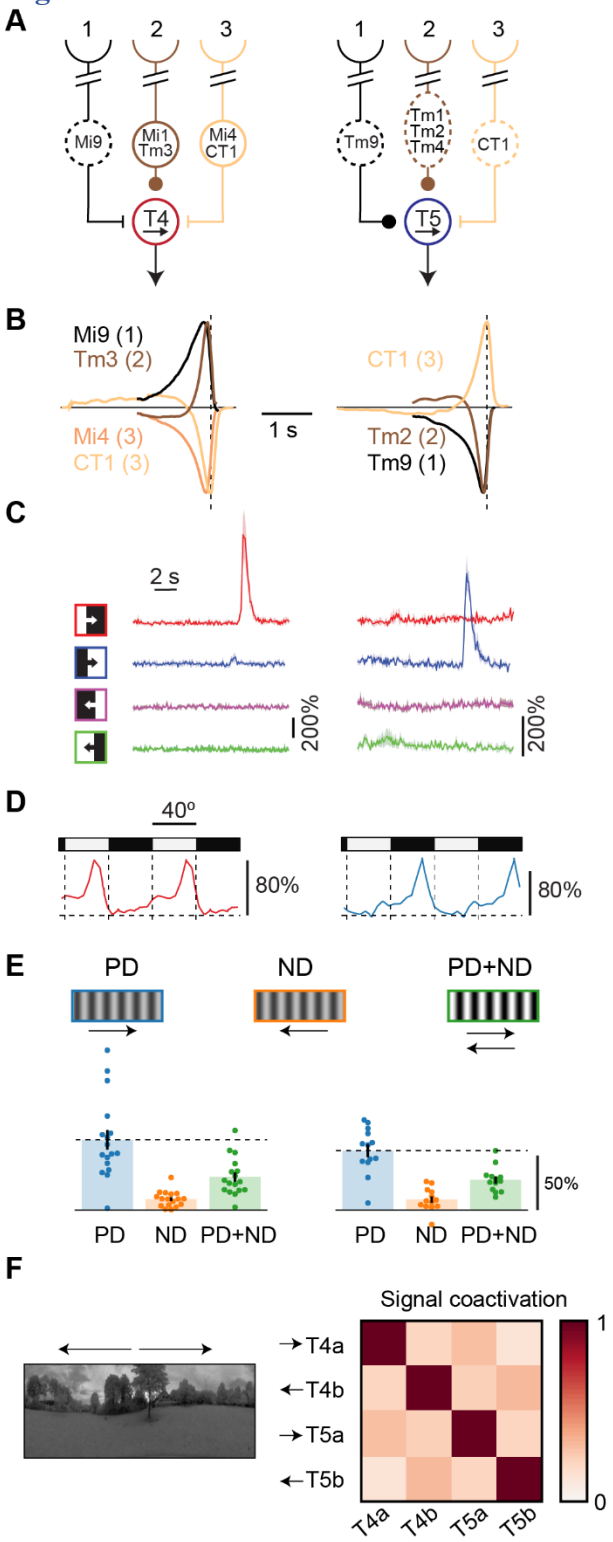


Figure 1. Non-canonical measured properties of primary motion detecting neurons in *Drosophila*.

A) Connectivity schematic of the three spatially separated inputs to T4 and T5 neurons, two parallel, primary motion detectors in *Drosophila*'s visual system. Dashed lines indicate

that a cell is in the OFF pathway. Round synapses indicate excitatory connections, while bars indicate inhibitory synapses.

B) For each cell immediately upstream of T4 and T5, we plot the linear model prediction of the calcium response to an impulse of light signed by their input to T4 and T5. Neurons in position 2 show fast dynamics compared to the neurons in flanking positions. Inputs from position 3 have the opposite influence on T4 and T5 from neurons in positions 1 and 2. Data from [24, 37].

C) Traces of T4 and T5 calcium responses to light and dark edges moving in the preferred (rightward) and null (leftward) directions. Data from [30].

D) Mean calcium responses of T4 and T5 neurons to a stationary square wave stimulus as a function of position, showing preferential responses at edges of specific polarity. Data from [30].

E) Mean calcium responses of T4 and T5 neurons to preferred direction (PD) and null direction (ND) drifting sinusoid gratings, as well as to their sum (PD+ND). The addition of null direction motion suppresses calcium responses in T4 and T5, a form of opponent suppression in primary motion detectors. Data from [29].

F) T4 and T5 calcium signals in response to naturalistic stimuli tend to be non-coactive. Arrows indicate the direction selectivity of the different neuron classes. Data from [28].

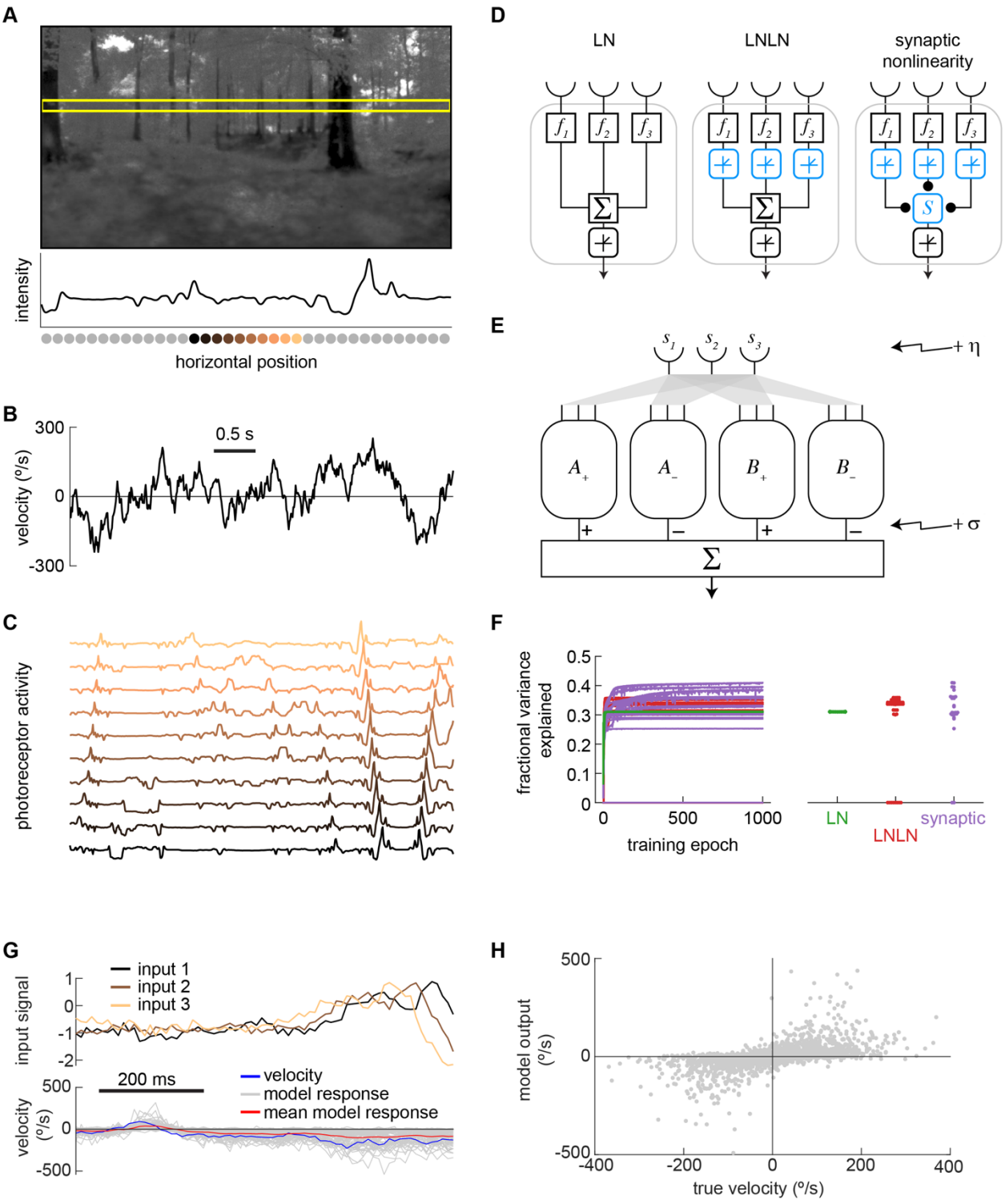


Figure 2. Models predicted velocities of naturalistic training data.

A) Panoramic natural scene from database [39]. Horizontal yellow box shows a 1-dimensional cut through the scene. The luminance trace of that cut is shown below the image, with the positions of simulated photoreceptors below the x-axis.

B) Dynamic velocities traces were drawn from a Gaussian distribution with a correlation time of 200 ms (see Methods).

C) Scenes were translated at the assigned velocities in order to generate a trace of inputs that mirrored the ommatidial inputs of a fly (see Methods). Each trace represents the activity of a photoreceptor located at the position of the photoreceptor in (A) with matching color.
D) Three different shallow network unit types were tested: a linear-nonlinear unit (LN), an LNLN unit, and a unit combining inputs using a biophysical nonlinearity (see Methods).
E) In the models, two units were each paired with a mirror symmetric version of themselves (A_+ with A_-, B_+ with B_-), and signals from the units were subtracted. A and B units had the same architecture but were trained with independent weights. The model output was the sum of these differences. Noise was added at the front-end of the model (η) and at the back end (σ, see Methods).
F) Models containing the three different unit types were trained to predict the scene velocity from ommatidial signal traces. The training converged (*left*) and the fully-trained models predicted 30-40% of the variance in the velocity (*right*). These traces show results for training with $\eta = \sigma = 1/8$.
G) Example traces of inputs and outputs of an LN model trained as in (F), as compared to the true input velocity (*blue*). Different model outputs (*gray*) are for different spatial locations in images, with the same velocity trace. The mean value of the model responses is plotted in red.
H) Scatter plot of individual instantaneous model outputs against the true velocity.

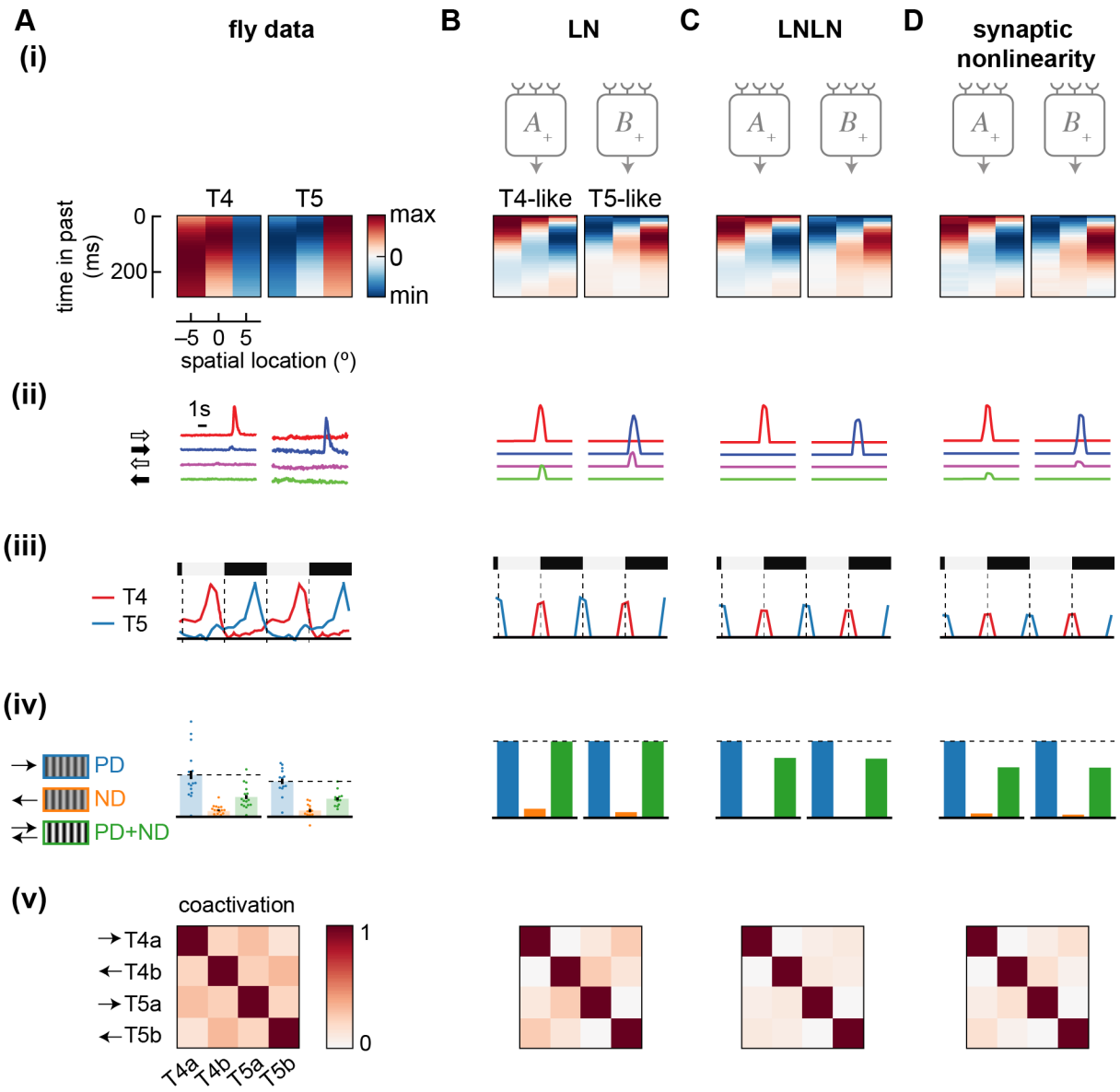
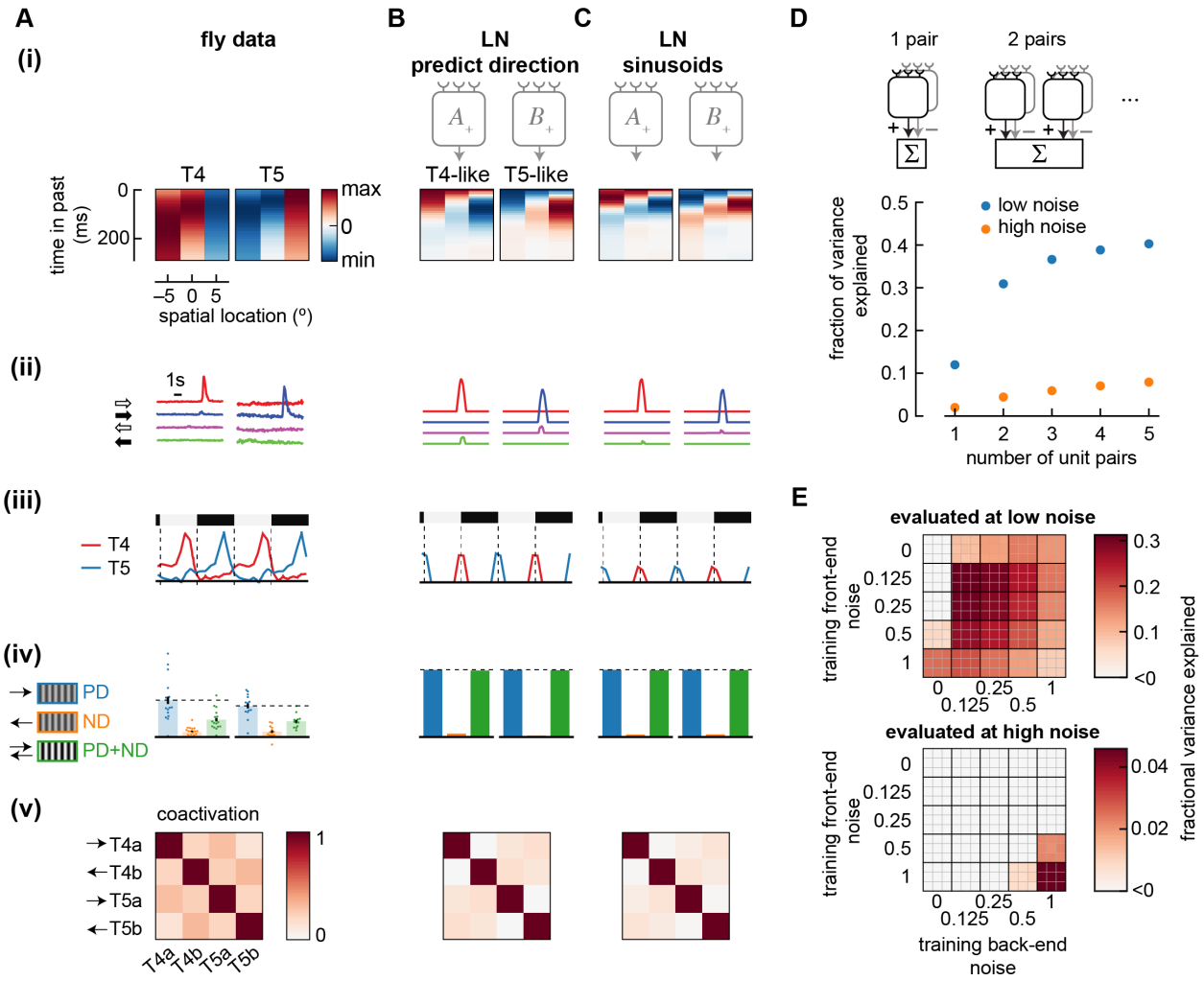


Figure 3. Models trained to predict naturalistic velocities possess many properties of the biological circuit. Data shown includes: (i) Spatiotemporal receptive fields composed of time traces of the filters of the 3 spatially separated inputs to T4 and T5 or to T4- and T5-like units. Each input filter is normalized. (ii) Responses to light and dark edges moving left and right. (iii) Responses to stationary square waves. For model responses, the full vertical extent of the dashed lines is the amplitude of responses to the preferred moving edge in (ii). (iv) Relative responses to preferred and null direction sinusoids, and their sum. (v) Coactivation of units in response to naturalistic stimuli.

A) Data from the fly, as in Figure 1.
B) As in (A), but for a trained LN model.
C) As in (A), but for a trained LNLN model.
D) As in (A), but for a trained synaptic nonlinearity model. All three models were trained with noise values of $\eta = \sigma = 1$.



930
931 **Figure 4. Effects of model loss function, training, architecture, and noise.**

932 A) Summary of properties measured in T4 and T5 (from **Figure 1**). Data shows false color
933 time traces of 3 spatially separated input filters (i), responses to light and dark edges
934 moving left and right (ii), responses to stationary square waves (iii), responses to
935 preferred and null direction sinusoids, and their sum (iv), and coactivation of units in
936 response to naturalistic stimuli (v).

937 B) As in (A), but showing the results of an LN model with an alternate loss function, in
938 which it was trained to predict *direction* of motion rather than predict *velocity* of motion.
939 Compare with **Fig. 3B**. Model was trained with noise of $\eta = \sigma = 1$.

940 C) As in (A), but showing the results of an LN model trained on sinusoidal gratings instead
941 of natural scenes. Compare with **Fig. 3B**. Model was trained with noise of $\eta = \sigma = 1$.

942 D) The number of mirror-symmetric, subtracted unit pairs was swept from one to five (*top*),
943 while measuring the fraction of variance explained for LN models trained and evaluated
944 in high and low noise conditions. All unit pairs received inputs from the same 3 spatial
945 locations. Throughout the rest of this study, two pairs were used.

946 E) Fraction of variance explained by models trained at a variety of front- and back-end noise
947 levels, then tested at low noise (*top*) and high noise (*bottom*). The top 9 models are shown
948 as a 3x3 grid at the coordinate of a specific parameter set. Low noise evaluation used
949 parameters $\eta = \sigma = 0.125$; high noise evaluation used parameters $\eta = \sigma = 1$.

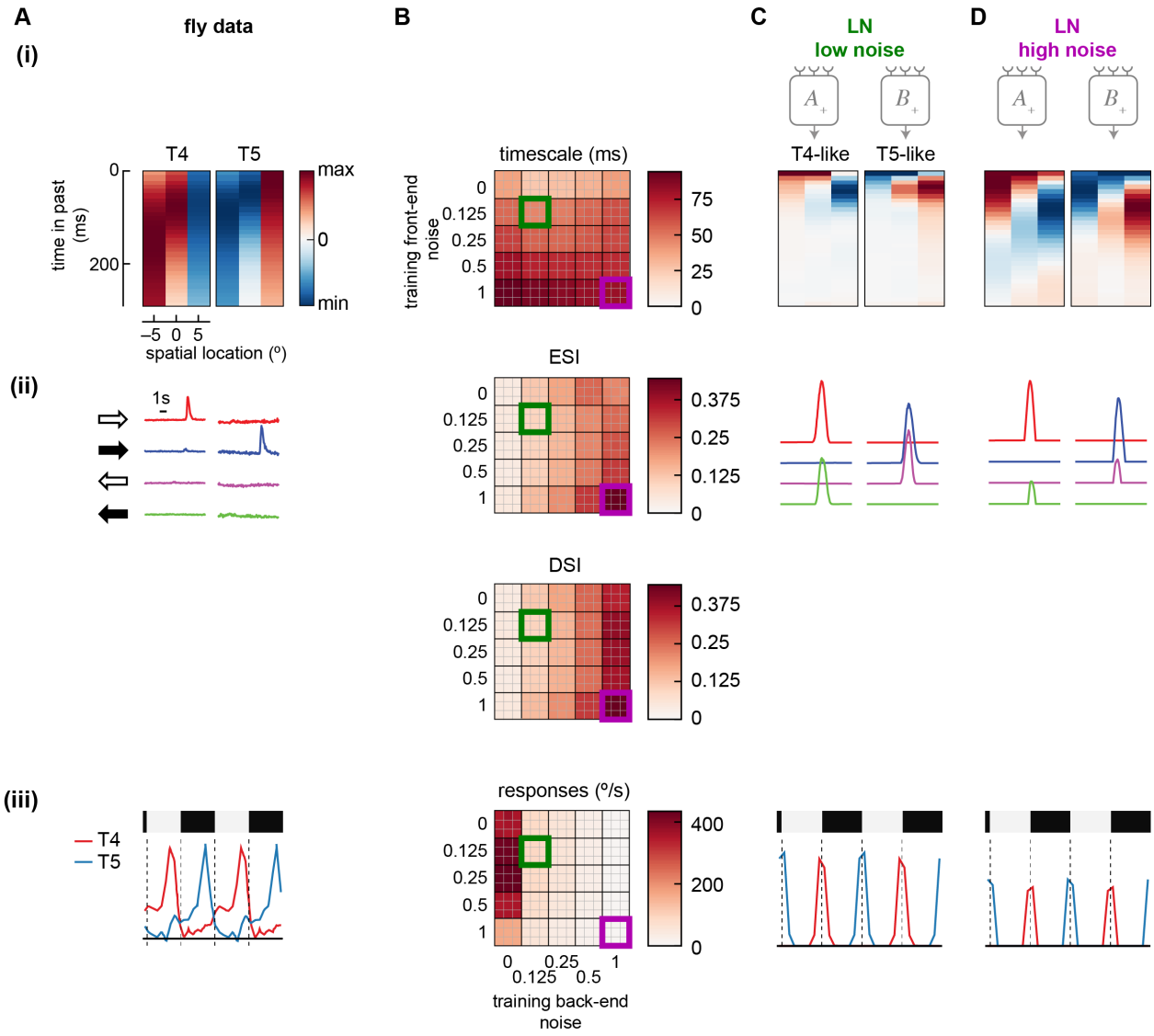


Figure 5. High training noise yields slower filter dynamics and stronger selectivity to moving edges.

A) Summary of properties measured in T4 and T5 (from **Fig. 1**). Data shows false color time traces of 3 spatially separated input filters (i), responses to light and dark edges moving left and right (ii), and responses to stationary square waves (iii).

B) Summary responses of models trained with different levels of front-end and back-end noise. Top 9 performing models of 50 trained are shown for each condition, measuring the center-of-mass of the filters (i), the ESI and DSI of the light and dark moving edge responses of each unit (ii, *top* and *bottom*), and the responses to stationary square waves of the units (iii).

C) Example traces of a low-noise trained model (green square in (B)). Shown are filters for each unit (i), traces of responses to left and right moving light and dark edges (ii), and responses to stationary square wave stimuli (iii).

D) As in (C) but with the high-noise trained model (purple square in (B)).

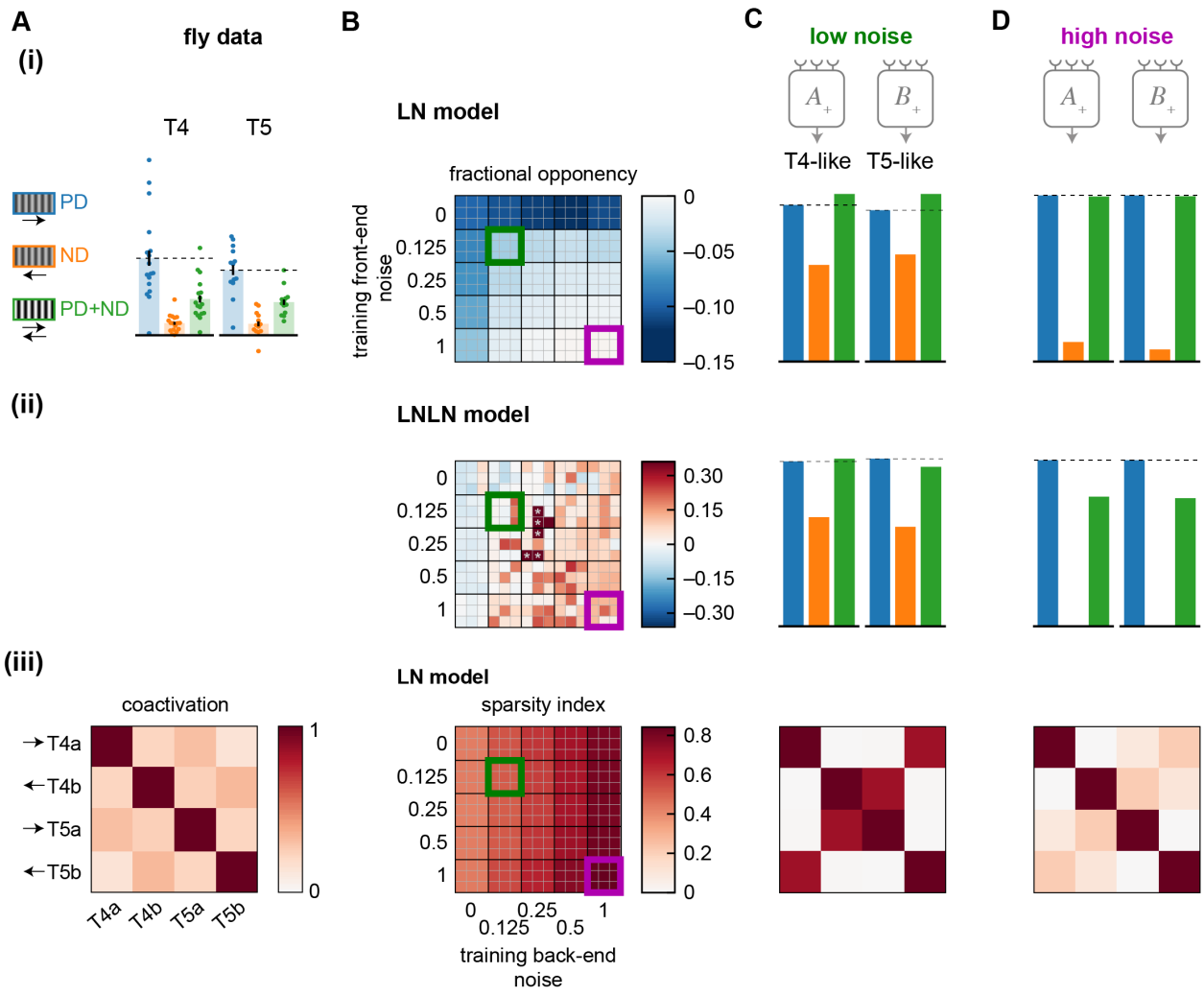


Figure 6. High training noise yields strong opponency and channel decorrelation.

A) Summary of properties measured in T4 and T5 (from Figure 1). Data shows responses to preferred and null direction sinusoids (PD, ND) and their sum (PD+ND) (i), and coactivation of units in response to naturalistic stimuli (iii).

B) Summary responses of models trained with different levels of front-end and back-end noise. The top 9 performing models of 50 trained are shown for each condition. Data shown is the opponency of LN models (i) and LNLN models (ii), where asterisks denote models with opponency near 1, out of the false color range. The sparsity index is shown for the LN model units in response to naturalistic stimuli (iii). The sparsity index is 1 when the coactivation matrix is the identity matrix and is 0 when all elements in the matrix are 1.

C) Example responses from a low-noise training protocol (green box in (B)). Opponency is shown for the LN model (i) and LNLN model (ii), while a coactivation matrix is shown for an LN model responding to naturalistic stimuli (iii).

D) As in (C) but for a high-noise training protocol (purple box in (B)).

Citations

1. Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619-8624.
2. Haesemeyer, M., Schier, A.F., and Engert, F. (2019). Convergent temperature representations in artificial and biological neural networks. Neuron *103*, 1123-1134. e1126.
3. Yamins, D.L., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356-365.
4. Yamins, D. (2020). An Optimization-Based Approach to Understanding Sensory Systems. The Cognitive Neurosciences *4*, 381.
5. Hasson, U., Nastase, S.A., and Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. Neuron *105*, 416-434.
6. Schnell, B., Raghu, S.V., Nern, A., and Borst, A. (2012). Columnar cells necessary for motion responses of wide-field visual interneurons in Drosophila. J. Comp. Physiol. A *198*, 389-395.
7. Maisak, M.S., Haag, J., Ammer, G., Serbe, E., Meier, M., Leonhardt, A., Schilling, T., Bahl, A., Rubin, G.M., Nern, A., et al. (2013). A directional tuning map of Drosophila elementary motion detectors. Nature *500*, 212-216.
8. Schilling, T., and Borst, A. (2015). Local motion detectors are required for the computation of expansion flow-fields. Biology open, bio. 012690.
9. Creamer, M.S., Mano, O., and Clark, D.A. (2018). Visual Control of Walking Speed in Drosophila. Neuron *100*, 1460-1473.
10. Shinomiya, K., Huang, G., Lu, Z., Parag, T., Xu, C.S., Aniceto, R., Ansari, N., Cheatham, N., Lauchie, S., Neace, E., et al. (2019). Comparisons between the ON-and OFF-edge motion pathways in the Drosophila brain. eLife *8*, e40025.
11. Takemura, S.-y., Nern, A., Chklovskii, D.B., Scheffer, L.K., Rubin, G.M., and Meinertzhagen, I.A. (2017). The comprehensive connectome of a neural substrate for 'ON'motion detection in Drosophila. Elife *6*.
12. Meinertzhagen, I., and O'Neil, S. (1991). Synaptic organization of columnar elements in the lamina of the wild type in Drosophila melanogaster. The Journal of comparative neurology *305*, 232-263.
13. Takemura, S.-y., Bharioke, A., Lu, Z., Nern, A., Vitaladevuni, S., Rivlin, P.K., Katz, W.T., Olbris, D.J., Plaza, S.M., Winston, P., et al. (2013). A visual motion detection circuit suggested by Drosophila connectomics. Nature *500*, 175-181.
14. Fisher, Y.E., Leong, J.C., Sporar, K., Ketkar, M.D., Gohl, D.M., Clandinin, T.R., and Silies, M. (2015). A class of visual neurons with wide-field properties is required for local motion detection. Curr. Biol. *25*, 3178-3189.
15. Yang, H.H., St-Pierre, F., Sun, X., Ding, X., Lin, M.Z., and Clandinin, T.R. (2016). Subcellular imaging of voltage and calcium signals reveals neural processing in vivo. Cell *166*, 245–257.
16. Leong, J.C.S., Esch, J.J., Poole, B., Ganguli, S., and Clandinin, T.R. (2016). Direction selectivity in Drosophila emerges from preferred-direction enhancement and null-direction suppression. J. Neurosci. *36*, 8078-8092.

17. Wienecke, C.F., Leong, J.C., and Clandinin, T.R. (2018). Linear Summation Underlies Direction Selectivity in Drosophila. Neuron.
18. Strother, J.A., Wu, S.-T., Wong, A.M., Nern, A., Rogers, E.M., Le, J.Q., Rubin, G.M., and Reiser, M.B. (2017). The emergence of directional selectivity in the visual motion pathway of Drosophila. Neuron *94*, 168-182. e110.
19. Gruntman, E., Romani, S., and Reiser, M.B. (2018). Simple integration of fast excitation and offset, delayed inhibition computes directional selectivity in Drosophila. Nat. Neurosci., 1.
20. Strother, J.A., Wu, S.-T., Rogers, E.M., Eliason, J.L., Wong, A.M., Nern, A., and Reiser, M.B. (2018). Behavioral state modulates the ON visual motion pathway of Drosophila. Proc. Natl. Acad. Sci. USA *115*, E102-E111.
21. Gruntman, E., Romani, S., and Reiser, M.B. (2019). The computation of directional selectivity in the Drosophila OFF motion pathway. eLife *8*.
22. Serbe, E., Meier, M., Leonhardt, A., and Borst, A. (2016). Comprehensive characterization of the major presynaptic elements to the Drosophila OFF motion detector. Neuron *89*, 829-841.
23. Haag, J., Arenz, A., Serbe, E., Gabbiani, F., and Borst, A. (2016). Complementary mechanisms create direction selectivity in the fly. Elife *5*.
24. Arenz, A., Drews, M.S., Richter, F.G., Ammer, G., and Borst, A. (2017). The temporal tuning of the Drosophila motion detectors is determined by the dynamics of their input elements. Curr. Biol. *27*, 929-944.
25. Haag, J., Mishra, A., and Borst, A. (2017). A common directional tuning mechanism of Drosophila motion-sensing neurons in the ON and in the OFF pathway. Elife *6*, e29044.
26. Clark, D.A., Bursztyn, L., Horowitz, M.A., Schnitzer, M.J., and Clandinin, T.R. (2011). Defining the computational structure of the motion detector in Drosophila. Neuron *70*, 1165-1177.
27. Salazar-Gatzimas, E., Chen, J., Creamer, M.S., Mano, O., Mandel, H.B., Matulis, C.A., Pottackal, J., and Clark, D.A. (2016). Direct measurement of correlation responses in Drosophila elementary motion detectors reveals fast timescale tuning. Neuron *92*, 227-239.
28. Salazar-Gatzimas, E., Agrochao, M., Fitzgerald, J.E., and Clark, D.A. (2018). The Neuronal Basis of an Illusory Motion Percept Is Explained by Decorrelation of Parallel Motion Pathways. Curr. Biol. *28*, 3748-3762. e3748.
29. Badwan, B.A., Creamer, M.S., Zavatone-Veth, J.A., and Clark, D.A. (2019). Dynamic nonlinearities enable direction opponency in Drosophila elementary motion detectors. Nat. Neurosci. *22*, 1318-1326.
30. Agrochao, M., Tanaka, R., Salazar-Gatzimas, E., and Clark, D.A. (2020). Mechanism for analogous illusory motion perception in flies and humans. Proc. Natl. Acad. Sci. *117*, 23044-23053.
31. Joesch, M., Schnell, B., Raghu, S., Reiff, D., and Borst, A. (2010). ON and OFF pathways in Drosophila motion vision. Nature *468*, 300-304.
32. Hassenstein, B., and Reichardt, W. (1956). Systemtheoretische Analyse der Zeit-, Reihenfolgen-und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers Chlorophanus. Zeits. Naturforsch. *11*, 513–524.
33. Adelson, E., and Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. JOSA A *2*, 284-299.

34. Potters, M., and Bialek, W. (1994). Statistical mechanics and visual signal processing. J. Physique *4*, 1755-1775.
35. Fitzgerald, J.E., Katsov, A.Y., Clandinin, T.R., and Schnitzer, M.J. (2011). Symmetries in stimulus statistics shape the form of visual motion estimators. Proc. Natl. Acad. Sci. USA *108*, 12909-12914.
36. Borst, A., and Egelhaaf, M. (1989). Principles of visual motion detection. Trends Neurosci. *12*, 297-306.
37. Meier, M., and Borst, A. (2019). Extreme Compartmentalization in a Drosophila Amacrine Cell. Curr. Biol. *29*, 1545-1550. e1542.
38. Fitzgerald, J.E., and Clark, D.A. (2015). Nonlinear circuits for naturalistic visual motion estimation. eLife, e09123.
39. Meyer, H.G., Schwegmann, A., Lindemann, J.P., and Egelhaaf, M. (2014). Panoramic high dynamic range images in diverse environments. B. University, ed.
40. Götz, K. (1964). Optomotorische untersuchung des visuellen systems einiger augenmutanten der fruchtfliege Drosophila. Biol. Cybern. *2*, 77-92.
41. Götz, K., and Wenking, H. (1973). Visual control of locomotion in the walking fruitfly Drosophila. J. Comp. Physiol. A *85*, 235-266.
42. Cafaro, J., Zylberberg, J., and Field, G.D. (2020). Global motion processing by populations of direction-selective retinal ganglion cells. J. Neurosci. *40*, 5807-5819.
43. Chen, J., Mandel, H.B., Fitzgerald, J.E., and Clark, D.A. (2019). Asymmetric ON-OFF processing of visual motion cancels variability induced by the structure of natural scenes. eLife *8*, e47579.
44. Leonhardt, A., Ammer, G., Meier, M., Serbe, E., Bahl, A., and Borst, A. (2016). Asymmetry of Drosophila ON and OFF motion detectors enhances real-world velocity estimation. Nat. Neurosci. *19*, 706–715.
45. Shoemaker, P.A., O'Carroll, D.C., and Straw, A.D. (2005). Velocity constancy and models for wide-field visual motion detection in insects. Biol. Cybern. *93*, 275-287.
46. Straw, A.D., Rainsford, T., and O'Carroll, D.C. (2008). Contrast sensitivity of insect motion detectors to natural images. J. Vis. *8*, 32-32.
47. Dror, R.O., O'Carroll, D.C., and Laughlin, S.B. (2001). Accuracy of velocity estimation by Reichardt correlators. JOSA A *18*, 241-252.
48. DeAngelis, B.D., Zavatone-Veth, J.A., and Clark, D.A. (2019). The manifold structure of limb coordination in walking Drosophila. eLife *8*, e46409.
49. Katsov, A.Y., Freifeld, L., Horowitz, M.A., Kuehn, S., and Clandinin, T.R. (2017). Dynamic structure of locomotor behavior in walking fruit flies. eLife *6*, e26410.
50. Behnia, R., Clark, D.A., Carter, A.G., Clandinin, T.R., and Desplan, C. (2014). Processing properties of ON and OFF pathways for Drosophila motion detection. Nature *512*, 427-430.
51. Stavenga, D. (2003). Angular and spectral sensitivity of fly photoreceptors. II. Dependence on facet lens F-number and rhabdomere type in Drosophila. Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology *189*, 189-202.
52. Fransen, J.W., and Borghuis, B.G. (2017). Temporally Diverse Excitation Generates Direction-Selective Responses in ON-and OFF-Type Retinal Starburst Amacrine Cells. Cell Rep. *18*, 1356-1365.

53. Kim, J.S., Greene, M.J., Zlateski, A., Lee, K., Richardson, M., Turaga, S.C., Purcaro, M., Balkam, M., Robinson, A., and Behabadi, B.F. (2014). Space-time wiring specificity supports direction selectivity in the retina. Nature *509*, 331-336.
54. Jagadeesh, B., Wheat, H.S., and Ferster, D. (1993). Linearity of summation of synaptic potentials underlying direction selectivity in simple cells of the cat visual cortex. Science *262*, 1901-1904.
55. Rust, N.C., Schwartz, O., Movshon, J.A., and Simoncelli, E.P. (2005). Spatiotemporal elements of macaque v1 receptive fields. Neuron *46*, 945-956.
56. Niell, C.M., and Stryker, M.P. (2008). Highly selective receptive fields in mouse visual cortex. J. Neurosci. *28*, 7520-7536.
57. Heeger, D.J. (1992). Half-squaring in responses of cat striate cells. Vis. Neurosci. *9*, 427-443.
58. Strother, J.A., Nern, A., and Reiser, M.B. (2014). Direct observation of ON and OFF pathways in the Drosophila visual system. Curr. Biol. *24*, 976-983.
59. Zavatone-Veth, J.A., Badwan, B., and Clark, D.A. (2020). A minimal synaptic model for direction selective neurons in Drosophila. J. Vis. *20*.
60. Borst, A. (2018). A biophysical mechanism for preferred direction enhancement in fly motion vision. PLoS Comp. Biol. *14*, e1006240.
61. Borst, A., Egelhaaf, M., and Haag, J. (1995). Mechanisms of dendritic integration underlying gain control in fly motion-sensitive interneurons. J. Comput. Neurosci. *2*, 5-18.
62. Mauss, A.S., Pankova, K., Arenz, A., Nern, A., Rubin, G.M., and Borst, A. (2015). Neural circuit to integrate opposing motions in the visual field. Cell *162*, 351-362.
63. Joesch, M., Plett, J., Borst, A., and Reiff, D. (2008). Response properties of motion-sensitive visual interneurons in the lobula plate of Drosophila melanogaster. Curr. Biol. *18*, 368-374.
64. Srinivasan, M., Laughlin, S., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. Proc. R. Soc. Lond., Ser. B: Biol. Sci. *216*, 427.
65. Van Hateren, J. (1992). Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation. J. Comp. Physiol. A *171*, 157-170.
66. Matulis, C.A., Chen, J., Gonzalez-Suarez, A., Behnia, R., and Clark, D.A. (2020). Heterogeneous temporal contrast adaptation in Drosophila direction-selective circuits. Curr. Biol.
67. Drews, M.S., Leonhardt, A., Pirogova, N., Richter, F.G., Schuetzenberger, A., Braun, L., Serbe, E., and Borst, A. (2020). Dynamic Signal Compression for Robust Motion Vision in Flies. Curr. Biol.
68. Juusola, M., Uusitalo, R., and Weckström, M. (1995). Transfer of graded potentials at the photoreceptor-interneuron synapse. J. Gen. Physiol. *105*, 117.
69. Juusola, M., and Hardie, R.C. (2001). Light Adaptation in Drosophila Photoreceptors I. Response Dynamics and Signaling Efficiency at 25° C. J. Gen. Physiol. *117*, 3-25.
70. Faisal, A.A., Selen, L.P., and Wolpert, D.M. (2008). Noise in the nervous system. Nat. Rev. Neurosci. *9*, 292-303.
71. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., and Isard, M. (2016). Tensorflow: A system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16). pp. 265-283.

72. Chen, T.-W., Wardill, T.J., Sun, Y., Pulver, S.R., Renninger, S.L., Baohan, A., Schreiter, E.R., Kerr, R.A., Orger, M.B., and Jayaraman, V. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature *499*, 295-300.
73. Schnell, B., Joesch, M., Forstner, F., Raghu, S.V., Otsuna, H., Ito, K., Borst, A., and Reiff, D.F. (2010). Processing of horizontal optic flow in three visual interneurons of the Drosophila brain. J. Neurophysiol. *103*, 1646-1657.
74. Maddess, T., and Laughlin, S.B. (1985). Adaptation of the motion-sensitive neuron H1 is generated locally and governed by contrast frequency. Proceedings of the Royal society of London. Series B. Biological sciences *225*, 251-275.
75. Haag, J., Denk, W., and Borst, A. (2004). Fly motion vision is based on Reichardt detectors regardless of the signal-to-noise ratio. Proc. Natl. Acad. Sci. USA *101*, 16333.
76. Clark, D.A., Fitzgerald, J.E., Ales, J.M., Gohl, D.M., Silies, M., Norcia, A.M., and Clandinin, T.R. (2014). Flies and humans share a motion estimation strategy that exploits natural scene statistics. Nat. Neurosci. *17*, 296-303.
77. Dickinson, M.H., Farley, C.T., Full, R.J., Koehl, M., Kram, R., and Lehman, S. (2000). How animals move: an integrative view. Science *288*, 100-106.
78. Lindemann, J.P., Kern, R., Van Hateren, J., Ritter, H., and Egelhaaf, M. (2005). On the computations analyzing natural optic flow: quantitative model analysis of the blowfly motion vision pathway. J. Neurosci. *25*, 6435-6448.
79. Kikuchi, A., Ohashi, S., Fuse, N., Ohta, T., Suzuki, M., Suzuki, Y., Fujita, T., Miyamoto, T., Aonishi, T., and Miyakawa, H. (2012). Experience-dependent plasticity of the optomotor response in Drosophila melanogaster. Dev. Neurosci. *34*, 533-542.
80. Du, S.S., Zhai, X., Poczos, B., and Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. arXiv preprint arXiv:1810.02054.
81. Stanley, K.O., Clune, J., Lehman, J., and Miikkulainen, R. (2019). Designing neural networks through neuroevolution. Nature Machine Intelligence *1*, 24-35.
82. Olshausen, B.A., and Field, D.J. (2004). Sparse coding of sensory inputs. Curr. Opin. Neurobiol. *14*, 481-487.
83. Wager, S., Wang, S., and Liang, P.S. (2013). Dropout training as adaptive regularization. In Advances in neural information processing systems. pp. 351-359.
84. Juusola, M., and Hardie, R.C. (2001). Light Adaptation in Drosophila Photoreceptors II. Rising Temperature Increases the Bandwidth of Reliable Signaling. J. Gen. Physiol. *117*, 27-42.
85. Jones, P.W., and Gabbiani, F. (2012). Impact of neural noise on a sensory-motor pathway signaling impending collision. J. Neurophysiol. *107*, 1067-1079.
86. Laughlin, S.B., van Steveninck, R.R.d.R., and Anderson, J.C. (1998). The metabolic cost of neural information. Nat. Neurosci. *1*, 36-41.
87. Mo, C.-H., and Koch, C. (2003). Modeling reverse-phi motion-selective neurons in cortex: double synaptic-veto mechanism. Neural Comput. *15*, 735-759.
88. Bahroun, Y., Chklovskii, D., and Sengupta, A. (2019). A Similarity-preserving Network Trained on Transformed Images Recapitulates Salient Features of the Fly Motion Detection Circuit. In Advances in Neural Information Processing Systems. pp. 14201-14212.
89. Klapoetke, N.C., Nern, A., Peek, M.Y., Rogers, E.M., Breads, P., Rubin, G.M., Reiser, M.B., and Card, G.M. (2017). Ultra-selective looming detection from radial motion opponency. Nature *551*, 237.

90. Scheffer, L.K., Xu, C.S., Januszewski, M., Lu, Z., Takemura, S.-y., Hayworth, K.J., Huang, G.B., Shinomiya, K., Maitlin-Shepard, J., and Berg, S. (2020). A connectome and analysis of the adult Drosophila central brain. Elife *9*, e57443.
91. Clark, D.A., and Demb, J.B. (2016). Parallel computations in insect and mammalian visual motion processing. Curr. Biol. *26*, R1062-R1072.
92. Borst, A., and Helmstaedter, M. (2015). Common circuit design in fly and mammalian motion vision. Nat. Neurosci. *18*, 1067-1076.
93. Euler, T., Detwiler, P.B., and Denk, W. (2002). Directionally selective calcium signals in dendrites of starburst amacrine cells. Nature *418*, 845-852.
94. Vaney, D.I., Sivyer, B., and Taylor, W.R. (2012). Direction selectivity in the retina: symmetry and asymmetry in structure and function. Nat. Rev. Neurosci. *13*, 194-208.
95. Moulden, B., and Begg, H. (1986). Some tests of the Marr-Ullman model of movement detection. Perception *15*, 139.
96. Hu, Q., and Victor, J.D. (2010). A set of high-order spatiotemporal stimuli that elicit motion and reverse-phi percepts. J. Vis. *10*.
97. Schiller, P.H., Finlay, B.L., and Volman, S.F. (1976). Quantitative studies of single-cell properties in monkey striate cortex. I. Spatiotemporal organization of receptive fields. J. Neurophysiol. *39*, 1288.
98. Mather, G., Moulden, B., and O'Halloran, A. (1991). Polarity specific adaptation to motion in the human visual system. Vision Res. *31*, 1013-1019.
99. Yildizoglu, T., Riegler, C., Fitzgerald, J.E., and Portugues, R. (2020). A Neural Representation of Naturalistic Motion-Guided Behavior in the Zebrafish Brain. Curr. Biol.
100. Gjorgjieva, J., Sompolinsky, H., and Meister, M. (2014). Benefits of Pathway Splitting in Sensory Coding. J. Neurosci. *34*, 12127-12144.
101. Gjorgjieva, J., Meister, M., and Sompolinsky, H. (2019). Functional diversity among sensory neurons from efficient coding principles. PLoS Comp. Biol. *15*, e1007476.
102. Fisher, Y.E., Silies, M., and Clandinin, T.R. (2015). Orientation selectivity sharpens motion detection in Drosophila. Neuron *88*, 390-402.
103. Molina-Obando, S., Vargas-Fique, J.F., Henning, M., Gür, B., Schladt, T.M., Akhtar, J., Berger, T.K., and Silies, M. (2019). ON selectivity in Drosophila vision is a multisynaptic process involving both glutamatergic and GABAergic inhibition. eLife *8*, e49373.
104. Freifeld, L., Clark, D.A., Schnitzer, M.J., Horowitz, M.A., and Clandinin, T.R. (2013). GABAergic lateral interactions tune the early stages of visual processing in Drosophila. Neuron *78*, 1075-1089.
105. Zheng, L., de Polavieja, G.G., Wolfram, V., Asyali, M.H., Hardie, R.C., and Juusola, M. (2006). Feedback network controls photoreceptor output at the layer of first visual synapses in Drosophila. J. Gen. Physiol. *127*, 495-510.
106. Zheng, L., Nikolaev, A., Wardill, T.J., O'Kane, C.J., de Polavieja, G.G., and Juusola, M. (2009). Network adaptation improves temporal representation of naturalistic stimuli in Drosophila eye: I dynamics. PLoS One *4*, e4307.
107. Fairhall, A.L., Lewen, G.D., Bialek, W., and van Steveninck, R.R.d.R. (2001). Efficiency and ambiguity in an adaptive neural code. Nature *412*, 787.
108. Eichner, H., Joesch, M., Schnell, B., Reiff, D.F., and Borst, A. (2011). Internal structure of the fly elementary motion detector. Neuron *70*, 1155-1164.

109. McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S., and Baccus, S. (2016). Deep learning models of the retinal response to natural scenes. In Advances in neural information processing systems. pp. 1369-1377.
110. Maheswaranathan, N., McIntosh, L., Kastner, D.B., Melander, J., Brezovec, L., Nayebi, A., Wang, J., Ganguli, S., and Baccus, S.A. (2018). Deep learning models reveal internal structure and diverse computations in the retina under natural scenes. bioRxiv, 340943.
111. Tschopp, F.D., Reiser, M.B., and Turaga, S.C. (2018). A connectome based hexagonal lattice convolutional network model of the Drosophila visual system. arXiv preprint arXiv:1806.04793.
112. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., and Smith, N.J. (2020). Array programming with NumPy. Nature *585*, 357-362.
113. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., and Bright, J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261-272.
114. Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. IEEE Annals of the History of Computing *9*, 90-95.
115. Torre, V., and Poggio, T. (1978). A synaptic mechanism possibly underlying directional selectivity to motion. Proc. R. Soc. Lond. B *202*, 409-416.
116. Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics. (JMLR Workshop and Conference Proceedings), pp. 249-256.
117. Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. Network: Comput. Neural Syst. *12*, 199-213.