bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Models of heterogeneous dopamine signaling in an insect learning and memory center

Linnie Jiang[1,2] and Ashok Litwin-Kumar[1,c]

[1]Mortimer B. Zuckerman Mind Brain Behavior Institute, Department of Neuroscience, Columbia University, New York, NY 10027, USA
[2]Department of Neurobiology, Stanford University, Stanford, CA 94305, USA
[c]Corresponding author. Email: ak3625@columbia.edu.

Abstract

The *Drosophila* mushroom body exhibits dopamine (DA) dependent synaptic plasticity that underlies the acquisition and retrieval of associative memories. Classic studies have recorded DA activity in this system and identified signals related to external reinforcement such as reward and punishment. However, recent studies have found that other factors including locomotion, novelty, reward expectation, and internal state also modulate DA neurons. This heterogeneous activity is at odds with typical modeling approaches in which DA neurons are assumed to encode a global, scalar error signal. How can DA signals support appropriate synaptic plasticity in the presence of this heterogeneity? We develop a modeling approach that infers a pattern of DA activity that is sufficient to solve a defined set of behavioral tasks, given architectural constraints informed by knowledge of mushroom body circuitry. Model DA neurons exhibit diverse tuning to task parameters while nonetheless producing coherent learned behaviors. Our results provide a mechanistic framework that accounts for the heterogeneity of DA signals observed during learning and behavior.

Introduction

Dopamine (DA) release modulates synaptic plasticity and learning across vertebrate and invertebrate species (Perisse et al., 2013; Watabe-Uchida et al., 2017). A standard view of DA activity, proposed on the basis of recordings in the mammalian midbrain dopaminergic system, holds that DA

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

neuron firing represents a "reward prediction error" (RPE), the difference between reward received and predicted reward (Schultz et al., 1997). This view is consistent with models of classical conditioning experiments and with reinforcement learning algorithms that choose sequences of actions to maximize reward received (Sutton and Barto, 1998). A standard assumption in these models is that the scalar RPE signal is globally broadcast to and gates the modification of synaptic connections involved in learning. However, recent studies in both vertebrates and invertebrates suggest that DA neuron activity is modulated by other variables in addition to RPE, and that this modulation is heterogeneous across populations of DA neurons (Watabe-Uchida and Uchida, 2019).

In the *Drosophila* mushroom body (MB), Kenyon cells (KCs) conveying sensory information, predominantly odor-related signals, send parallel fibers that contact the dendrites of output neurons (MBONs). The activation of specific MBONs can bias the organism toward particular actions (Aso et al., 2014a). MBON dendrites define discrete anatomical regions, known as "compartments," each of which is innervated by distinct classes of dopaminergic neurons (DANs; we use the term DAN to refer specifically to mushroom body dopaminergic neurons). If the KCs and DANs that project to a given MBON are both active within a particular time window, KC-to-MBON synapses are strengthened or weakened depending on the relative timing of KC and DAN activation (Hige et al., 2015a; Aso and Rubin, 2016; Handler et al., 2019). The resulting synaptic modifications permit flies to learn and update associations between stimuli and reinforcement.

Early studies identified DAN activity in the MB related to reward and punishment, although whether this activity reflects prediction errors is unclear (Schwaerzel et al., 2003; Kim et al., 2007; Aso et al., 2010, 2012; Burke et al., 2012). More recently DANs have been shown to encode additional variables, including novelty (Hattori et al., 2017), reward prediction (Felsenberg et al., 2017, 2018), and locomotion-related signals (Cohn et al., 2015). DA signals related to movement, novelty and salience, and separate pathways for rewards and punishment have also been identified in mammalian midbrain regions (Steinfels et al., 1983; Ljungberg et al., 1992; Horvitz et al., 1997; Rebec et al., 1997; Lak et al., 2016; Bromberg-Martin et al., 2010; Menegas et al., 2017; Howe and Dombeck, 2016; Engelhard et al., 2019; Watabe-Uchida and Uchida, 2019). These observations call for extensions of classic models that assume DA neurons are globally tuned to RPE. How can DA signals gate appropriate synaptic plasticity and learning if their responses are modulated by mixed sources of information?

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To address this question, we develop a modeling approach that constructs a recurrent network that produces DA signals suited to learning a particular class of tasks. The network is constrained by the well-characterized anatomy of the MB and knowledge of the DA-dependent synaptic plasticity rules that modify its connections (Aso et al., 2014b; Handler et al., 2019). Comprehensive synapse-level wiring diagrams for the output circuitry of the MB will soon be available, which will allow the connectivity of models constructed with our approach to be further constrained by data (Eichler et al., 2017; Takemura et al., 2017; Zheng et al., 2018; Eschbach et al., 2019). The models we construct can solve complex behavioral tasks and generalize to novel stimuli while using only experimentally constrained plasticity rules. They can form associations based on limited numbers of stimulus/reinforcement pairings and are capable of continual learning, which are often challenging for artificial neural networks (Finn et al., 2017; Kirkpatrick et al., 2017). We use these models to predict DAN activity patterns that are suitable for learning the tasks we consider and find that different model DANs exhibit diverse tuning to task-related variables. Our approach uncovers the mechanisms behind the observed heterogeneity of DA signals in the MB and suggests that the "error" signals that support associative learning may be more distributed than is often assumed.

Results

Modeling recurrent mushroom body output circuitry

The diversity of DAN activity challenges models of MB learning that assume DANs convey global reward or punishment signals. Part of this discrepancy is likely due to the intricate connectivity among MBONs, DANs, and other neurons that form synapses with them (Aso et al., 2014b; Eschbach et al., 2019). We therefore modeled these neurons and their connections, which we refer to collectively as the MB "output circuitry," as a recurrent neural network (Fig. 1A). Recurrent connections within this network are defined by a matrix of synaptic weights $\mathbf{W}_{\text{recur}}$. Synapses from KCs onto MBONs provide the network with sensory information and are represented by $\mathbf{W}_{\text{KC}\rightarrow\text{MBON}}$. Separate pathways convey signals such as reward or punishment from external regions, via weights $\mathbf{W}_{\text{ext}}$. The objective of the network is to generate a desired pattern of activity in a readout that represents the behavioral bias produced by the MB. The readout decodes this desired output through weights $\mathbf{W}_{\text{readout}}$.

Recurrent network modeling approaches in neuroscience typically fix all of these synaptic weight

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

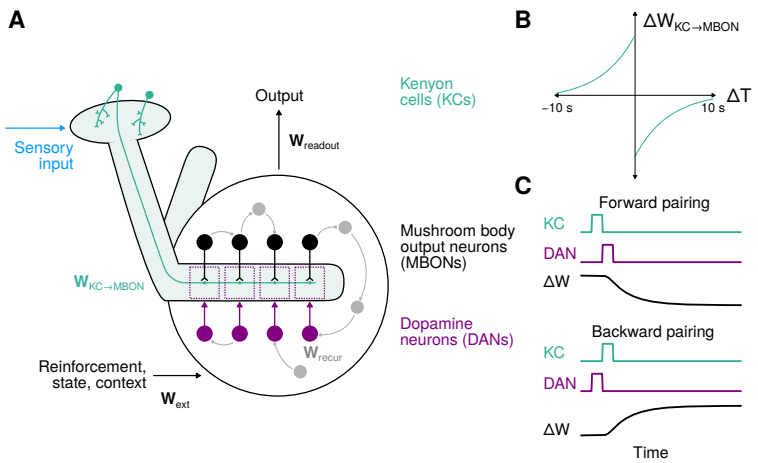


FIGURE 1: Diagram of the mushroom body (MB) model. **(A)** Kenyon cells (KCs) respond to stimuli and project to mushroom body output neurons (MBONs) via weights $\mathbf{W}_{\text{KC}\to\text{MBON}}$. These connections are dynamic variables that are modified according to a synaptic plasticity rule gated by dopamine neurons (DANs). MBONs and DANs are organized into compartments (dotted rectangles). External signals convey, e.g., reward, punishment, or context to the MB output circuitry according to weights $\mathbf{W}_{\text{ext}}$. A linear readout of the output circuitry with weights $\mathbf{W}_{\text{readout}}$ is used to determine the behavioral output of the system. Connections among MBONs, DANs, and feedback neurons (gray) are determined by weights $\mathbf{W}_{\text{recur}}$. **(B)** The form of the DAN-gated synaptic plasticity rule operative at KC-to-MBON synapses. ΔT is the time difference between KC activation and DAN activation. **(C)** Illustration of the change in KC-to-MBON synaptic weight ΔW following forward and backward pairings of KC and DAN activity.

matrices after optimizing them to produce a desired behavior. However, connections between KCs and MBONs are known to exhibit DA-gated synaptic plasticity. This plasticity is dependent on the relative timing of KC and DAN activation (notably, it does not appear to depend on the postsynaptic MBON firing rate; Hige et al., 2015a) and can drive substantial changes in evoked MBON activity even after brief KC-DAN pairings (Handler et al., 2019). We modeled this plasticity by assuming that each element w of $\mathbf{W}_{\text{KC}\to\text{MBON}}$ is a dynamic quantity that is modified according to the following update rule:

$$\frac{dw}{dt} = \alpha \left(\bar{r}_{\text{DAN}}(t) r_{\text{KC}}(t) - \bar{r}_{\text{KC}}(t) r_{\text{DAN}}(t) \right), \tag{1}$$

where r_{KC} and r_{DAN} are the firing rates of the KC and the DAN that innervate the corresponding compartment, $\bar{r}_{\text{KC}}$ and $\bar{r}_{\text{DAN}}$ are synaptic eligibility traces constructed by low-pass filtering r_{KC} and r_{DAN}, and α is a constant that determines the magnitude of synaptic plasticity. The time constants of the low-pass filters used to generate the eligibility traces determine the time window within which pairings of KC and DAN activity elicit appreciable changes of w. When KC and DAN firing rates are modeled as pulses separated by a time lag ΔT, the dependence of the change in w on ΔT takes the form of a biphasic timing-dependent function (Fig. 1B,C), consistent with a recent experimental characterization (Handler et al., 2019). The seconds-long timescale of this curve is compatible with

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the use of continuous firing rates rather than discrete spike timing to model KC-to-MBON plasticity, as we have done in Eq. 1.

Importantly, the weight update rule in Eq. 1 is a smooth function of network firing rates, allowing networks with this update rule to be constructed using standard gradient descent algorithms used in machine learning. Such an approach has been recently used to augment networks with adjustable connections ("fast weights"; Ba et al., 2016), although plasticity rules of the form of Eq. 1 have not been examined. To construct our networks, we use gradient descent to modify $\mathbf{W}_{\text{recur}}$, $\mathbf{W}_{\text{ext}}$, and $\mathbf{W}_{\text{readout}}$ (the connections describing the MB output circuitry) to optimize the performance of the network on a given set of behavioral tasks. We refer to the gradient descent modification of these weights as the "optimization" phase of constructing our networks. This optimization represents the evolutionary and developmental processes that produce a network capable of efficiently learning new associations (Zador, 2019). After this optimization is complete, the output circuitry is fixed but KC-to-MBON weights are subject to synaptic plasticity according to Eq. 1. To begin, we assume that KC-to-MBON weights are set to their baseline values at the beginning of each trial in which new assocations are formed. Later, we will consider the case of continual learning of many associations.

Models of associative conditioning

We begin by considering models of classical conditioning, which involve the formation of associations between a conditioned stimulus (CS) and unconditioned stimulus (US) such as reward or punishment. A one-dimensional readout of the MBON population is taken to represent the stimulus valence, which measures whether the organism prefers (valence > 0) or avoids (valence < 0) the CS. In the model, CS are encoded by the activation of a random ensembles of KCs. Rewards and punishments are encoded by external inputs to the network.

To construct the model, we optimized the MB output circuitry to produce a target valence in the readout during presentation of CS+ that have been paired with US (first-order conditioning; Fig. 2A,B, top). After optimization, the valence of the associated US is reported for CS+ but not unconditioned stimulus (CS-) presentations. The activities of subsets of model MBONs are suppressed following conditioning, indicating that the network learns to modify its responses for CS+ but not CS- responses (Fig. 2A,B, bottom) This form of classical conditioning requires an appropriate mapping from US pathways to DANs, but recurrent MB output circuitry is not required; networks

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

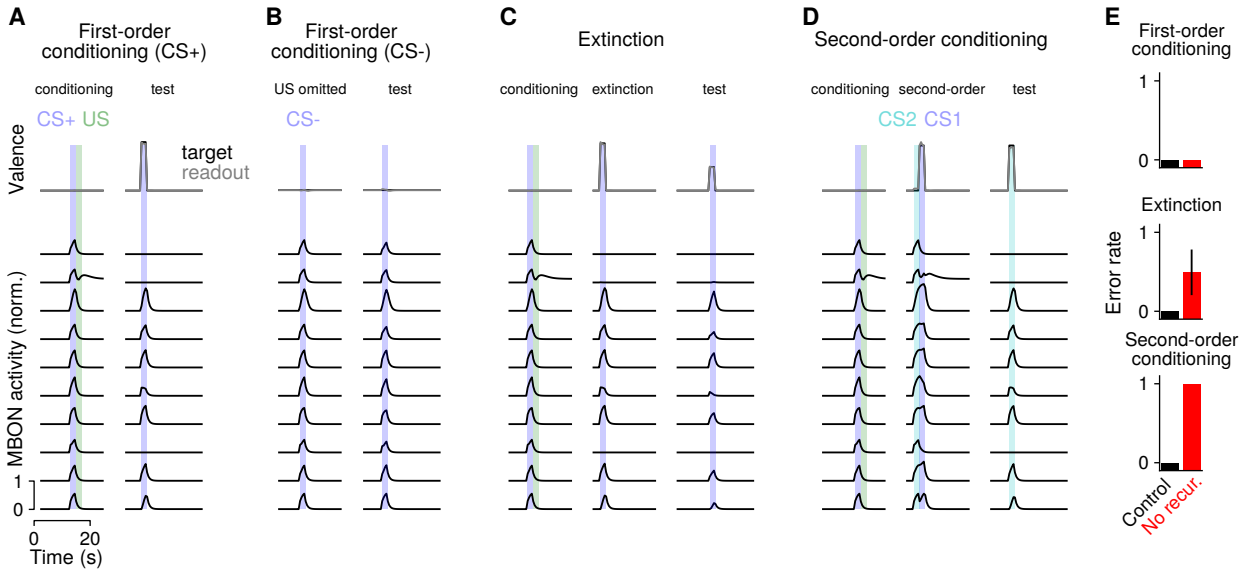


FIGURE 2: Behavior of network during reward conditioning paradigms. **(A)** Behavior of MBONs during first-order conditioning. During training, a CS+ (blue) is presented, followed by a US (green). Top: The network is optimized so that a readout of the MBON activity during the second CS+ presentation encodes the valence of the conditioned stimulus (gray curve). Black curve represents the target valence and overlaps with the readout. Bottom: Example responses of MBONs. **(B)** Same as **A**, but for a CS- presentation without US. **(C)** Same as **A**, but for extinction, in which a second presentation of the CS+ without the US partially extinguishes the association. **(D)** Same as **A**, but for second-order conditioning, in which a second stimulus (CS2) is paired with a conditioned stimulus (CS1). **(E)** Error rate averaged across networks in different paradigms. An error is defined as a difference between reported and target valence with magnitude greater than 0.2 during the test period. Networks optimized with recurrent MB output circuitry (control; black) are compared to networks without recurrence (no recur.; red).

without recurrence also produce the target valence (Fig. 2E, top). We therefore considered a more complex set of tasks. Networks were optimized to perform first-order conditioning, to extinguish associations upon repeated presentation of a CS+ without US, and also to perform second-order conditioning.

During extinction, the omission of a US following a previously conditioned CS+ reduces the strength of the learned association (Fig. 2C). In second-order conditioning, a CS (CS1) is first paired with a reward or punishment (Fig. 2D, left), and then a second CS (CS2) is paired with CS1 (Fig. 2D, middle). Because CS2 now predicts CS1 which in turn predicts reward or punishment, the learned valence of CS1 is transferred to CS2 (Fig. 2D, right). In both extinction and second-order conditioning, a previously learned association must be used to instruct either the modification of an existing association (in the case of extinction) or the formation of a new association (in the case of second-order conditioning). We hypothesized that recurrent output circuitry would be required in these cases. Indeed, non-recurrent MB networks are unable to solve these tasks, while recurrent networks are (Fig. 2E, middle, bottom). Thus, for complex relationships between stimuli beyond

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

first-order conditioning, recurrent output circuitry provides a substantial benefit.

Comparison to networks without plasticity

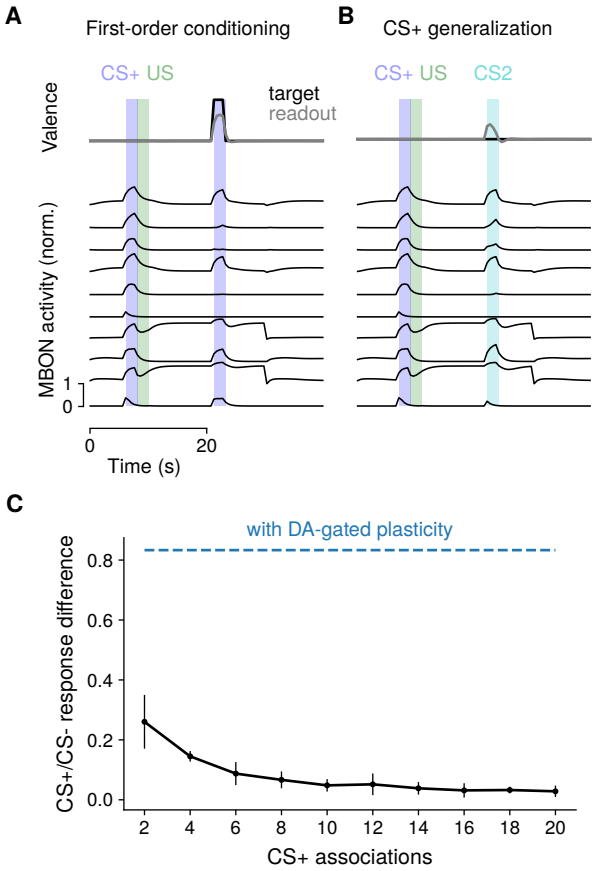


FIGURE 3: Comparison to networks without DA-gated plasticity. **(A)** Behavior during first-order conditioning, similar to Fig. 2A, but for a non-plastic network. Because of the need for non-plastic networks to maintain information using persistent activity, performance degrades with longer delays between training and testing phases. We therefore chose this delay to be shorter than in Fig. 2A. **(B)** Same as **A**, but for a trial in which a CS-US pairing is followed by the presentation of a neutral CS. **(C)** Difference in response (reported valence) for CS+ and CS- as a function of the number of CS+ associations. Each CS+ is associated with either a positive or negative US. A difference of 0 corresponds to overgeneralization of the CS+ valence to neutral CS-. For comparison, the corresponding response difference for networks with DA-gated plasticity is shown in blue.

Standard recurrent neural networks can maintain stimulus information over time through persistent neural activity, without modification of synaptic weights. This raises the question of whether the DA-gated plasticity we implemented is necessary to recall CS-US associations, or if recurrent MB output circuitry alone is sufficient. We therefore compared the networks described above to networks lacking this plasticity. For non-plastic networks, connections from KCs to MBONs are set to fixed, random values (reflecting the fact that these weights are not specialized to specific odors;

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hige et al., 2015b). Networks are optimized to associate a limited number of CS+ with either a positive or negative valence US, while not responding to CS-.

Non-plastic networks can form CS-US associations (Fig. 3A). Compared to networks with DA-gated plasticity (Fig. 2A), MBONs exhibit stronger persistent activity following a CS-US pairing. This activity retains information about the learned association as an "attractor" of neural activity (Hopfield, 1982). However, non-plastic networks exhibit a high degree of overgeneralization of learned associations to neutral CS- stimuli (Fig. 3B). This likely reflects a difficulty in constructing a large number of attractors, corresponding to each possible CS-US pairing, that do not overlap with patterns of activity evoked by other CS- stimuli. Consistent with this, as the number of CS+ increases, the difference between the reported valence for CS+ and CS- decreases, reflecting increasing overgeneralization (Fig. 2C). Networks with DA-gated plasticity do not suffer from such overgeneralization, as they can store and update the identities of stimuli in plastic weights.

In total, the comparison between plastic and non-plastic networks demonstrates that the addition of DA-gated plasticity at KC-to-MBON synapses improves capacity and reduces overgeneralization. Furthermore, plastic networks need not rely solely on persistent activity in order to store associations (compare Fig. 2A and Fig. 3A), likely prolonging the timescale over which information can be stored without being disrupted by ongoing activity.

Distributed representations across DANs

We next examined the responses of DANs to neutral, unconditioned, and conditioned stimuli in the networks we constructed, to examine the "error" signals responsible for learning (Fig. 4A). DANs exhibited heterogeneity in their responses. We performed hierarchical clustering to identify groups of DANs with similar response properties (Fig. 4B, gray). This procedure identified two broad groups of DANs—one that responds to positive-valence US and another that responds to negative-valence US—as well as more subtle features in the population response.

While some DANs increase their firing only for US, most also respond to conditioned CS. In some cases, this response includes a decrease in firing rate in response to the omission of a predicted US that would otherwise cause an increase in rate, consistent with a reward prediction error. In other cases, neurons respond only with increases in firing rate for US of a particular valence, and for

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

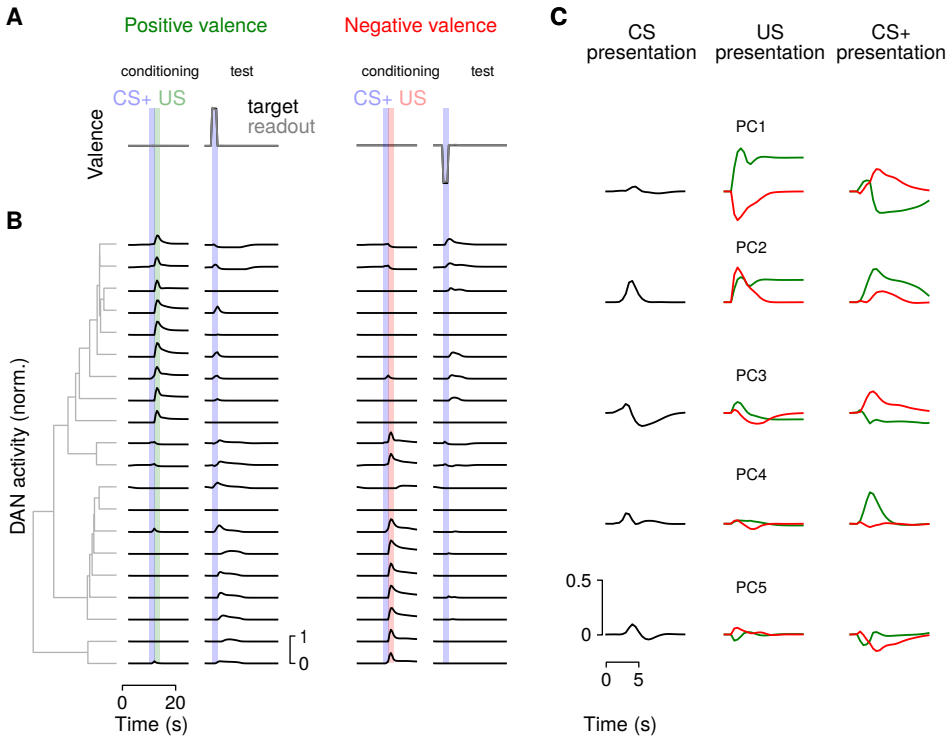


FIGURE 4: Population analysis of DAN activity. Principal components analysis of DAN population responses during presentation of neutral CS. **(A)** Responses are shown for CS+ conditioning with a US of positive (left) or negative (valence), followed by a test presentation of the conditioned CS+ without US. **(B)** Responses of model DANs from a single network. DANs are sorted according to hierarchical clustering (illustrated with gray dendrogram) of their responses. **(C)** Principal components analysis (PCA) of DAN population activity. Left: Response to a neutral CS. Middle: Response to a positive (green) or negative (red) valence US. Right: Response to a previously conditioned US.

omitted US of the opposite valence, consistent with cross-compartmental interactions supporting the prediction of valence (Felsenberg et al., 2017). The presence of both reward prediction error-like responses and valence-specific omission responses suggests that multiple learning mechanisms are employed by the network to perform tasks such as extinction and second-order conditioning.

Our examination of DAN responses demonstrates that DANs in our models are diversely tuned to CS and US valence. This tuning implies that KC-to-MBON synapses change in a heterogeneous manner in response to CS and US presentations, but that these changes are sufficient to produce an appropriate behavioral response collectively. Consistent with this idea, principal components analysis (PCA) of DAN responses identified modes of activity with interpretable, task-relevant dynamics. The first principal component (PC1; Fig. 4C) reflected US valence and predicted CS+ valence, while rapidly changing sign upon US omission, consistent with a reward prediction error. Subsequent PCs included components that responded to CS and US of both valences (PC2) or tuned primarily to a single stimulus, such as a positive valence CS+ (PC4).

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To further explore how DAN responses depend on the task being learned, we extended the model to require encoding of novelty and familiarity, inspired by a recent study that showed that the MB is required for learning and expressing an alerting behavior driven by novel CS (Hattori et al., 2017). We added a second readout that reports CS novelty, in addition to the readout of valence described previously. Networks optimized to report both variables exhibit enhanced CS responses and a large novelty-selective component in the population response identified by PCA (Supplemental Fig. 1), compared to networks that only report valence (Fig. 4B). These results suggest that DANs collectively respond to any variables relevant to the task for which the output circuitry is optimized, which may include variables distinct from reward prediction. Furthermore, the distributed nature of this representation implies that individual variables may be more readily decoded from populations of DANs than from single DANs.

Continual learning of associations

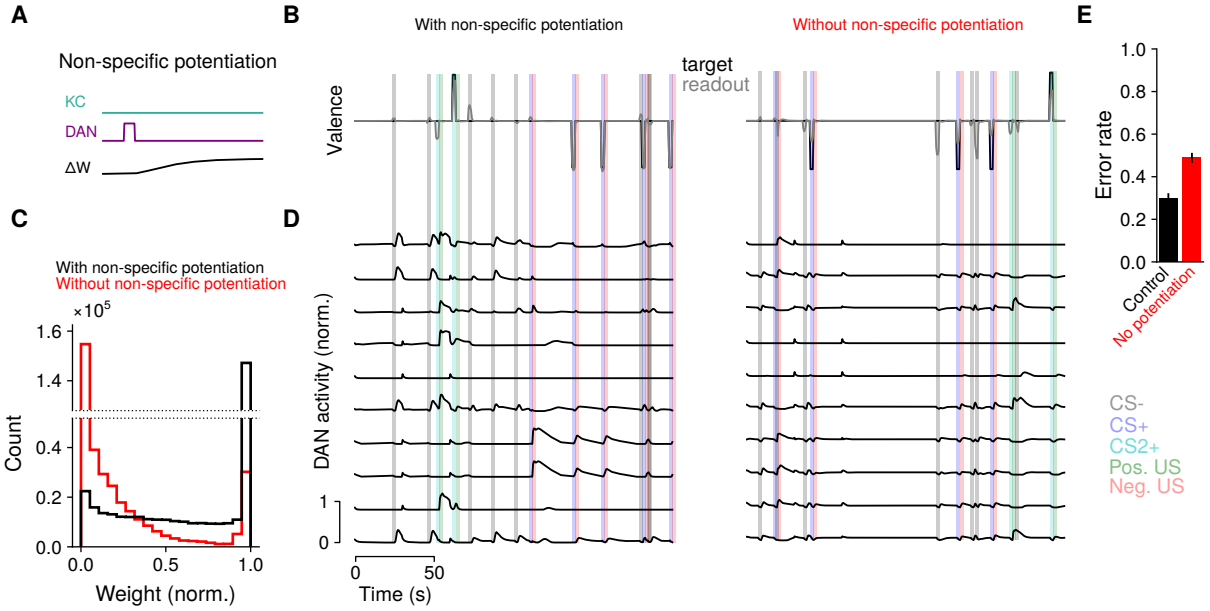


FIGURE 5: Model behavior for long sequences of associations. **(A)** Illustration of non-specific potentiation following DAN activity (compare with Fig. 1C). **(B)** Histogram of synaptic weights after a long sequence of CS and US presentations for networks with (black) and without (red) non-specific potentiation. Weights are normalized to their maximum value. **(C)** Top: Example sequence of positive and negative associations between two odors CS+ and CS2+ and US. Neutral gray odors (CS-) are also presented randomly. Bottom: DAN responses for the sequence of CS and US presentations. **(C)** Same as **C**, but for a network without non-specific potentiation. Such networks are less likely to report the correct valence for conditioned CS+ and also exhibit a higher rate of false positive responses to CS-. **(E)** Error rate (defined as a difference between reported and target valence with magnitude greater than 0.5 during a CS presentation) for networks with (black) and without (red) non-specific potentiation.

In the previous sections, we modeled the dynamics of networks during individual trials containing

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a limited number of associations. We next ask whether these networks are capable of continual learning, in which long sequences of associations are formed, with recent associations potentially overwriting older ones. Such learning is often challenging, particularly when synaptic weights have a bounded range, due to the tendency of weights to saturate at their minimum or maximum value after many associations are formed (Fusi and Abbott, 2007). To combat this, a homeostasic process that prevents such saturation is typically required. We therefore if our optimized networks can implement such homeostasis.

In certain compartments of the MB, it has been shown that the activation of DANs in the absence of KC activity leads to potentiation of KC-to-MBON synapses (Aso and Rubin, 2016). This provides a mechanism for the erasure of memories formed following synaptic depression. We hypothesized that this non-specific potentiation could implement a form of homeostasis that prevents widespread synaptic depression after many associations are formed. We therefore augmented our DA-gated synaptic plasticity rule (Fig. 1C) with such potentiation (Fig. 5A). The new synaptic plasticity rule is given by:

$$\frac{dw}{dt} = \alpha \left(\bar{r}_{\mathrm{DAN}}(t) r_{\mathrm{KC}}(t) - \bar{r}_{\mathrm{KC}}(t) r_{\mathrm{DAN}}(t)\right) + \beta \bar{r}_{\mathrm{DAN}}(t), \tag{2}$$

where β represents the rate of non-specific potentiation (compare with Eq. 1). We allowed β to be optimized by gradient descent individually for each compartment.

We modeled long sequences of associations in which CS+, CS-, and US are presented randomly (Fig. 5B). We then examined the distribution of KC-to-MBON synaptic weights after such sequences. Without non-specific potentiation, most synaptic weights are clustered near 0 (Fig. 5C, red). However, the addition of this potentiation substantially changes the synaptic weight distribution, with many weights remaining potentiated even after thousands of CS and US presentations (Fig. 5C, black).

We also examined performance and DAN responses in the two types of networks. Without non-specific potentiation, DAN responses are weaker and the reported valence less accurately tracks the target valence, compared to networks with such potentiation (Fig. 5D,E). In total, we find that our approach can construct models that robustly implement continual learning if provided with homeostatic mechanisms that can maintain a stable distribution of synaptic weights.

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Associating stimuli with changes in internal state

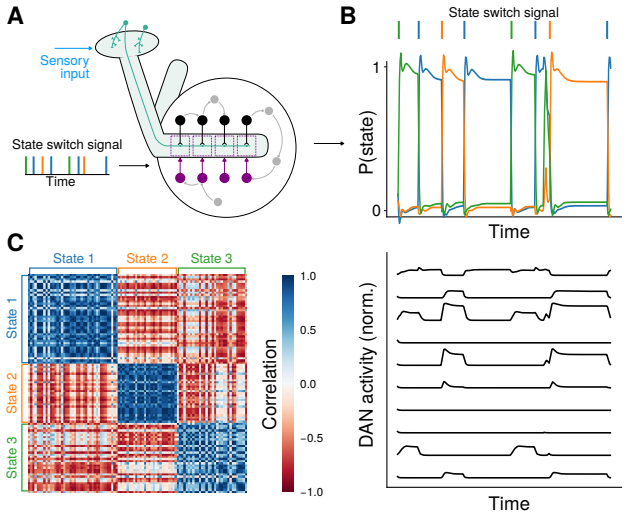


FIGURE 6: **(A)** Diagram of a network whose activity transitions between a sequence of discrete states. Brief pulse inputs to the network signal that a switch to a new state should occur. **(B)** Top: A linear readout of MBON activity can be used to decode the network state. Bottom: DAN activity exhibits state-dependent fluctuations. **(C)** Decoding of stimuli that predict state transitions. Heatmap illustrates the correlation between MBON population responses to the presentation of different stimuli that had previously been presented prior to a state transition. Stimuli are ordered based on the state transitions that follow their first presentation. Blue blocks indicate that stimuli that predict the same state transition evoke similar MBON activity.

In the previous sections, we focused on networks whose DANs exhibited transient responses to the presentation of relevant external cues. Recent studies have found that DANs also exhibit continuous fluctuations that track the state of the fly, even in the absence of overt external reinforcement. These fluctuations are correlated with transitions between, for example, movement and quiescence (Cohn et al., 2015), or hunger and satiation (Krashes et al., 2009). Understanding the functional role of these DAN fluctuations is a major challenge for models of DA-dependent learning. We hypothesized that such activity could permit the association of stimuli with the internal state of the organism. This could allow downstream networks to read out whether a stimulus has previously been experienced in conjuction with a particular change in state, which might inform an appropriate behavioral response to that stimulus.

To test this hypothesis, we constructed a network that transitioned between a set of three discrete states, triggered on input pulses that signal the identity of the next state (Fig. 6A). This input represents signals from other brain areas that drive state transitions. We optimized the output circuitry to continuously maintain a state representation, quantified by the ability of a linear readout of MBON activity to decode the current state (Fig. 6B, top). This led to widespread state-dependent

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

activity throughout the network, including among DANs (Fig. 6B, bottom).

We next examined MBON responses to the presentation of stimuli that had previously preceded a transition to some state. If a transition to a given state reliably evokes a particular pattern of DAN activity, then KC-to-MBON synapses that are activated by any stimulus preceding such a transition will experience a similar pattern of depression or potentiation. Consistent with this prediction, the pattern of MBON activity evoked by a stimulus that predicts a transition to state S_1 is more similar to the corresponding activity for other stimuli that predict the same state than any other state S_2 (Fig. 6C). The representations of state-transition-predictive stimuli are thus "imprinted" with the identity of the predicted state. This could allow circuits downstream of the MB to consistently produce a desired behavior that depends on the internal state, instead of or in addition to the external reinforcement, that is predicted by a stimulus. Our model thus provides a hypothesis for the functional role of state-dependent DAN activity.

Mixed encoding of reward and movement in models of navigation

We also examined models of dynamic, goal directed behaviors. An important function of olfactory associations in *Drosophila* is to enable navigation to the sources of reward-predicting odor cues, such as food odors (Gaudry et al., 2012). We therefore optimized networks to control the forward and angular velocity of a simulated organism in a two-dimensional environment. The environment contains multiple odor sources that produce odor plumes that the the organism encounters as it moves. The organism is first presented with a CS+/reward pairing and then is placed in the two-dimensional environment and must navigate to the rewarded odor (Fig. 7A, top). This is a complex behavior that requires storing the identity of the rewarded odor, identifying the upwind direction for that odor, moving toward the odor source using concentration information, and ignoring neutral odors. We assumed that the MB output circuitry supports these computations by integrating odor concentration input from KCs and information from other brain areas about wind direction relative to the organism's orientation (Fig. 7A, bottom; Suver et al., 2019).

The simulated organism can successfully navigate to the rewarded odor source (Fig. 7B), and successful navigation requires plasticity during conditioning that encodes the CS+/US pairing (Supplemental Fig. 2). We wondered whether DA-gated plasticity might also be operative during navigation, based on recent findings that recorded ongoing DAN fluctuations correlated with movement

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

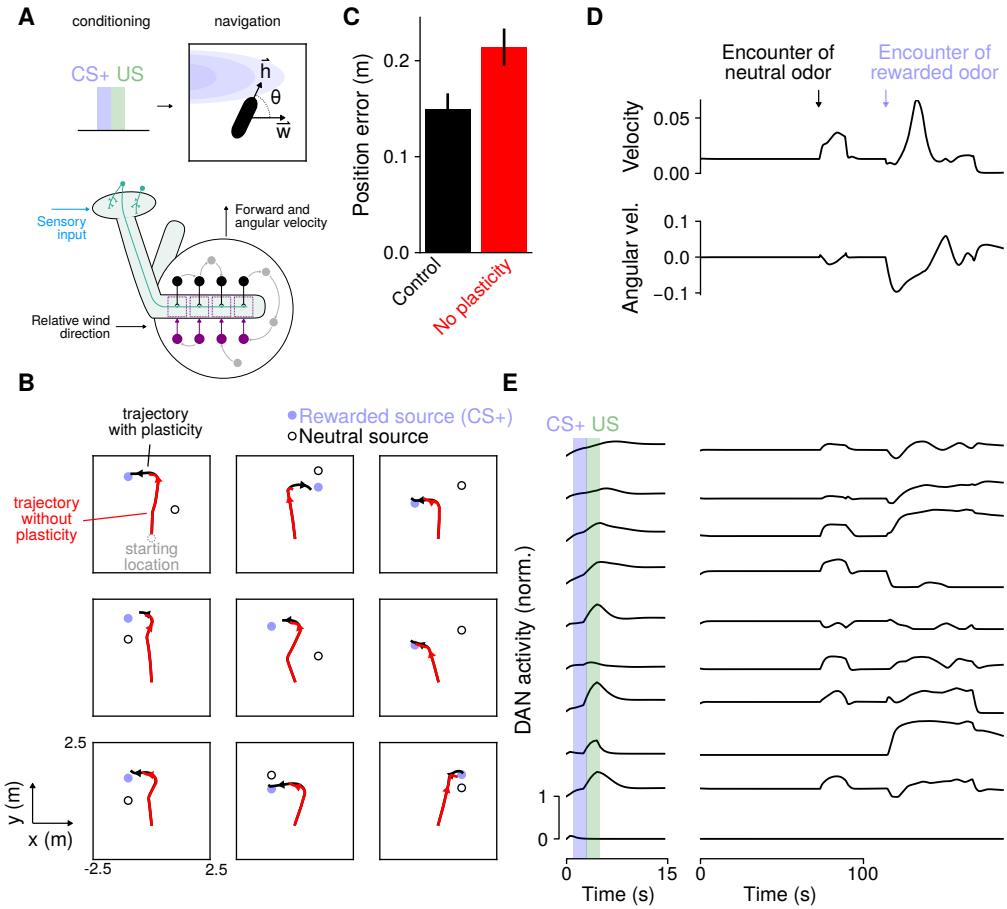


FIGURE 7: **(A)** Top: Schematic of navigation task. After conditioning, the simulated organism uses odor concentration input (blue) and information about wind direction **w** relative to its heading **h**. Bottom: Diagram of a network that uses these signals to compute forward and angular velocity signals for navigation. Velocity signals are read out from other neurons in the MB output circuitry (gray), rather than MBONs. **(B)** Position of the simulated organism as a function of time during navigation. Black: Simulation with intact DA-gated plasticity during navigation; Red: Simulation with plasticity blocked. Arrowheads indicate direction of movement. In the top left plot, the starting location (gray circle) is indicated. **(C)** Position error (mean-squared distance from rewarded odor source at the end of navigation) for control networks and networks without DA-gated plasticity. **(D)** Forward (top) and angular (bottom) velocity as a function of time during one example navigation trial. **(E)** DAN activity during the same trial as in **D**.

(Cohn et al., 2015). We asked whether such plasticity during navigation is important for the behavior of the model by examining the performance of networks in which this plasticity is blocked after the networks are optimized. Blocking plasticity during navigation impairs performance, suggesting that it contributes to the computation being performed by the MB output circuitry (Fig. 7C). In particular, networks lacking plasticity often exhibit decreased forward velocity after entering a plume corresponding to a rewarded odor (Fig. 7B), suggesting that ongoing plasticity may reinforce salient odors as they are encountered and promote odor-seeking.

We also examined the relationship of DAN activity with movement variables during navigation. The

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

simulated organism exhibits increased forward velocity and turning upon the encounter of an odor, with greater increases for rewarded than neutral odors (Fig. 7D). Model DANs exhibit activity during navigation that correlates with movement (Fig. 7E). Many of the same DANs also exhibit reward-related activity, demonstrating that they multiplex reward and movement-related signals. Thus, our model accounts for DAN tuning to these two types of signals, a feature present in recordings that traditional modeling approaches do not capture (Cohn et al., 2015).

Discussion

We have developed models of the MB that use a biologically plausible form of DA-gated synaptic plasticity to solve a variety of learning tasks. By optimizing the MB output circuitry for task performance, these models generate patterns of DAN activity sufficient to produce the desired behaviors. Model DAN responses are distributed, tuned to multiple task-relevant variables, and exhibit rich temporal fluctuations. This diversity is a result of optimizing our models only for task performance rather than assuming that DANs uniformly represent a particular quantity of interest, such as a global reward prediction error signal (Schultz et al., 1997). Our results predict that individual DANs may exhibit diverse tuning while producing coherent activity at the population level. They also provide the first unified modeling framework that can account for valence and reward prediction (Fig. 4), novelty (Supplemental Fig. 1), and movement-related (Fig. 7) DAN responses that have been recorded in experiments.

Relationship to other modeling approaches

To construct our MB models, we took advantage of recent advances in recurrent neural network optimization to augment standard network architectures with DA-gated plasticity. Our approach can be viewed as a form of “meta-learning” (Finn et al., 2017), or “learning to learn,” in which a network learns through gradient descent to use a differentiable form of synaptic plasticity (Eq. 1) to solve a set of tasks. As we have shown, this meta-learning approach allows us to construct networks that exhibit continual learning and can form associations based on single CS-US pairings (Fig. 5). Recent studies have modeled networks with other forms of differentiable plasticity, including Hebbian plasticity (Ba et al., 2016; Miconi et al., 2018; Orhan and Ma, 2019). In our case, detailed knowledge of the site and functional form of plasticity (Handler et al., 2019) allowed us to investigate

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

specific predictions about DAN responses. Similar approaches may be effective for modeling other brain areas in which the neurons responsible for conveying "error" signals can be identified, such as the cerebellum or basal ganglia (Ito et al., 1982; Watabe-Uchida et al., 2017).

Another recent study used a meta-learning approach to model DA activity and activity in the prefrontal cortex (PFC) of mammals (Wang et al., 2018). Unlike our study, in which the "slow" optimization is taken to represent evolutionary and developmental processes that determine the MB output circuitry, in this study the slow component of learning involved DA-dependent optimization of recurrent connections in PFC. This process relied on gradient descent in a recurrent network of long short-term memory (LSTM) units, leaving open the biological implementation of such a learning process. Like in actor-critic models of the basal ganglia (Barto, 1995), DA was modeled as a global RPE signal.

Heterogeneity of DA signaling in mammals

Numerous recent studies have described heterogeneity in DA signals of the mammalian midbrain dopaminergic system reminiscent of the heterogeneity across DANs in the MB (Watabe-Uchida and Uchida, 2019). These include reports detailing distinct subtypes of DA neurons that convey positive or negative valence signals or respond to salient signals of multiple valences (Matsumoto and Hikosaka, 2009; Bromberg-Martin et al., 2010), novelty responses (Steinfels et al., 1983; Ljungberg et al., 1992; Horvitz et al., 1997; Rebec et al., 1997; Lak et al., 2016; Menegas et al., 2017), responses to threat (Menegas et al., 2018), and modulation of DA neurons by movement (Howe and Dombeck, 2016; Engelhard et al., 2019). In many cases, these subtypes are defined by their striatal projection targets, suggesting a compartmentalization of function similar to that of the MB (Watabe-Uchida and Uchida, 2019). However, the logic of this compartmentalization is not yet clear.

Standard reinforcement learning models of the basal ganglia, such as actor-critic models, assume that DA neurons are globally tuned to reward prediction error (RPE) signals (Barto, 1995). Proposals have been made to account for heterogeneous DA responses, including that different regions produce sensory prediction errors based on access to distinct state information (Lau et al., 2017), or that DA neurons implement an algorithm for learning the statistics of transitions between states (Gardner et al., 2018). Our results are compatible with these theories, but different in that our

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

model does not assume that all DANs encode prediction errors. Instead, prediction error coding by particular modes of population activity emerges in our model as a consequence of optimizing for task performance (Fig. 4).

Connecting mushroom body architecture and function

The identification of groups of DANs that respond to positive and negative valence US (Schwaerzel et al., 2003), MBONs whose activity promotes approach or avoidance (Aso et al., 2014a), and DA-gated plasticity of KC-to-MBON synapses (Aso and Rubin, 2016; Handler et al., 2019) has led to effective models of first-order appetitive and aversive conditioning in *Drosophila*. A minimal model of such learning requires only two compartments of opposing valence and no recurrence among MBONs or DANs. The presence of extensive recurrence (Aso et al., 2014b; Eschbach et al., 2019) and DANs that are modulated by other variables (Cohn et al., 2015; Hattori et al., 2017; Felsenberg et al., 2017, 2018) suggests that the MB modulates learning and behavior along multiple axes.

The architecture of our model reflects the connectivity between KCs and MBONs, compartmentalization among MBONs and DANs, and recurrence of the MB output circuitry. While the identities of MBONs and DANs have been mapped anatomically (Aso et al., 2014b), the feedback pathways have not, so the feedback neurons in our model (gray neurons in Fig. 1A) represent any neurons that participate in recurrent loops involving the MB, which may involve paths through other brain areas. As electron-microscopy reconstructions of these pathways become available, effective interactions among compartments in our model may be compared to anatomical connections, and additional constraints may be placed on model connectivity. By modifying its architecture, our model could be used to test the role of other types of interactions, such as recurrence among KCs, connections between KCs and DANs (Eichler et al., 2017), or direct depolarizing or hyperpolarizing effects of DA on MBONs (Takemura et al., 2017). There is evidence that DA-gated synaptic plasticity rules are heterogeneous across compartments, which could also be incorporated into future models (Hige et al., 2015a; Aso and Rubin, 2016). While we have primarily focused on the formation of associations over short timescales because the detailed parameters of compartment-specific learning rules have not been described, such heterogeneity will likely be particularly important in models of long-term memory (Trannoy et al., 2011; Aso et al., 2012).

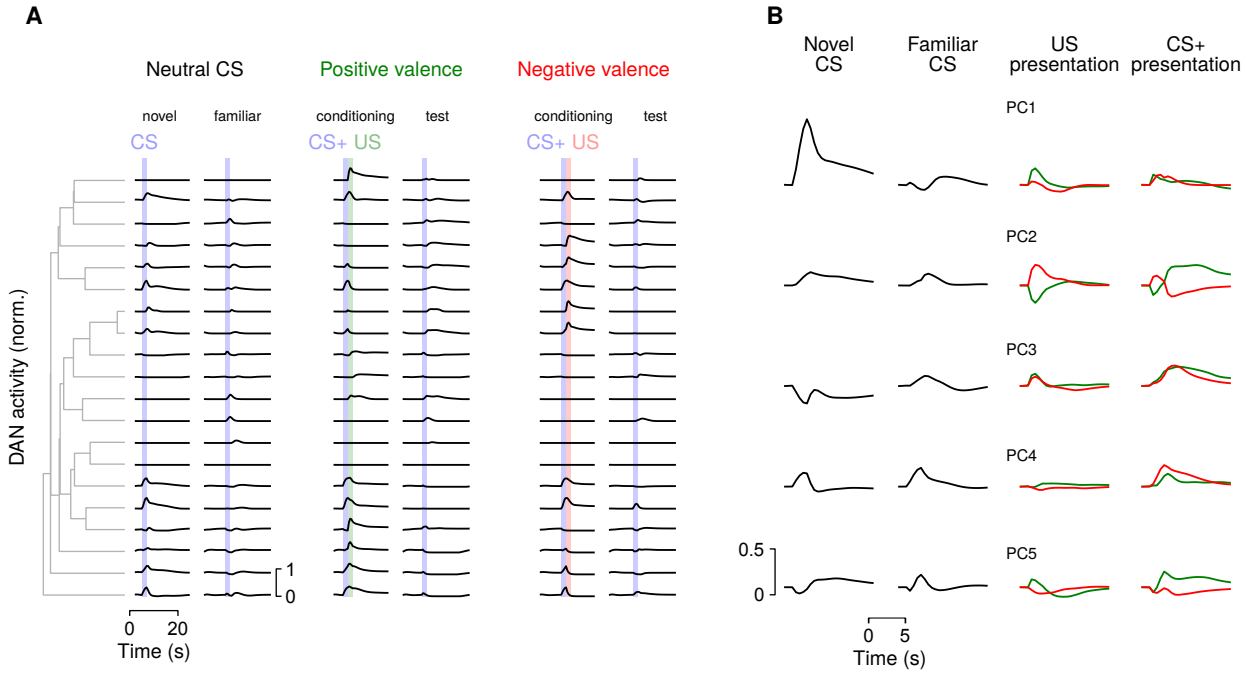
It is unlikely that purely anatomical information, even at the level of a synaptic wiring diagram,

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

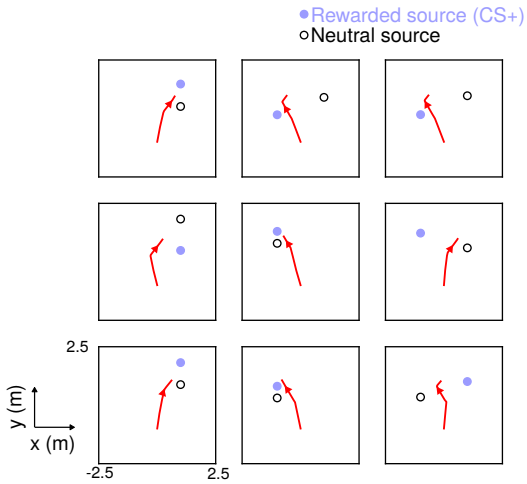
will be sufficient to infer how the MB functions (Bargmann and Marder, 2013). We have used anatomical information and parameterized synaptic plasticity rules along with hypotheses about which behaviors the MB supports to build "task-optimized" models, related to approaches that have been applied to sensory systems (Yamins and DiCarlo, 2016). The success of these approaches for explaining neural data relies on the availability of complex tasks that challenge and constrain the computations performed by the models. Therefore, experiments that probe the axes of fly behavior that the MB supports, including behaviors that cannot be described within the framework of classical conditioning, will be a crucial complement to connectivity mapping efforts as models of this system are refined.

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplemental figures



SUPPLEMENTAL FIGURE 1: Similar to Fig. 4, but for a network with two readouts that encode both valence and novelty. The novelty readout is active for the first presentation of a given CS and zero otherwise. **(A)** The addition of novelty as a readout dimension introduces DAN responses that are selective for novel CS. Compare with Fig. 4B. **(B)** The first principal component (PC1) for the network in **A** is selective for CS novelty. Compare with Fig. 4C.



SUPPLEMENTAL FIGURE 2: Similar to Fig. 7B, but for a network lacking KC-to-MBON synaptic plasticity during both conditioning and navigation. The model organism is unable to identify the rewarded odor and navigate toward it. Trajectories tend toward points located between the two odor sources.

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Network dynamics

The networks consist of 20 mushroom body ouput neurons (MBONs), 20 dopamine neurons (DANs), and 60 feedback neurons (FBNs), which we collectively refer to as the MB output circuitry. Stimulus input is provided by 200 Kenyon cells (KCs). The behavior of neuron i belonging to the MB output circuitry is given by:

$$\tau \frac{d\mathbf{r}_i}{dt} = -\mathbf{r}_i(t) + \left[\sum_j \mathbf{W}_{ij}^{\text{recur}} \mathbf{r}_j(t) + \mathbf{b}_i + \mathbf{I}_i(t) \right]_+ , \tag{3}$$

where $[\cdot]_+$ represents (elementwise) positive rectification. For computational efficiency and ease of training, we assume τ = 1 s and simulated the system with a timestep of $\Delta t = 0.5$ s, but our results do not depend strongly on these parameters. The bias $\mathbf{b}_i$ determines the excitability of neuron i, while $\mathbf{I}_i(t)$ represents its external input. We do not constrain $\mathbf{W}_{ij}^{\text{recur}}$, except that entries corresponding to connections from DANs to MBONs are set to zero, based on the assumption that these connections modulate plasticity of KC-to-MBON synapses rather than MBON firing directly (see Discussion).

If neuron i is an MBON, then $\mathbf{I}_i(t) = \sum_k \mathbf{W}_{ik}^{\text{KC}\to\text{MBON}} \mathbf{r}_k^{\text{KC}}$, representing input from KCs. If neuron i is a FBN, then $\mathbf{I}_i(t) = \sum_k \mathbf{W}_{ik}^{\text{ext}} \mathbf{r}_k^{\text{ext}}$, representing reinforcement, context, or state-dependent input from other brain regions. For DANs, $\mathbf{I}_i(t) = 0$. For tasks in which the predicted valence of a stimulus is read out, the activity of the readout is given by $\sum_i \mathbf{W}_i^{\text{readout}} \mathbf{r}_i$, where $\mathbf{W}_i^{\text{readout}}$ is nonzero only for MBONs. Readouts for other tasks are described below.

Aside from KC-to-MBON synaptic weights $\mathbf{W}_{\text{KC}\to\text{MBON}}$, other model parameters, specifically $\mathbf{W}_{\text{recur}}$, $\mathbf{b}$, $\mathbf{W}_{\text{ext}}$, and $\mathbf{W}_{\text{readout}}$ are optimized using gradient descent. For KC-to-MBON synapses, each weight is initially set to its maximum value of 0.05 and subsequently updated according to Eq. 1, with the updates of $\mathbf{W}_{\text{KC}\to\text{MBON}}$ low-pass filtered with a timescale of $\tau_W = 5$ s to account for the timescale of LTD or LTP. KC-to-MBON weights are constrained to lie between 0 and 0.05.

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Optimization

Parameters are optimized using PyTorch with the RMSprop optimizer (`www.pytorch.org`) with a learning rate of 0.001 and batch size of 30. The cost to be minimized is equal to the squared distance between the actual and target PI averaged over timesteps, plus a regularization term for DAN activity. The regularization term equals $\alpha_{\text{DAN}} \sum_{t,i\in\text{DAN}} [\mathbf{r}_i(t) - 0.1]_+^2$, which penalizes DAN activity that exceeds a baseline level of 0.1.

All optimized weights are initialized as zero mean Gaussian variables. To initialize $\mathbf{W}_{\text{recur}}$, weights from a neuron belonging to neuron type X (where X = MBON, DAN, or FBN) have 0 mean and variance equal to $\frac{1}{\sqrt{2}N_X}$, where N_X equals the number of neurons of type X. For $\mathbf{W}_{\text{readout}}$, the variance is $1/N_{\text{MBON}}$ while for $\mathbf{W}_{\text{ext}}$, the variance is 1. Bias parameters are initialized at 0.1. At the beginning of each trial $\mathbf{r}_i(t)$ is 0 for MBONs and 0.1 for DANs or FBNs, to permit these neurons to exhibit low levels of baseline activity.

Conditioning tasks

For conditioning tasks in which the predicted valence of a conditioned stimulus (CS) is reported (such as first- and second-order conditioning and extinction), each CS is encoded by setting 10% of the entries of $\mathbf{r}_{\text{KC}}$ to 1 and the rest to 0. Unconditioned stimuli (US) are encoded by $\mathbf{r}_{\text{ext}}$ which is two-dimensional, with entries equal to 0 or 1 based on on the presence or absence of positive or negative-valence US. CS and US are presented for 2 s. Tasks are split into 30 s intervals (for example conditioning and test intervals; see Fig. 2). Stimulus presentation occurs randomly between 5 s and 15 s within these intervals. Firing rates are reset at the beginning of each interval, which prevents networks from using persistent activity to maintain associations.

When optimizing networks in Fig. 2, random extinction and second-order conditioning trials were drawn. For half of these trials, CS or US are randomly omitted (and the target valence updated accordingly) in order to prevent the networks from overgeneralizing to unconditioned CS. Optimization progressed for 5000 epochs for networks trained to perform extinction and second-order conditioning. For networks trained only for first-order conditioning, (Fig. 2E, top; Fig. 3), only first-order conditioning trials were drawn, and optimization progressed for 2000 epochs.

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Principal components of DAN activity (Fig. 4) were estimated using 50 randomly chosen trials of extinction and second-order conditioning in previously optimized networks. To order DANs based on their response similarity (Fig. 4A), hierarchical clustering was performed using the Euclidean distance between the vector of firing rates corresponding to pairs of DANs during these trials.

For networks also trained to report stimulus novelty (Supplemental Fig. 1), an additional readout dimension that is active for the first presentation of a given CS and inactive otherwise is added. Adding this additional readout does not significantly impact the performance of the networks for classical conditioning tasks.

Networks without DA-gated plasticity

For networks without DA-gated plasticity, KC-to-MBON synaptic weights are drawn randomly from a uniform distribution between 0 and 0.05 and then fixed. The time of CS+ presentation is chosen uniformly between 5 s and 15 s, and the second CS presentation occurrs uniformly between 20 s and 30 s. Networks are optimized to perform first-order conditioning with positive and negative valence US for a fixed set of CS+ stimuli numbering between 1 and 10 (2 to 20 possible associations). On half of the trials, a random CS is presented instead of the second CS+ presentation (Fig. 3B) and networks are optimized to not respond to this CS.

Continual learning

To model continual learning (Fig. 5), networks were augmented with non-specific potentiation gated by DAN activity according to Eq. 2. The potentiation parameter β is compartment-specific and updated through gradient descent. Each parameter is initialized at 0.01 and constrained to be positive.

Trials consist of 200 s intervals, during which two CS+ and two CS- odors are presented randomly. For each CS, the number of presentations in this interval is chosen from a Poisson distribution with a mean of 2 presentations. Unlike other networks, for these networks the values of $\mathbf{W}_{\mathrm{KC}\to\mathrm{MBON}}$ at the end of one trial are used as the initial condition for the next trial. To prevent weights from

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

saturating early in optimization, the weights at the beginning of trial t are set equal to:

$$w_t = (1 - x)w_0 + xw_{t-1}, \tag{4}$$

where $w_0 = 0.05$ corresponds to the initial weight at the beginning of optimization, and x increases linearly from 0 to 1 during the first 2500 epochs of optimization. Networks were optimized for a total of 5000 epochs.

Networks that encode changes in state

For networks that encode changes in state (Fig. 6), a three-dimensional readout of MBON activity is optimized to encode the state (at each moment in time, the target is equal to 1 readout dimension and 0 for the others). The external input $\mathbf{r}_{\text{ext}}$ is three-dimensional and signals state transitions using input pulses of length 2 s. The length of time between pulses ΔT_{state} is a random variable distributed according to $\Delta T_{\text{state}} \sim 10\text{ s} \cdot (1 + \text{Exp}(1))$. For these networks, we did not impose a penalty on DA neuron firing rates ($\alpha_{\text{DAN}} = 0$). Networks were optimized for 500 epochs.

To test how state-dependent DAN dynamics affect stimulus encoding, a CS is presented for 2 s, beginning 8 s prior to the second state change of a 300 s trial. Afterward, the same CS is presented for 5 s. This was repeated for 50 CS, and the correlation coefficient between MBON responses during the second 5 s presentation was calculated (Fig. 6C).

Models of navigation

To model navigation toward a rewarded odor source (Fig. 7), a CS+/US pairing is presented at $t = 2$ s in a 20 s training interval with a US strength of $\mathbf{r}_i^{\text{ext}} = 0.1$. This is followed by a 200 s interval during which the model organism navigates in a two-dimensional environment.

During navigation, two odor sources are present, one CS+ and one neutral CS. The sources are randomly placed at $x = \pm 1$ m and y chosen uniformly between 0 m and 2 m, with a minimum spacing of 0.5 m. Associated with each odor source is a wind stream that produces an odor plume that the model organism encounters as it navigates. These are assumed to be parallel to the x axis and oriented so that the odor plume diffuses toward the origin, with a height of 0.5 m and centered

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

on the y position of each odor source. For locations within these plumes and downwind of an odor source, the concentration of the odor is given by:

$$c(\Delta x, \Delta y) = \frac{1}{1 + 0.5\Delta x} \exp\left(-(\Delta y)^2/(0.1\Delta x)\right), \tag{5}$$

where Δx and Δy are the x and y displacements from the odor source in meters. This equation expresses a Gaussian odor plume with a width that increases and magnitude that decreases with distance from the odor source.

During navigation, when the model organism encounters an odor plume, KC activity is assumed to be proportional to the pattern of activity evoked by an odor (a random pattern that activates 10% of KCs) scaled by $c(\Delta x, \Delta y)$. The network further receives 4-dimensional wind direction input via $\mathbf{W}_{\text{ext}}$. Each input is given by $[\mathbf{w} \cdot \mathbf{h}_i]_+$, where $\mathbf{w}$ is a unit vector representing wind direction and $\mathbf{h}_i$ for $i = 1 \ldots 4$ is a unit vector pointing in the anterior, posterior, or lateral directions with respect to the model organism.

The organism is initially placed at the origin and at an angle distributed uniformly on The range $[\frac{\pi}{2}(1-\gamma), \frac{\pi}{2}(1+\gamma)]$, with γ increasing linearly from 0 to 0.5 during the optimization. The movement of the organism is given by two readouts of the FBNs. The first determines the forward velocity $v(t) = \text{Softplus}(\mathbf{W}_v \cdot \mathbf{r}(t) + b_v)$, and the second determines the angular velocity $\omega(t) = \mathbf{W}_\omega \cdot \mathbf{r}(t) + b_\omega$. The weights and bias parameters of these readouts are optimized using gradient descent. For these networks, we did not impose a penalty on DA neuron firing rates ($\alpha_{\text{DAN}} = 0$). For each trial, the loss is determined by the Euclidean distance of the model organism from the rewarded odor source at the end of the navigation interval. Networks were optimized for 500 epochs.

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgments

We wish to thank L. F. Abbott, R. Axel, V. Ruta, M. Zlatic, and A. Cardona for insightful discussions and comments on the manuscript. We are particularly grateful to L. F. Abbott for discussions during the development of this study. Research was supported by a Columbia University Class of 1939 Summer Research Fellowship (L. J.), the Columbia Science Research Fellows Program (L. J.), the Burroughs-Wellcome Foundation (A. L.-K.), the Simons Collaboration on the Global Brain (A. L.-K.), the Gatsby Charitable Foundation (L. J. and A. L.-K.), and NSF NeuroNex Award DBI-1707398 (L. J. and A. L.-K.).

References

Aso, Y. and Rubin, G.M. (2016). Dopaminergic neurons write and update memories with cell-type-specific rules. eLife *5*, e16135.

Aso, Y., et al. (2010). Specific dopaminergic neurons for the formation of labile aversive memory. Current Biology *20*, 1445–1451.

Aso, Y., et al. (2012). Three dopamine pathways induce aversive odor memories with different stability. PLOS Genetics *8*, e1002768.

Aso, Y., et al. (2014a). Mushroom body output neurons encode valence and guide memory-based action selection in Drosophila. eLife *3*, e04580.

Aso, Y., et al. (2014b). The neuronal architecture of the mushroom body provides a logic for associative learning. eLife *3*, e04577.

Ba, J., Hinton, G.E., Mnih, V., Leibo, J.Z., and Ionescu, C., Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems*, volume 29, 4331–4339 (2016).

Bargmann, C.I. and Marder, E. (2013). From the connectome to brain function. Nature Methods *10*, 483–490.

Barto, A.G., Adaptive critics and the basal ganglia. In *Models of Information Processing in the Basal Ganglia*, Computational Neuroscience, 215–232 (The MIT Press, Cambridge, MA, 1995).

Bromberg-Martin, E.S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in motivational control: Rewarding, aversive, and alerting. Neuron *68*, 815–834.

Burke, C.J., et al. (2012). Layered reward signalling through octopamine and dopamine in Drosophila. Nature *492*, 433–437.

Cohn, R., Morantte, I., and Ruta, V. (2015). Coordinated and compartmentalized neuromodulation

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

shapes sensory processing in {Drosophila}. Cell *163*, 1742–1755.

Eichler, K., et al. (2017). The complete connectome of a learning and memory centre in an insect brain. Nature *548*, 175–182.

Engelhard, B., et al. (2019). Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons. Nature *570*, 509.

Eschbach, C., et al. (2019). Multilevel feedback architecture for adaptive regulation of learning in the insect brain. bioRxiv 649731.

Felsenberg, J., Barnstedt, O., Cognigni, P., Lin, S., and Waddell, S. (2017). Re-evaluation of learned information in Drosophila. Nature *544*, 240–244.

Felsenberg, J., et al. (2018). Integration of parallel opposing memories underlies memory extinction. Cell *175*, 709–722.

Finn, C., Abbeel, P., and Levine, S., Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 1126–1135 (2017).

Fusi, S. and Abbott, L.F. (2007). Limits on the memory storage capacity of bounded synapses. Nature Neuroscience *10*, 485–493.

Gardner, M.P.H., Schoenbaum, G., and Gershman, S.J. (2018). Rethinking dopamine as generalized prediction error. Proceedings. Biological Sciences *285*.

Gaudry, Q., Nagel, K.I., and Wilson, R.I. (2012). Smelling on the fly: Sensory cues and strategies for olfactory navigation in Drosophila. Current Opinion in Neurobiology *22*, 216–222.

Handler, A., et al. (2019). Distinct dopamine receptor pathways underlie the temporal sensitivity of associative learning. Cell *178*, 60–75.e19.

Hattori, D., et al. (2017). Representations of novelty and familiarity in a mushroom body compartment. Cell *169*, 956–969.e17.

Hige, T., Aso, Y., Modi, M.N., Rubin, G.M., and Turner, G.C. (2015a). Heterosynaptic plasticity underlies aversive olfactory learning in Drosophila. Neuron *88*, 985–998.

Hige, T., Aso, Y., Rubin, G.M., and Turner, G.C. (2015b). Plasticity-driven individualization of olfactory coding in mushroom body output neurons. Nature *526*, 258–262.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences *79*, 2554–2558.

Horvitz, J.C., Stewart, T., and Jacobs, B.L. (1997). Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. Brain Research *759*, 251–258.

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Howe, M.W. and Dombeck, D.A. (2016). Rapid signalling in distinct dopaminergic axons during locomotion and reward. Nature *535*, 505–510.

Ito, M., Sakurai, M., and Tongroach, P. (1982). Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar purkinje cells. Journal of Physiology *324*, 113–134.

Kim, Y.C., Lee, H.G., and Han, K.A. (2007). D1 dopamine receptor DDA1 is required in the mushroom body neurons for aversive and appetitive learning in Drosophila. Journal of Neuroscience *27*, 7640–7647.

Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences *114*, 3521–3526.

Krashes, M.J., et al. (2009). A neural circuit mechanism integrating motivational state with memory expression in Drosophila. Cell *139*, 416–427.

Lak, A., Stauffer, W.R., and Schultz, W. (2016). Dopamine neurons learn relative chosen value from probabilistic rewards. eLife *5*, e18044.

Lau, B., Monteiro, T., and Paton, J.J. (2017). The many worlds hypothesis of dopamine prediction error: Implications of a parallel circuit architecture in the basal ganglia. Current Opinion in Neurobiology *46*, 241–247.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. Journal of Neurophysiology *67*, 145–163.

Matsumoto, M. and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature *459*, 837–841.

Menegas, W., Akiti, K., Amo, R., Uchida, N., and Watabe-Uchida, M. (2018). Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. Nature Neuroscience *21*, 1421–1430.

Menegas, W., Babayan, B.M., Uchida, N., and Watabe-Uchida, M. (2017). Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. eLife *6*, e21886.

Miconi, T., Clune, J., and Stanley, K.O. (2018). Differentiable plasticity: Training plastic neural networks with backpropagation. arXiv:1804.02464 .

Orhan, A.E. and Ma, W.J. (2019). A diverse range of factors affect the nature of neural representations underlying short-term memory. Nature Neuroscience *22*, 275.

Perisse, E., Burke, C., Huetteroth, W., and Waddell, S. (2013). Shocking revelations and saccharin sweetness in the study of Drosophila olfactory memory. Current Biology *23*, R752–R763.

bioRxiv preprint doi: https://doi.org/10.1101/737064; this version posted August 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rebec, G.V., Christensen, J.R.C., Guerra, C., and Bardo, M.T. (1997). Regional and temporal differences in real-time dopamine efflux in the nucleus accumbens during free-choice novelty. Brain Research *776*, 61–67.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Schwaerzel, M., et al. (2003). Dopamine and octopamine differentiate between aversive and appetitive olfactory memories in Drosophila. Journal of Neuroscience *23*, 10495–10502.

Steinfels, G.F., Heym, J., Strecker, R.E., and Jacobs, B.L. (1983). Behavioral correlates of dopaminergic unit activity in freely moving cats. Brain Research *258*, 217–228.

Sutton, R.S. and Barto, A.G., *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).

Suver, M.P., et al. (2019). Encoding of wind direction by central neurons in Drosophila. Neuron *102*, 828–842.e7.

Takemura, S., et al. (2017). A connectome of a learning and memory center in the adult Drosophila brain. eLife *6*, e26975.

Trannoy, S., Redt-Clouet, C., Dura, J.M., and Preat, T. (2011). Parallel processing of appetitive short- and long-term memories in Drosophila. Current Biology *21*, 1647–1653.

Wang, J.X., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. Nature Neuroscience *21*, 860–868.

Watabe-Uchida, M., Eshel, N., and Uchida, N. (2017). Neural circuitry of reward prediction error. Annual Review of Neuroscience *40*, 373–394.

Watabe-Uchida, M. and Uchida, N. (2019). Multiple dopamine systems: Weal and woe of dopamine. Cold Spring Harbor Symposia on Quantitative Biology 037648.

Yamins, D.L.K. and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience *19*, 356–365.

Zador, A.M. (2019). A critique of pure learning: What artificial neural networks can learn from animal brains. bioRxiv 582643.

Zheng, Z., et al. (2018). A complete electron microscopy volume of the brain of adult Drosophila melanogaster. Cell *174*, 730–743.