Expectation-based learning rules underlie dynamic foraging in *Drosophila*

Adithya E. Rajagopalan [1,2], Ran Darshan [1], James E. Fitzgerald [1], Glenn C. Turner [1,]*

1 - Howard Hughes Medical Institute, Janelia Research Campus, Ashburn, VA 20147, USA
2 - The Solomon H. Snyder Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
** Correspondence : turnerg@hhmi.org*

Abstract

Foraging animals must use decision-making strategies that dynamically account for uncertainty in the world. To cope with this uncertainty, animals have developed strikingly convergent strategies that use information about multiple past choices and reward to learn representations of the current state of the world. However, the underlying learning rules that drive the required learning have remained unclear. Here, working in the relatively simple nervous system of Drosophila, we combine behavioral measurements, mathematical modeling, and neural circuit perturbations to show that dynamic foraging depends on a learning rule incorporating reward expectation. Using a novel olfactory dynamic foraging task, we characterize the behavioral strategies used by individual flies when faced with unpredictable rewards and show, for the first time, that they perform operant matching. We build on past theoretical work and demonstrate that this strategy requires the existence of a covariance-based learning rule in the mushroom body - a hub for learning in the fly. In particular, the behavioral consequences of optogenetic perturbation experiments suggest that this learning rule incorporates reward expectation. Our results identify a key element of the algorithm underlying dynamic foraging in flies and suggest a comprehensive mechanism that could be fundamental to these behaviors across species.

Introduction

In natural environments, animals often have to accumulate rewards by making choices between options that are unreliable predictors of these rewards[1,2]. Studies in both vertebrates and invertebrates have found that animals employ decision-making strategies that account for this unpredictability, allowing them to maximize reward accumulation in a variety of dynamic and unpredictable environments[3–11]. Three key features are typically thought to be involved in the successful strategies that have been identified: i) a representation of information relevant for making choices, such as option value; ii) a tally of past choices and rewards; and iii) a learning rule by which these two are combined to update the decision variables based on experience[10,12–14]. Signatures of option value have been identified in a variety of brain structures and can serve as estimates of the reward probabilities associated with different options[5,10,15–22]. This can then be used to guide an animal's decisions. At the same time, neural recordings and phenomenological models of behavior have provided evidence that animals tally reward history in the form of a reward expectation[6,10,23–27]. For example, in the dopaminergic neurons in the vertebrate midbrain reward expectations are implicitly encoded by reward prediction error (RPE)[23,28–30]. On the other hand, because of the complexity of the animal models used to study foraging behavior, the learning rules that connect these option value and reward history representations have been challenging to pinpoint, and so a complete understanding of how these computations are carried out to produce particular behavioral strategies has yet to arise.

The fruit fly, *Drosophila melanogaster*, is equipped with great anatomical and behavioral tools with which to dissect the computations that could underlie such behavioral strategies. On the anatomical front, a long history of research in genetics has allowed for the targeted manipulation of neurons with unparalleled specificity[31,32]. On the behavioral front, research over the past several decades has transformed the fruit fly into a key model organism for the study of associative learning and choice behavior[33–35]. A rich body of experiments has shown that flies can learn a variety of Pavlovian conditioning paradigms[36–42] and exhibit features common in vertebrate memory, such as forgetting, re-evaluation, and consolidation, amongst others[43–50]. However, most existing tasks rely on studying the place preference behavior of groups of flies. This limits the ability of the experimenter to provide reward contingent on actions of any one fly, making it hard to study behavior in response to probabilistic reward or measure choice distributions over time (but see[51–55]).

In this study we develop a novel olfactory decision-making task for individual *Drosophila melanogaster*, inspired by tasks in vertebrates[5,10,21]. It allows us to explore the behavior of flies in more complex environments and study the decision-making strategy that the fly brain uses to cope with reward uncertainty. First, we show that flies can learn associations between cues and probabilistic rewards. We find that they are able to learn multiple probabilistic cue-reward associations simultaneously and distinguish between different reward probabilities. We also find that the decisions made by flies depend on the history of rewards and choices over multiple trials. We build on these results by developing a paradigm in which the probabilistic cue-reward associations change over time. Using this paradigm we show for the first time that flies follow the operant matching law when faced with such dynamically varying environments. Operant

matching is a common strategy observed across species, and it leads animals to divide their choices between options in proportion to the ratio of rewards received from each option[3–6,10,21,56–58].

Our discovery of operant matching in flies provides unprecedented opportunities for studying the synaptic learning mechanisms of this canonical animal behavior. This is particularly promising as insightful theoretical work from Loewenstein and Seung already pinpointed the features of learning that are necessary for matching[59]. This theory specifically predicts that decision variables must be updated by a learning rule that incorporates the covariance between reward and relevant neural activity. The long history of studying olfactory learning in the fly suggests that the behaviors we observe in this task depend on the mushroom body (MB) - a key center for learning and memory in the fly[33–35,60]. We build on the work of Lowenstein and Seung and develop an MB-inspired model of matching behavior that incorporates various learning rules and well-known MB circuitry. We use this model to generate de novo simulated behavior as well as fit fly behavior. We find that the covariance learning rules suggested by Loewenstein and Seung were more consistent with the observed sequential choice behavior than non-covariance learning rules. This theoretical framework suggests that reward expectation might modulate learning in the MB. We test this by manipulating the way reward is delivered to the flies and provide behavioral evidence that is indeed the case. These experiments also hint that this reward expectation might be represented in the activity of dopaminergic neurons (DANs).

Results

Individual flies show learning in a olfactory dynamic foraging task

To study decision-making strategies in dynamic environments we designed an olfactory Y-arena for individual flies (Fig. 1A; Methods; Supp. Info. 1). This design was inspired by earlier related assays for flies[52–54], and foraging tasks in vertebrates[5,10,16,21]. In our Y-arena, a single fly begins a trial in an arm filled with clean air and can choose between two odor cues that are randomly assigned to the other two arms (see Supp. Fig. 1 for estimates of odor boundaries). The fly can freely enter, exit and re-enter each arm, and a choice is determined when a fly enters far enough into an odorized arm to cross into the reward zone (Fig. 1A). Once a decision is made, we optogenetically activate reward-related neurons according to predefined reward contingencies. The Y-arena then resets, with the arm chosen by the fly filled with clean air and the other two arms randomly filled with the two odors. This novel task design provides us with the power to repeatedly monitor the choices of a *single* fly between predictors of reward (odors) over multiple trials, setting us up nicely to study the choice strategies used by individual flies.

We first probed fly behavior in a paradigm where odor cues were paired with reliable rewards. Flies were first presented with two unpaired odors - 3-octanol (OCT) and 4-methylcyclohexanol (MCH) - and showed no particular preference for one over the other (Fig. 1B naive). We investigated the trajectories taken by flies by plotting the displacement of the fly from center as a function of time (Fig. 1C). Flies spent a lot of time in the air arm and made variable choices with no obvious preference for either odor (Fig. 1C left, example fly). We then modified reward contingencies by rewarding one of the choices (100:0 paradigm) with activation of Gr64f receptor expressing sugar sensory neurons. These neurons have been shown to be necessary for detection of sugar in Drosophila and have been found to be capable of inducing cue-reward associations[61–63]. Flies changed their behavior following the switch in reward contingencies and were more likely to choose the rewarded odor (Fig. 1B rewarded; Wilcoxon signed-rank test: $p = 0.0039$, $n = 9$; Fig. 1C right). The averaged performance across flies can be plotted as a function of time to more clearly observe the change in preference over time (Fig. 1D). We also observed that choices became faster the more familiar flies became with the task (Fig. 1E; Wilcoxon signed-rank test: $p = 6.292*10^{-4}$, $n = 18$). This behavior change manifests in fly movement as a reduction in long meandering trajectories and an increase in quick excursions towards the rewarded odor (example fly in Fig. 1C).

To study these changes in fly behavior in more detail, we analyzed flies' choices to walk into (approach) or turn away from (reversal) an odor upon experiencing it (Methods). We specifically asked how this aspect of behavior varied between rewarded and unrewarded odors in our task. Flies were around four times more likely to turn away from unrewarded odors than they were to walk into them (Fig. 1F; Mann-Whitney rank-sum test: $p = 3.2278 * 10^{-7}$, $n = 18$). In contrast, flies were much more likely to walk toward the rewarded odor (Fig. 1F; Mann-Whitney rank-sum test: $p= 3.2166 * 10^{-7}$, $n = 18$). Turns away from the rewarded odor decreased with time (Fig. 1G; Wilcoxon signed-rank test: $p =0.0171$, $n = 18$, comparing bins 15-65 with 66-115 minutes to exclude late times when many flies had already finished the task), while turns away from the

unrewarded odor stayed constant (Fig. 1G; Wilcoxon signed-rank test: $p = 0.1839$, $n = 18$, same as rewarded odor). Interestingly, most turns made by flies occurred partway up the arm (Fig. 1H), suggesting that it takes flies a little time to make and commit to their decision to turn away. Changes in turn frequency over time (Fig. 1G) can also be observed at the level of single flies (Fig. 1I). In agreement with the data averaged across flies (Fig. 1G), the example fly became less likely to turn when experiencing the rewarded odor and remained highly likely to turn when experiencing the unrewarded odor.

Flies can learn multiple probabilistic cue-reward associations simultaneously

We next asked how fly behavioral strategies changed if the odor cues were linked with unreliable rewards. We incorporated reward baiting into our probabilistic reward task[5,10], which means that rewards probabilistically become available and then persist until the reward is foraged with a correct choice (Methods). We began with experiments in which a single odor was rewarded with one of several different probabilities - 1, 0.8 or 0.4. Flies showed a preference towards the rewarded odor in all cases (Fig. 2A; Mann-Whitney rank-sum test: probability = 1, $p = 4.4500*10^{-8}$, $n = 18$; probability= 0.8, $p = 5.8927*10^{-5}$, $n = 10$; probability = 0.4, $p = 0.0014$, $n = 10$). The extent of the preference varied with the probability of reward - a higher probability of reward led flies to show a stronger preference. In a different set of experiments where both odors were paired with probabilistic reward, we compared trial durations between experiments where the sum of reward probabilities associated with both options was 1 or 0.5. We found that flies made faster choices in situations where rewards were more probable (Fig. 2B; Mann-Whitney rank-sum test: $p = 4.7653*10^{-4}$, $n = 18$ for summed probability = 1; $n = 20$ for summed probability = 0.5).

Since having multiple cues linked with probabilistic rewards is important for the study of decision-making strategies, we designed a paradigm that incorporated 3 odors: OCT, MCH and ethyl lactate (EL) (Fig. 2C top). Flies first made 80 unrewarded choices (naive). This consisted of 40 choices between OCT and EL and 40 choices between MCH and EL. In the next 80 (reward) trials, one of OCT or MCH was assigned to be the high rewarding odor with a reward probability of 0.8 and the other the low rewarding odor with a reward probability of 0.4. EL was always unrewarded. OCT versus EL and MCH versus EL trials were interleaved over both trial blocks to ensure both relationships would be learnt simultaneously. After pairing, flies tended to prefer the rewarded odors, both for cases of high and low rewarding odors (Fig. 2C bottom; Wilcoxon signed-rank test: $p = 1.7962*10^{-4}$ for high rewarding odor; $n = 10$ $p = 0.0124$ for low rewarding odor, $n = 10$).

This choice preference was also reflected in the turning behavior (Fig. 2D). The turning behavior in the high reward versus unrewarded trials were very similar to that observed in the experiments with a single reliable predictor of reward (Fig. 1F), with flies exhibiting a clear preference for walking into the rewarding odor as compared to turning away (Fig. 2D right; Mann-Whitney rank-sum test: $p= 1.8267*10^{-4}$, $n = 10$) and vice-versa for the unrewarded odor (Fig. 2D right; Mann-Whitney rank-sum test: $p= 1.8267*10^{-4}$, $n = 10$). Flies also exhibited a statistically insignificant trend towards decreased turning away when experiencing the highly

rewarded odor in the rewarded block compared to the naive block (Fig. 2D right; Mann-Whitney rank-sum test: p= 0.089, n = 10).

Interestingly, in the low reward versus unrewarded trials, flies showed an increased percentage of turning away from both odors when compared to naive trials (Fig. 2D left; Mann-Whitney rank-sum test: unrewarded odor: $p = 1.8165*10^{-4}$, n = 10; rewarded odor: $p = 7.6854*10^{-4}$, n = 10). However, flies were more likely to turn away from the unrewarded odor than the low reward option (Fig. 2D left; Mann-Whitney rank-sum test: p = 0.0172, n =10). This highlights the possibility that the flies retain information about the high rewarding option even when it is unavailable to them. The increased turning might be the result of a local search strategy to identify the location of this high rewarding odor.

Flies perform operant matching

The final element needed to study behavioral strategies in a dynamic environment is to test the fly's ability to make decisions in the face of changing reward contingencies. We did this in a paradigm in which single flies made 240 choices, divided into three blocks of 80 trials each. In each block, flies made choices between two odors (OCT and MCH) that were paired with different probabilities of reward (Methods). These probabilities remained fixed within a block and changed across blocks (Fig. 3A, example reward probabilities and fly behavior). When faced with tasks like the one just described, a wide variety of animals have been found to approximately follow the operant matching law[3–6,10,21,56–58]. This strategy produces behavior in which animals divide their choices between options (choice ratio) in proportion to the ratio of rewards received from each option (reward ratio).

We calculated the instantaneous choice ratio in a rolling 10 trial window (Fig. 3A, blue) and found that this ratio was correlated to the reward ratio received by the fly in the same window (Fig. 3A, black). Another way of visualizing this data is to look at the cumulative number of choices made towards each option and plotting them against each other (Fig. 3B). The slope of the curves (blue) is a measure of the instantaneous choice ratio, which can be overlaid with the block average reward ratios by plotting lines (black) with slope equal to the reward ratio (Fig. 3B, same example fly in 3A). The slopes of the choice ratio curves and reward ratio lines were very similar after an initial lag following a block change, further supporting the idea that *Drosophila* matched their choice ratio with their reward ratio. To show that this holds true across flies, we compared the block averaged choice ratios against their respective reward ratios (Fig. 3C). In such a plot the matching law predicts that all points will fall along the unity line (a line with slope equal to one). We found that flies appear to approximately follow the matching law, with a slight amount of undermatching. This is signified by a decrease in the slope of the choice ratio v reward ratio plot from unity. This deviation is commonly observed across species[5,6,10,21,58], and several reasons have been suggested for why animals tend to slightly undermatch[58,64] (see Discussion).

It took flies, on average, around 15-20 trials to reach a new steady state choice behavior following block switches (Fig. 3D). This timescale indicated that flies might be basing decisions

on information obtained from multiple past trials, rather than just the most recent experience. To test this hypothesis, we first looked at the decision made by flies on a given choice as a function of their past choices and outcomes. We focussed on choices following particular triplets of past experiences (Fig. 3E). For example, on trials following three unrewarded choices towards the green odor, flies tended to choose between options roughly at random (Fig. 3E left - top). However, on trials where the green option was rewarded on the most recent (Fig. 3E left - middle) or more distant trials (Fig. 3E - bottom) trials, the flies' choices were biased towards the green option. Other examples show that the flies' tendency to switch back to the green odor after a single unrewarded orange trial depends on how green was rewarded in the recent past (Fig. 3E - right).

There are many such patterns one could analyze. Instead, we systematically estimated the contribution of reward and choice histories on sequential choice behavior by fitting logistic regression models to our data, with reward and choice histories as predictors. This analysis supported our hypothesis that reward information from multiple trials in the past was relevant to flies when making a choice (Fig. 3F top). Rewards on more recent trials had higher regression coefficients and were more important than distant rewards, but all 15 reward coefficients were significant (Student's t-test: *** - $p < 0.001$, ** - $p < 0.01$, * - $p<0.05$, n = estimates from 21 flies) . On the other hand, the dependence of present choice on choice history seemed to be limited to a shorter history of 5 past choices (Fig. 3F bottom). This suggests that the fly might exhibit a small amount of stickiness to previously chosen odors in its choice behavior. To further confirm that reward and choice histories of more than one trial in the past influenced choice, we predicted fly data with an intuitive 'Hit-stay Miss-switch' model that used only the most recent trial's information to determine choice (Methods). We compared the predictions from this model with predictions of our logistic regression model and a standard Q-learning model that tallies action values over trials and is commonly used to fit choice data of this type (Methods). The two models that made use of multiple trials worth of information were better fits for the behavior (Fig. 3G). An example of such a fit, using the Q-learning model to predict fly choice behavior is shown in Fig. 3H (example fly - blue; fit - purple).

A mushroom body inspired model of matching behavior

The analyses of behaviors in the novel olfactory Y-arena showed that *Drosophila* use a canonical operant matching strategy to make choices in dynamically varying environments. To probe the mechanistic basis of this behavior, we turned to a theory developed by Loewenstein & Seung that constrains the synaptic learning rules underlying matching behavior[59]. The theory assumes that decision variables are represented in the form of synaptic weights. The authors proved a theorem that shows that a synaptic plasticity rule that modifies these weights according to covariation of reward and neural activity is necessary and sufficient to produce matching. Such a rule modifies synaptic weights proportional to the product of the involved neural activity and reward, with at least one of the terms being mean subtracted (Fig. 4A - box). This allows for bi-directional plasticity depending on the outcome of the choice and leads to a behavior that obeys the matching law.

Additionally, Loewenstein & Seung used a simple neural circuit model to demonstrate that their theory holds true (Fig. 4A). In this model, the cues that the animal has to choose between are represented by two sensory neurons (S_1 and S_2; Fig. 4A). These neurons synapse onto two motor neurons, whose activity is compared to determine action via a winner-take-all computation (A_1 and A_2; Fig. 4A). The value associated with each option, therefore, is represented in the weight of the synapses between the sensory and motor neurons (and hence in the activity of A_1 and A_2). This can change over time according to the learning rule (Fig. 4A - box). Matching behavior can only arise in this circuit when synaptic weights are updated according to a learning rule that incorporates the covariation of reward and neural activity[59] (Supp. Fig. 2A-D).

Importantly, in the context of the fly, the mushroom body (MB), contains all the features deemed to be necessary for matching according to Loewenstein & Seung's theory (Fig. 4B). It is a key center for learning and memory in the fly brain that likely plays a role in the behavioral strategies we have observed[33–35,60]. The architecture of the MB shares many similarities with the model circuit in Figure 4A. The odors that the fly has to choose between are represented by the kenyon cells (KC) - with different odors activating different sub-populations of the entire set of the KCs[44,65–67]. The mushroom body output neurons (MBONs) receive input from these KCs and are thought to represent valence that guides action[18,19]. There are multiple MBONs, each innervating a particular compartment of the KC axon bundle[68]. Crucially, the synapses between the KC and the MBONs are thought to represent the values associated with each odor and are modified by a reward-based learning rule that requires the coincident activity of odor representing KCs and release of dopamine by reward-related dopaminergic neurons (DANs)[19,20,46,69–75] (Fig. 4B - box).

These remarkable similarities between the model in Figure 4A and the architecture of the MB circuit (Fig. 4B) allowed us to modify the L&S model, tailoring it to our task structure and to the MB circuitry. In the original model (Fig. 4A), both stimuli were present at the time of choice and reward. In our task, at any given time, the fly experienced only one of the two odor cues as it explored the Y-arena. This means that only one of the odors was present when the fly walked the reward zone and the reward was either provided or omitted (Fig 4C top). Differences with the model described by Loewenstein & Seung (Fig. 4A) also arise at the circuit level. Rather than two competing output neurons, a single compartment of the MB only contains a single output neuron that mediates the resulting behaviors, such as reversal or approach[18,19,68]. Therefore, instead of comparing the activity of two output neurons to determine choice, in the model, the activity of an MBON should represent the likelihood to turn away from or move towards the odor source (Fig. 4C bottom). Furthermore, to account for possible stochasticity in this process, we replaced the winner-take-all competition in the original model with a sigmoidal nonlinearity to determine the behavior (Fig. 4C; Methods). Finally, rather than two single input neurons, odors are represented by populations of KCs (Fig. 4C). These population responses can be noisy, and could have some amount of overlap between odors. We thus parameterized our input representation to incorporate these features of KC odor responses (Methods).

We modified the model to incorporate the relevant new features of MB circuitry and our decision making task and simulated the model with different learning rules (Fig. 4C). All three covariance-based rules that produced matching previously continued to produce matching in our modified model (Fig. 4D; Supp Fig. 2 B-F). Consistent with the theory of Loewenstein & Seung, a rule that does not incorporate either reward or stimulus expectation did not produce matching (Fig. 4E; Supp Fig. 2A).

To confirm that covariance-based learning rules could also replicate the trial-by-trial behavior of flies, we simulated data using all 4 rules in the paradigm shown in Fig 3A. Covariance-based rules (Fig. 4F; Supp Fig. 3A-C) produced behavior that qualitatively resembled fly behavior. The choice behavior resulting from these rules, tracked changes in the reward contingencies across blocks, with the resulting instantaneous choice ratio being biased towards the more rewarded option in each block. In contrast, simulations with the non-covariance rule were dissimilar to fly behavior (Supp. Fig. 3A,D), with the instantaneous choice ratio and reward ratio curves deviating from each other. In particular, choices were made roughly equally to both options throughout . Combining our observation of operant matching in *Drosophila* with the behavior of the simulated MB-inspired model in the context of Loewenstein and Seung's theory, we hypothesized that flies made use of a covariance-based learning rule to generate their learned behaviors.

Identifying learning rules underlying dynamic foraging in the mushroom body

We next tested more systematically if, in addition to producing matching at a global level, covariance-based rules were also quantitatively better predictors of choice on a trial-by-trial basis. To do this, we designed a logistic regression model (Fig. 5A) that used the sensory inputs experienced by the fly, $S_i(t)$, and learned synaptic weights, $W_i(t)$, to calculate a correlate of MBON activity, $\widehat{A} = \Sigma_i W_i(t) \cdot S_i(t)$. As MBON activity guides behavior, the output of this computation, $\widehat{A}$, predicts the probability of each action. This model incorporated a learning rule, $\Delta W_i(t)$, that changed the synaptic weights whenever the animal chose an odor (Fig. 5A - right). The rule consisted of four possible components that may contribute to the learning rule at the KC-MBON synapse (Fig. 4B, 5A) - a constant term, stimulus, reward, and their combination. It also consisted of a bias term that is related to the initial value of the weights. The stimulus and reward variables could be subtracted by their respective expectation values or not. As the learning rule at this synapse is known to be independent of the activity of the MBON itself[35], no terms related to MBON activity were included. The regression would then identify which of these terms was most important to best predict behavior. By building models that incorporated these expectations in various combinations, we could use this regression model to investigate if covariance-based rules were better at predicting sequential choice behavior than a non-covariance-based rule. This would further support the hypothesis that flies were making use of a covariance-based learning rule in the task. Furthermore, we could also ask if one of the covariance-based rules was a better fit for fly behavior than the others.

We first verified the feasibility of the approach in simulations that were well-matched to the details of the regression model. In Fig. 5B, the example simulated data (blue) was produced by a model using a covariance learning rule that incorporated reward expectation via an $[R - E(R)]$ term. Here, as in Fig. 1I, the probability of approach is plotted as a function of time and calculated separately for each odor. As expected, a regression that incorporated reward expectation fit the data better than one that did not (Fig. 5B). We verified that this was the case for all 50 simulations by comparing the log-likelihoods between the regression models that used the covariance rule with reward expectation and the non-covariance rule (Supp. Fig. 4A, Wilcoxon signed-rank test : $p = 6.7595 * 10^{-9}$). Moreover, across 50 example simulations, the largest regression coefficients were assigned to the $S_i \cdot [R - E(R)]$ term in the learning rule (Fig. 5C). Note here, that the model does assign non zero values to the $[R - E(R)]$ term as well. These terms are correlated and require a lot of data to dissociate, suggesting a potential limitation when applying this model to fly behavior (see also Fig. 5G).

The MB model was simulated assuming a certain amount of overlap of KC representations to calculate sensory inputs, and the fitting procedure likewise used this parameter when inferring the learning rule. Additionally noise was added to these representations (Methods). Since the overlap is imprecisely specified biologically, we verified that the regression model was able to similarly fit the simulated data when the parameter fed into the regression was shifted from the true value (Supp. Fig. 4B). We also found that the extent of matching produced by the model was unaffected by the amount of overlap (Supp. Fig. 4C-F). Given these findings, we assumed for simplicity that the overlap was 0 for all future simulations and when fitting regression models to fly data.

We applied our approach to fit data from flies performing the behavioral paradigm introduced in Fig. 3. We chose this task over the paradigms in Figs. 1-2 because it produced a larger number of trials per fly and thus provided more data from which the regression model could decipher the learning rule. Regressions that utilized covariance-based learning rules usually tracked fly behavior better, as can be seen in the example shown (Fig. 5D, Supp. Fig. 5B). We also systematically compared the fits for every fly that resulted from the three covariance rules with fits from the non-covariance rule, by subtracting the log-likelihood measure from the covariance rules with that from the non-covariance rule (Fig. 5E, Supp. Fig 5A). The difference in log-likelihood between non-covariance and covariance models showed that covariance rules were better fits for fly behavior overall (Fig. 5E; Wilcoxon signed-rank test: p = 0.0386; Supp. Fig 5A; Wilcoxon signed-rank test: for $R \cdot [S_i - E(S_i)]$ rule : p = 0.04; for $[R - E(R)] \cdot [S_i - E(S_i)]$ rule : p=0.0386), although the regression using the non-covariance rule did produce a better fit for a few flies. However, no statistical differences could be observed between the various covariance-based rules (Fig. 5E & Supp. Fig 5A; Wilcoxon signed-rank test: $[R - E(R)] \cdot S_i$ & $[R - E(R) \cdot S_i - E(S_i)]$, p = 0.5202; $R \cdot [S_i - E(S_i)]$ & $[R - E(R)] \cdot [S_i - E(S_i)]$, p = 0.4330; $R \cdot [S_i - E(S_i)]$ & $[R - E(R)] \cdot S_i$, p = 0.4549).

We next examined the regression coefficients resulting from this analysis to assess which terms were important for learning. When a rule that incorporated reward expectation via an $[R - E(R)]$ term was used to fit fly behavior, we found that the best fit model parameters had a similar profile as when this rule was used to fit simulated data in Fig. 5B. The largest regression coefficients were assigned to the $S_i \cdot [R - E(R)]$ term in the learning rule with the model assigning non zero values to the $[R - E(R)]$ term as well (Fig. 5F). Yet, either of these parameters could be small, or even negative, in some flies. Other parameters were on average assigned values near zero (Fig. 5F). Across rules, the term that involved the product of sensory input and reward was given the most importance (Fig. 5F, Supp. Fig. 5B,D,F), which is in agreement with the nature of the KC-MBON synaptic plasticity rule. Together, these results further supported the idea that a covariance-based rule was implemented in the MB.

It should be noted, however, that the best-fit model parameters in all covariance-based rules varied across flies (Fig. 5F, Supp. Fig. 5B,D,F). To better understand the patterns of this variability, we calculated the correlation coefficients between the best-fit model parameters across the population of flies (Fig. 5G, Supp. Fig. 5C,E,G). This revealed that many pairs of parameters were strongly correlated or anti-correlated with each other. Such covariation is typical in statistical models when multiple model parameters values produce similar fits to the data. This finding also makes intuitive sense in the current setting because individual terms in the learning rule involve overlapping combinations of stimulus and reward that are hard to dissociate statistically. More interpretable parameters values may arise from statistical fits to larger data sets or alternate behavioral paradigms.

Behavioral evidence of reward expectation in DANs

While the regression analysis confirmed that covariance-based learning rules were better at predicting trial-by-trial behavior in flies over the non-covariance rule, it could not distinguish between the covariance-based rules. However the mathematical differences between the three rules suggested a way forward. In particular, the rules differ by which terms - sensory input or reward - incorporate an expectation. Experimental manipulations that prevent the circuit from incorporating sensory or reward expectations would therefore change the learning rule in some cases but not others. Here we leveraged the genetic tools available in the fly to manipulate the activity of DANs during the task. It allowed us to override the natural reward signal represented by the DAN with one that was, potentially, independent of reward expectation.

We specifically did this by providing reward via optogenetic activation of the reward-related protocerebral anterior medial (PAM) DANs. This would potentially activate the DANs to the same level whenever a reward is received, independent of whether it was expected or not. According to the theoretical framework we have adopted, this manipulation would affect the behavior if the following conditions were met: i) the task-relevant learning rule depended on the activity of the DANs and KCs; ii) the activity of DANs incorporated reward expectation under natural conditions; iii) the activity of KCs does not incorporate sensory expectation. If all three conditions are met, the manipulation of providing reward by direct PAM DAN activation would change the learning rule from a covariance-based rule to a non-covariance rule - resulting in

modified behavior. We focused on a task consisting of a two blocks (naive and reward) of 60 trials in which the reward ratio protocol in the reward block was either 100:0 (brown curves), or 80:20 (orange curves).

We first predicted how the behavior during the reward block would differ between covariance-based and non-covariance rules using simulations that incorporated all four learning rules. (Fig. 6A,B, Supp. Fig. 6). The behavior of the model with any covariance-based rule was similar to the fly behavior when it was rewarded using the sugar neurons (Fig. 6C). This agrees with our previous results and reinforces our hypothesis that one of these three rules is being used by the fly. When rewards were provided via DAN activation, behavior now looked strikingly like that produced by the non-covariance rule (Fig. 6D). The behavior produced with DAN activation reward led to much lower plateau performance in both tasks when compared to sugar sensory rewards (Fig. 6E,F; Wilcoxon signed-rank test: Gr64f - 100:0, $p = 0.0039$, $n = 9$; Gr64f - 80:20, $p = 0.0312$, $n = 6$; DAN - 100:0, $p = 0.0391$, $n = 8$; DAN - 80:20, $p = 0.1875$, $n = 8$). This reduced performance might seem surprising, but the saturation in performance, around 75% in the 100:0 experiments and 50% in the 80:20 experiments, were also predicted by the non-covariance learning rule model (Methods).

These results suggested that providing reward with DAN activation changed the learning rule from a covariance-based form to a non-covariance rule. It implied that flies use a covariance-based rule that incorporates reward expectations, but not sensory stimulus expectations. To examine this hypothesis quantitatively, we fit our MB-inspired regression models (Fig. 5A) to fly data produced with DAN reward (Fig. 6D). As expected, we found that the non-covariance rule was the best fit (Fig. 6G,H).

Next, we trained flies in the paradigm of Fig. 3 using DAN activation as reward, instead of activating the sugar sensory neurons. In contrast to the fly behavior with sugar sensory neurons activations, flies tend to match their choice and reward ratios much less (Fig. 6I,J). When plotting choice ratio against reward ratio (Fig. 6I), we found that the behavior was best fit by a line that had a much lower slope than before (Fig. 6I blue vs. red). The instantaneous choice ratio and reward ratio of an example fly (Fig. 6J) showed that this undermatching stems from an inability to produce behavior that was strongly biased towards the rewarded odor. This result is consistent with the hypothesis that DANs activation leads to a non-covariance learning rule (Fig. 4E). Indeed, matching was almost abolished in simulations that used a non-covariance rule in this 3-block paradigm (Fig. 6K, Supp. Fig. 3C). All together, our modeling and DAN activation experiments suggest that reward expectation is incorporated in learning rules that underlie matching behaviors in flies, possibly in the activity of MB DANs.

Discussion

We combined a novel behavioral task for individual fruit flies, with mathematical modeling and manipulations of underlying neural circuitry to gain insight into the decision-making strategies used by flies and their underlying learning rules. We showed, for the first time, that flies follow the operant matching law when making choices in dynamic and unreliable environments. Our results suggested that a covariance-based learning rule underlies this behavior. In particular, our findings indicated that such a rule requires the calculation of an expected reward signal and that this reward expectation is likely represented in the activity of the DANs. This finding provides a novel insight into the role of DANs in the MB.

Behavioral similarities between flies and other animals

The signatures of learnt behavior and decision-making strategies we observed in flies have a lot in common with behaviors observed in a variety of other animals. For example, we found that the choices made by flies depended on choice and reward information received over multiple past trials (Fig 3E-F). This is in agreement with what has been observed in mice and monkeys [6,10]. Most strikingly, this history dependence led flies to make use of an operant matching strategy where they divided their choices in proportion to the ratio of rewards they received from each option (Fig 3A-C). This strategy is common and has been observed in a variety of animals, including bees, pigeons, mice, primates, and humans [3–6,10,21,56–58]. Interestingly, none of these animals (including flies) perfectly follow the matching law; rather they appear to undermatch, characterized by points in a choice ratio versus reward ratio plot falling on a line with slope smaller than one[5,10,21,56,58]. Another key signature of learnt behavior in mice and monkeys is an increase in the speed of choice with increased reward probability [10,21]. We observed this dependency in flies as well (Fig. 2B). These striking similarities suggest that there could exist common rules underlying behavior in all of these animals.

Possible circuit mechanisms for calculating reward expectation

Our results suggest that flies can perform operant matching because reward expectation is incorporated into the learning rule present in the MB circuit. These results also support a model in which this calculation is implemented in the activity of the DANs.

A reward expectation signal is implicitly represented as a reward prediction error (RPE) in mammalian dopaminergic neurons and has been shown to play an important role in learnt behaviors [23,28–30].One potential mechanism suggests that DANs could represent reward expectation in its activity as an RPE, and involves a feedback loop from MBONs. In the MB circuit, MBONs have been shown to represent odor valence, which is related to the expectation of reward associated with a given odor[18,19]. An indirect feedback loop, via an inhibitory interneuron for example, could serve as the conduit of this reward expectation related information between the MBONs and DANs. In this scenario, a negative expected rewarded signal from the interneuron, and a positive reward signal from the sugar neuron could be combined at the level of the DANs. This would allow the DANs' activity to signal an RPE. Based

on the recently released hemi-brain connectome we know that many direct and indirect feedback connections between MBONs and DANs exist that could support such a computation [76]. Additionally, signatures of an RPE signal in DAN activity have been found in *Drosophila* larvae, further supporting this hypothetical mechanism [77].

However, such RPE-based mechanisms are not exactly in line with the framework of covariance-based learning. In this framework, reward expectation is a single stimulus-independent value, whereas RPE is stimulus dependent. Furthermore, recordings from punishment-related DANs in adult *Drosophila* have not found signals that resemble the classical RPE seen in mammals [78,79]. This has led several groups to propose alternative mechanisms by which reward and punishment expectations can be incorporated into the MB circuit without requiring that they be represented in the activity of DANs [48,60,80,81]. Further modeling efforts, and neural recordings targeted at reward related DANs, are required to provide clarity into the nature of DAN activity, its dependence on reward expectations, and the consequences for learning at the KC-MBON synapse.

Value learning in multiple mushroom body compartments could explain deviations from matching

We observed that flies, like several other animals, don't perfectly follow the matching law; rather they undermatch. Two hypotheses have been proposed to account for this deviation from the matching law. The first option suggests that animals might be making use of a learning rule that deviates from a covariance-based rule[64]. A second idea stems from the fact that animals learn over both short and long time scales[58]. This idea suggests that in a dynamic environment, past experiences act as a bias that prevents the animal from correctly estimating the present cue-reward relationships.

Interestingly, both these ideas can be explained within the architecture of the MB. Multiple MBONs, divided into different compartments, guide behavior, with each compartment being innervated by a unique DAN[68]. It is possible that some compartments make use of a covariance-based learning rule, while others do not. This might lead to undermatching. Moreover, past work has shown that the memory formed in these different compartments have different time scales and persist for different amounts of times[46]. Whether either of these mechanisms contribute to undermatching in flies can be studied in future experiments by manipulating different compartments of the MB circuitry and analyzing the effect of such a manipulation on undermatching.

Approaches for inferring learning rules from behavior

The inference of learning rules is important for many areas of neuroscience. Here we introduced a novel statistical method for using logistic regression to estimate learning rules from behavioral data, and we applied our approach to infer learning rules for the fly mushroom body. This methodology could be similarly applied to model other complex learned behaviors in the fly and other animals. While the current work considered learning rules that only depended on the

current sensory stimulus and reward, our methodology would straightforwardly generalize to the inference of longer-timescale learning rules that depend on the history of sensory input and reward. In contrast, the logistic regression formalization would break down entirely for learning rules that depend on the magnitudes of synaptic weights or post-synaptic activity, because these would induce different nonlinear dependencies between the choice sequence and learning rule parameters. Many biological learning rules depend on these quantities. Future work should explore more flexible methodology from modern machine learning to develop more generally applicable fitting routines.

The importance of synergy between experiments and theory for providing mechanistic explanations of complex behaviors

Our study speaks strongly to the importance of theory in understanding biological mechanisms. Here we provided behavioral evidence that learning in the mushroom body depends on reward expectations. This finding is inextricably linked to the theory put forward by Loewenstein and Seung, and their theoretical insights fundamentally shaped our experimental and modeling approach. Moreover, our mushroom body model incorporates the fundamental principles of the Loewenstein and Seung theory and makes circuit-level predictions. These predictions are experimentally accessible, and their verification would add tremendously to our understanding of learning in the fly. Even if the predictions are not supported by future studies, then one of several important conclusions necessarily follow that would also add to our understanding of fly learning. Indeed, perhaps the greatest utility of the Loewenstein and Seung theory is that its rigor, generality, and specificity permit a level of falsifiability that guarantees conceptual advance. Its clarity and significance provide a wonderful example of what theory should aspire to in biology.

Such insightful theory is best paired with an experimental system that can readily test the predictions put forward. In the context of neuroscience, recent advances in the study of *Drosophila* have made great strides in making the fly brain such a system to tackle mechanistic questions. Work from multiple groups have allowed for the production of a wiring diagram of the entire brain[82,83]. This can serve as an incredibly useful tool for understanding and predicting the neural circuitry underlying behavior. Furthermore, most neurons identified in the connectome can be specifically manipulated with the help of the various genetic tools available to the fly community [31,32]. However it has remained unclear if flies can perform many of the complex cognitive tasks studied in other animals. We show that flies are capable of performing complex decision making tasks, and in fact use some of the same strategies to solve these tasks as other animals do. Many of the features of behavior we have observed in flies, such as reward-history based decisions and an analog sense of reward probability, are key features underlying other complex behaviors such as risky decision making, the effect of cost or punishment on choice, and planning. This pairing of theoretical neuroscience with experiments in *Drosophila* promises to be important for generating mechanistic explanations of complex behaviors.

Materials & Methods

Fly strains and rearing:

Drosophila melanogaster used for behavior experiments were raised on standard cornmeal food supplemented with 0.2 mM all-trans-retinal at 25℃ [for Gr64f lines - see following table] or 21℃ [for other lines - see following table] with 60% relative humidity and kept in dark throughout. The details of all flies used for experiments in this manuscript can be found in the table below:

Genotype	**Expression target/reporter description**	**Bloomington stock number (reference)**
w; Gr64f-Gal4/CyO; Gr64f-Gal4/TM3 X 20XUAS-CsChrimson-mVenus attp18	Optogenetic activation of Gr64f expressing sugar sensory neurons in behavior experiments	(Haberkern et al. 2019)[63] X
w; +; 58E02-Gal4 X 20XUAS-CsChrimson-mVenus attp18	Optogenetic activation of PAM cluster DAN in behavior experiments	BDSC:41347(Jenett et al. 2012) [31] X

Cross progeny were sorted to identify the correct phenotypes by cold anesthesia of around 4℃ on a cold plate and females of the appropriate genotype were transferred to starvation vials. Starvation vials contained nutrient free 1% agarose to prevent desiccation. Flies were starved between 28 - 42 hrs before being aspirated into the Y-arena for experiments.

Y-arena:

All behavior experiments were performed in a novel olfactory Y-arena designed by the Janelia Experimental Technology team (jET).

Apparatus design:

A detailed description of the apparatus is provided in Supplementary Information 1. The Y arena shares many design features with a circular arena used for monitoring learning with groups of flies [18]. The Y chamber is mounted above an LED board that provides infrared illumination to monitor the fly's movements, and red (617nm) light for optogenetic activation (Supplementary Information 2). Each arm of the Y has a corresponding odor delivery system, capable of delivering up to 5 odors. Experiments were recorded from above the Y using a single USB3 camera (Flea3, model: FL3-U3-13E4M-C: 1.3 MP, 60 FPS, e2v EV76C560, Mono; Teledyne FLIR, with longpass filter of 800 nm).

The Y chamber consists of two layers of white opaque plastic. The bottom is a single continuous circular layer and serves as the floor of the Y that flies navigate. The top is a circular layer with a Y shaped hole in the middle. The length of each arm from center to tip is 5cm and the width of each arm is 1cm. These two layers are placed underneath an annulus of black aluminum. A transparent glass circle is located in the center of this annulus and acts as the ceiling of the Y - allowing for video recording of experiments. This transparent disk is rotatable and contains a small hole used to load flies. The black annulus houses three clamps that hold the circular disk in place. All three layers are held together and made airtight with the help of 12 screws that connect the layers.

The LED board consists of a square array of red (617 nm peak emission, Red-Orange LUXEON Rebel LED, 122 lm at 700mA) and infrared (IR) LEDs which shine through an acrylic diffuser to illuminate flies from below.

For our experiments olfactometers injected air/odor streams into each arm at a flow rate of 100 ml/min. A crisp boundary between odors and air is formed at the center of the Y (Supp. Fig. 1A). Odors and concentrations used for each experiment are detailed in the behavioral task structure and design section of the methods. The center of the Y contains an exhaust port connected to a vacuum which was set at 300 ml/min using a flow meter (Dwyer, Series VF Visi-Float® acrylic flowmeter) - matching total input flow in our experiments.

Fly tracking and operation:

We wrote custom MATLAB code (MATLAB 2018b, Mathworks) to control the Y-arena and run experiments. The data collected by the USB3 camera was loaded into MATLAB in real time and the fly's location was identified using the MATLAB image processing toolbox. A background image was calculated just before beginning the experiment by averaging multiple frames as the fly moved around in the Y. This background was subtracted from the frame being processed and the resulting image was thresholded, leaving the fly as a white shape on a black background. The location of the centroid of the fly was estimated and then assigned to 1 of 6 regions in the Y (Supp. Fig. 1B). If the fly was located in one of the choice zones, a trial was deemed complete and rewards were provided by switching on the red LEDs as defined by the reward contingencies of the task. The arena was then reset and odors and air were randomly reassigned to the three arms (Fig. 1B). The location of the fly along with other information such as reward presence and odor-arm assignments were saved as a .mat file for further analysis. All analysis in figures 1, 2 and 3 are based on this saved information.

Behavioral experiments:

Odorant information:

For all experiments in the paper, two or three of the following odorants were used to form cue-reward relationships :

1. 3-Octanol (OCT) [Sigma-Aldrich 218405] diluted in paraffin oil [Sigma-Aldrich 18512] at a 1:500 concentration and then air-diluted to a fourth of this concentration.

2. 4-Methyl-cyclo-hexanol (MCH) [Sigma-Aldrich 153095] diluted in paraffin oil [Sigma-Aldrich 18512] at a 1:500 concentration and then air-diluted to a fourth of this concentration.

3. Ethyl Lactate (EL) [Sigma-Aldrich W244015] diluted in paraffin oil [Sigma-Aldrich 18512] at a 1:5000 concentration and then air-diluted to a fourth of this concentration.

Behavioral task structure and design :

Tasks with reliable cue-reward relationships:

In the single consistently rewarded odor task analyzed in Fig. 1 and part of Fig. 6, flies were inserted randomly into one of the three arms. This arm was injected with a clean airstream and OCT and MCH were randomly assigned to the other two arms. For a given fly one of OCT or MCH was arbitrarily paired with reward 100% of the time. Once a fly reached the choice zone of either the odor arm a choice was considered to have been made. If the rewarded odor was chosen, the fly was rewarded with a 500ms flash of red (617 nm) LED to activate the appropriate reward-related neurons. The arena was then reset with the arm chosen by the fly injected with clean air and OCT and MCH randomly assigned to the other two arms. This was repeated for several trials. In the version of this task seen in Fig. 1C, flies were allowed to choose between OCT and MCH for 30 minutes when neither option was rewarding and then for 30 minutes when one of the options was consistently rewarding. In the version of this task used for analysis in Fig. 1D, 1I and Fig. 7, flies made 60 naive choices where neither option was rewarding and 60 rewarded choices where one option was consistently rewarded. In the version used for analysis in Figs. 1E-H, flies made 120 rewarded choices where one option was consistently rewarded.

Tasks with probabilistic reward:

In a probabilistic version of this task used in parts of Figs. 2 and 6, rather than one of the options being consistently rewarded and the other not, both options were rewarded probabilistically. The sum of reward probabilities for both options for this version of the task was either 1 or 0.5, with OCT being rewarded with probability X and MCH with probability 1-X. Whenever probabilistic rewards were included in our tasks reward baiting was also incorporated as follows. If an unchosen action would have been rewarded, the reward was delivered on the next choice of that alternative with 100% certainty. This means that the likelihood an odor cue holds a reward increases over time if it is unchosen for many trials. Importantly the fly performing the task is never informed as to whether the unchosen option would have been rewarding on any given trial.

In Fig. 2, we analyzed a version of this task where flies learnt multiple simultaneous probabilistic cue-reward pairings. Here every fly experienced all three odors. For a given fly OCT and MCH were arbitrarily paired with reward 80 and 40% (or 40 and 80%) of the time respectively. EL was unpaired. Any given trial of these experiments consisted of a choice between OCT and EL or MCH and EL. These two options were interleaved together. Flies experienced 40 naive trials of each combination where no rewards were provided. This was followed by 40 rewarded trials.

The dynamic foraging task in Fig. 3 was adapted from monkey and mouse versions [5,10,16,21] and incorporated features such as baiting and probabilistic reward described earlier. OCT and MCH were the two odors used in this task. An example protocol of this task with details about the three block structure and number of trials per block can be seen in Fig. 3A. The relative reward ratios between the two odors for a given block were drawn from the following set [1:1, 1:2, 1:4, 1:8]. The sum of reward probabilities for both options for this version of the task was either 1 or 0.5.

Quantitative analysis and behavioral modeling:

All analysis and modeling was performed using MATLAB 2020b **(**Mathworks).

Analysis of fly movement and choices:

The *(x,y)* coordinates of the fly were analyzed to calculate: i) the distance of the fly from the center of the Y; ii) when the fly entered and exited a given odor; and iii) the time taken to make a choice. These metrics were then used to produce the plots in Fig. 1, 2, 3 and 7.

Distance from center was calculated by projecting the *(x,y)* location of the fly (pt) onto a skeleton of the Y. The skeleton consisted of three lines running down the middle of each arm to the center of the Y (v_0). Based on which arm the fly was located in (see Supp. Fig. 1B), its *(x,y)* position was projected onto the appropriate (i^{th}) skeleton line using the following equations for projecting a point onto a line, where v_i is the *(x,y)* coordinates of the end of the ith arm:

$$D_i = \frac{||\bar{b} \times \bar{a}||}{||\bar{a}||}; \text{ where: } \bar{b} = pt - v_0 \text{ and } \bar{a} = v_i - v_0 \quad (1)$$

The entries/exits of a fly into/from a particular odorant or air were estimated by tracking the region that the fly was located in at every time point and comparing it to the known odor-arm identity map (stored in the experiment .mat file). A turn (reversal) was considered to have been made whenever a fly entered an odor and then exited this odor without reaching the reward zone. An approach was considered to have been made whenever a fly entered and odor arm and then traveled all the way into the reward zone of that same arm without ever exiting it.

To calculate the time taken per trial, we made use of the timestamp vector that we saved along with the *(x,y)* vector. Time taken from the entire trial was calculated by subtracting the timestamp for the frame that the previous trial was completed by the timestamp of the frame

when the current trial was completed. Time taken from first exit of the air arm was calculated by subtracting the timestamp of the frame that the fly first exited the air arm after a trial began by the timestamp of the frame when the current trial was completed.

Choices themselves were identified by determining the arm that the fly completed the choice in and mapping that arm to its assigned odor on that trial.

Logistic regression to estimate influence of past rewards and choices on behavior:

To estimate the role of choice and reward histories in determining fly choices in the dynamic foraging task, we fit the following logistic regression for each fly:

$$\ln\left(\frac{P(C(t)=1)}{1-P(C(t)=1)}\right) = \sum_{i=1}^{T} \beta_i^R \, R(t-i) + \sum_{i=1}^{T} \beta_i^C \, C(t-i) + \beta_0, \qquad (2)$$

where: t is the present trial; i is variable used to iterate over the past T trials; $C(t)$ = 1 if the chosen odor was OCT and -1 if the chosen odor was MCH; $R(t)$ = 1 if chosen OCT option produced reward, -1 if chosen MCH option produced reward, and 0 otherwise; β_0 represents the weight assigned to the bias term; β_i^C represents the weight assigned to the ith past choice; β_i^R represents the weight assigned to the ith past reward; and we chose to look at the past T = 15 trials to align with previous studies [6,10]. The regression coefficients generated were 10-way cross-validated and the regression model included an elastic net regularization (MATLAB function - lassoglm). These fly-specific regression coefficients could be combined with the flies reward and choice histories to predict choice probability on a trial by trial basis and estimate the log-likelihood:

$$LL = \sum_{i=1}^{T_{max}} \sum_{j=1}^{2} p(x_{ji}) \cdot x_{ji} \qquad (3)$$

where: T_{max} is the total number of trials in the data being fit; i is a variable that iterates through the number of trials; j is a variable that iterates through the number of possible options ; $p(x_{ji})$ is the probability with which the model predicts that choice j occurs on trial i; x_{ji} is the choice that actually took place on trial i.

Predictive Q-learning:

We also developed a model to predict behavior in the dynamic foraging task using the Q-learning reinforcement learning algorithm [14]. This model determines choices on a given trial by comparing Q-values, which are values assigned to each option the agent has to choose between based on choice and reward history.

The Q-values were calculated for a given trial ‘t’ using the following equations.If OCT is chosen by the model, values are updated according to:

$$Q_{OCT}(t+1) = \zeta \cdot Q_{OCT}(t) + \alpha \cdot \delta_{OCT}(t), \quad (4)$$

$$Q_{MCH}(t+1) = \zeta \cdot Q_{MCH}(t), \quad (5)$$

where α is the learning rate, ζ is a decay constant, and $\delta_{OCT}(t) = R(t) - Q_{OCT}(t)$. Similarly, if MCH is chosen by the model, values are updated according to:

$$Q_{OCT}(t+1) = \zeta \cdot Q_{OCT}(t), \quad (6)$$

$$Q_{MCH}(t+1) = \zeta \cdot Q_{MCH}(t) + \alpha \cdot \delta_{MCH}(t), \quad (7)$$

where $\delta_{MCH}(t) = R(t) - Q_{MCH}(t)$.

These Q-values are then compared and passed through a sigmoidal non-linearity to determine the probability of each choice,

$$P(C(t) = OCT) = \frac{1}{1+e^{(-\beta \cdot (Q_{OCT}(t) - Q_{MCH}(t)))}}. \quad (8)$$

$$P(C(t) = MCH) = 1 - P(C(t) = OCT) \quad (9)$$

The probability value generated by this operation is compared with a value drawn from a uniform distribution over the [0,1] interval to determine whether the resulting choice is OCT or MCH. These predicted choices could be compared to fly behavior to compute the model’s log-likelihood. The parameters α, β and ζ are fit for each fly so as to maximize the log-likelihood.

Win - stay, Lose - switch model:

A third model to predict behavior incorporated information only about the most recent choice made by the fly unlike the logistic regression and Q-learning alternatives. In this “stay-switch” model the agent chooses randomly on the first trial. If the chosen option produces a reward the agent picks the option again on the next trial (stays). If it doesn’t produce a reward the agent picks the other option on the next trial (switches). This procedure repeats to generate a sequence of choices.

Neural circuit model of dynamic foraging:

We designed a neural circuit model, inspired by work from Loewenstein and Seung [59], that was used to simulate behavior in a dynamic foraging task. Two versions of this model were used.

Original Loewenstein and Seung Version:

The first version aimed to directly replicate the model used by Loewenstein and Seung (Fig.4A). It generated behavior on a trial-by-trial basis in the dynamic foraging task. The number of trials to simulate were input by the user prior to simulation (typically 60, 240 or 2000 trials). The model consisted of two sensory neurons (S_1 and S_2) whose activity was drawn at the beginning of each trial from a normal distribution with mean 1 and standard deviation 0.1. These neurons synapsed with weights (W_1 and W_2) onto two motor neurons (M_1 and M_2) ,whose activity was given by $\sum_{i=1}^{2} W_i \cdot S_i$. The activity of M_1 and M_2 were compared and the choice was made to correspond to whichever neuron had the larger activity.

Once the choice was made, rewards were provided as determined by the reward contingencies of the task. These were input by the user prior to running the simulation. The weights between S and M were updated after each choice and followed the following rule:

$$\Delta W_i(t) = \eta \cdot \overline{R}(t) \cdot \overline{S_i}(t), \quad (7)$$

$$W_i(t) = W_i(t-1) + \Delta W_i(t), \quad (8)$$

where $\overline{R} = R$ or $\overline{R} = R - E(R)$, and $\overline{S_i} = S_i$ or $\overline{S_i} = S_i - E(S_i)$, based on the learning rule. Note that $E(R)$ and $E(S_i)$ depended on time and were calculated in one of two ways: i) by calculating the mean over the last 10 trials, ii) by filtering the entire history with an exponential filter with exponential timescale (τ) of 3.5 trials. The various covariance and non-covariance rules were achieved by selecting the appropriate combinations of $\overline{R}$ and $\overline{S_i}$.

Task and Mushroom-body Inspired Version:

The second version incorporated modifications to the model that made it more appropriate for the task we designed for fruit flies (Fig.4C). This model consisted of two sensory inputs that represented activity of populations of kenyon cells (KCs). However, this version of the model looped through odor experiences, rather than looping through trials determined by two-alternative forced choices. Therefore the activity of the sensory neurons was drawn differently. Rather than both values being drawn from normal distribution with mean 1 and standard deviation $\sigma = 0.1$, this was only true for the odor that was deemed to have been "experienced" by the model on a given odor experience. The activity of the other neuron was drawn from a normal distribution with mean $\alpha = 0.1$ (Fig. 4 & 5; Supp. Fig. 2 & 3) and standard deviation $\sigma = 0.1$ Here, α represents the similarity, or overlap, between the two inputs. This was included because the KC representations of the two different odors used in our task are thought to have some amount of overlap [44]. However, we found that modulating this term did not affect the resulting matching behavior (Supp. Fig 4C) and so for Fig. 6, we chose $\alpha = 0$. We also explored incorporating noise covariance between the two sensory inputs (with correlation coefficient c = 0.1), but this correlation was empirically unimportant and we usually set c = 0.

Another difference is that an odor experience could lead to either an approach (choice) or a turn away. The behavior chosen by the model on any given odor experience depended on the response of the single output neuron incorporated into this model. The activity of this output neuron (M) was the weighted sum of the two inputs $\sum_{i=1}^{2} W_i \cdot S_i$. This was then passed through a sigmoidal non-linearity of the form:

$$P(Y = Approach) = \frac{1}{1+e^{-\beta \cdot (M-a)}}, \tag{9}$$

where $\beta = 4$ and $a = 1$ (this mean value was chosen to encourage exploration at the beginning of learning). A random number from the interval [0,1] was drawn and compared to Y to determine whether an approach/choice or turn was made. If a turn was made, no reward was provided, and the weights remained unchanged. The model then experienced a new odor and the process repeated. If a choice was made then a reward was provided based on the choice contingencies and weights were updated according to the rules in eqs. 7 and 8.

Logistic regression model for estimating learning rules:

To determine the learning rules that best predict fly behavior, we designed a logistic regression model that made use of the known relationship between MBON activity and behavior. This model predicted behavior using the known relationships between input and weights that give rise to MBON activity . This was done as follows:

$$\widehat{Y}(t) = \sum_{i=1}^{2} W_i(t) \cdot S_i(t) \tag{10}$$

$$W_i(t) = W_i(0) + \sum_{T \mid Y=1} \Delta W_i(T) \tag{11}$$

Now $\Delta W_i(T)$ depends on the learning rule that is used by the circuit and it was here that we wanted to have the regression model to identify the rule that provided the best fit to actual data. To do this we allowed the model to use a learning rule with 4 different terms whose weights could be modified,

$$\Delta W_i(T) = a + b \cdot \overline{S}_i(T) + c \cdot \overline{R}(T) + d \cdot \overline{R}(T) \cdot \overline{S}_i(T). \tag{12}$$

We used eq. 10, 11 and 12 to calculate the inputs to the regression model. The resulting inputs could be calculated as follows :

$$I_1(T) = \sum_{i=1}^{2} (t-1) \cdot \overline{S}_i(T) \tag{13}$$

$$I_2(T) = \sum_{i=1}^{2} \overline{S}_i(T) \cdot \sum_{t=1}^{T-1} \overline{S}_i(t) \tag{14}$$

$$I_3(T) = \sum_{i=1}^{2} \overline{S}_i(T) \cdot \sum_{t=1}^{T-1} \overline{R}(t) \tag{15}$$

$$I_4(T) = \sum_{i=1}^{2} \overline{S}_i(T) \cdot \sum_{t=1}^{T-1} \overline{S}_i(t) \cdot \overline{R}(t) \tag{16}$$

The weights assigned to each of the four inputs (a,b,c,d) and a bias could then be used to identify the learning rule that the model predicted as the best estimate for producing the behavior that was tested.

Acknowledgements

This work was supported by the Howard Hughes Medical Institute. We thank Igor Negrashov, Tobias Goulet, Peter Polidoro, Steven Sawtelle, Jon Arnold and others at jET, Janelia Research Campus for helping design and fabricate the Y-arena. We thank Todd Laverty and others in the Janelia Drosophila resources team for providing support with fly lines and media. We also thank all members of the Turner, Fitzgerald and Aso groups and other members of the Janelia community for support, discussions and feedback. AER would like to thank The Neuroscience Graduate Training Program at The Solomon H. Snyder Department of Neuroscience, Johns Hopkins University and thesis committee members Vivek Jayaraman, Ann Hermundstad, Jeremiah Cohen, Christopher Potter, Yoshi Aso and Erik Snapp for their guidance and support.

References

1. Stephens, D.W., and Krebs, J.R. (1986). Foraging Theory (Princeton University Press).

2. Kagel, J.H., Battalio, R.C., Battalio, L., Green, L., and Others (1995). Economic choice theory: An experimental analysis of animal behavior (Cambridge University Press).

3. Herrnstein, R.J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. J. Exp. Anal. Behav. *4*, 267–272.

4. Greggers, U., and Menzel, R. (1993). Memory dynamics and foraging strategies of honeybees. Behav. Ecol. Sociobiol. *32*, 17–29.

5. Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2004). Matching behavior and the representation of value in the parietal cortex. Science *304*, 1782–1787.

6. Lau, B., and Glimcher, P.W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. J. Exp. Anal. Behav. *84*, 555–579.

7. Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron *69*, 1204–1215.

8. Miller, K.J., Botvinick, M.M., and Brody, C.D. (2017). Dorsal hippocampus contributes to model-based planning. Nat. Neurosci. *20*, 1269–1276.

9. Costa, V.D., Mitz, A.R., and Averbeck, B.B. (2019). Subcortical Substrates of Explore-Exploit Decisions in Primates. Neuron *103*, 533–545.e5.

10. Bari, B.A., Grossman, C.D., Lubin, E.E., Rajagopalan, A.E., Cressy, J.I., and Cohen, J.Y. (2019). Stable Representations of Decision Variables for Flexible Behavior. Neuron.

11. Beron, C.C., Neufeld, S.Q., Linderman, S.W., and Sabatini, B.L. (2022). Mice exhibit stochastic and efficient action switching during probabilistic decision making. Proc. Natl. Acad. Sci. U. S. A. *119*, e2113961119.

12. Soltani, A., and Wang, X.-J. (2006). A biophysically based neural model of matching law behavior: melioration by stochastic synapses. J. Neurosci. *26*, 3731–3744.

13. Glimcher, P.W., and Fehr, E. (2013). Neuroeconomics: Decision Making and the Brain (Academic Press).

14. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction (MIT Press).

15. Padoa-Schioppa, C., Jandolo, L., and Visalberghi, E. (2006). Multi-stage mental process for economic choice in capuchins. Cognition *99*, B1–B13.

16. Lau, B., and Glimcher, P.W. (2008). Value representations in the primate striatum during matching behavior. Neuron *58*, 451–463.

17. Wang, A.Y., Miura, K., and Uchida, N. (2013). The dorsomedial striatum encodes net expected return, critical for energizing performance vigor. Nat. Neurosci. *16*, 639–647.

18. Aso, Y., Sitaraman, D., Ichinose, T., Kaun, K.R., Vogt, K., Belliart-guérin, G., Plaçais, P.-Y., Robie, A.A., Yamagata, N., Schnaitmann, C., et al. (2014). Mushroom body output neurons encode valence and guide memory-based action selection in Drosophila. Elife, 1–42.

19. Owald, D., Felsenberg, J., Talbot, C.B., Das, G., Perisse, E., Huetteroth, W., and Waddell, S. (2015). Activity of defined mushroom body output neurons underlies learned olfactory behavior in Drosophila. Neuron *86*, 417–427.

20. Hige, T., Aso, Y., Modi, M.N., Rubin, G.M., and Turner, G.C. (2015). Heterosynaptic Plasticity Underlies Aversive Olfactory Learning in Drosophila Article. Neuron *88*, 985–998.

21. Tsutsui, K.-I., Grabenhorst, F., Kobayashi, S., and Schultz, W. (2016). A dynamic code for economic object valuation in prefrontal cortex neurons. Nat. Commun. *7*, 12554.

22. Kuwabara, M., Kang, N., Holy, T.E., and Padoa-Schioppa, C. (2020). Neural mechanisms of economic choices in mice. Elife *9*.

23. Schultz, W., Dayan, P., and Montague, P.R. (1997). A Neural Substrate of Prediction and Reward. Science *275*, 1593–1599.

24. Corrado, G.S., Sugrue, L.P., Seung, H.S., and Newsome, W.T. (2005). Linear-Nonlinear-Poisson models of primate choice dynamics. J. Exp. Anal. Behav. *84*, 581–617.

25. Glimcher, P.W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc. Natl. Acad. Sci. U. S. A. *108 Suppl 3*, 15647–15654.

26. Keiflin, R., and Janak, P.H. (2015). Dopamine Prediction Errors in Reward Learning and Addiction: From Theory to Neural Circuitry. Neuron *88*, 247–263.

27. Lowet, A.S., Zheng, Q., Matias, S., Drugowitsch, J., and Uchida, N. (2020). Distributional Reinforcement Learning in the Brain. Trends Neurosci. *43*, 980–997.

28. Schultz, W. (2007). Behavioral dopamine signals. Trends Neurosci. *30*, 203–210.

29. Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85–88.

30. Parker, N.F., Cameron, C.M., Taliaferro, J.P., Lee, J., Yoon, J., Davidson, T.J., Daw, N.D., and Witten, I.B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nat. Neurosci. *19*, 845–854.

31. Jenett, A., Rubin, G.M., Ngo, T.-T.B., Shepherd, D., Murphy, C., Dionne, H., Pfeiffer, B.D., Cavallaro, A., Hall, D., Jeter, J., et al. (2012). A GAL4-Driver Line Resource for Drosophila Neurobiology. Cell Rep. *2*, 991–1001.

32. Riabinina, O., Luginbuhl, D., Marr, E., Liu, S., Wu, M.N., Luo, L., and Potter, C.J. (2015). Improved and expanded Q-system reagents for genetic manipulations. Nat. Methods *12*, 219–22, 5 p following 222.

33. Heisenberg, M. (2003). Mushroom Body Memoir: From Maps to Models. Nature *4*,

266–275.

34. Davis, R.L. (2005). Olfactory Memory Formation in Drosophila : From Molecular to Systems Neuroscience. 275–304.

35. Modi, M.N., Shuai, Y., and Turner, G.C. (2020). The Drosophila Mushroom Body: From Architecture to Algorithm in a Learning Circuit. Annu. Rev. Neurosci. *43*, 465–484.

36. Tempel, B.L., Bonini, N., Dawson, D.R., and Quinn, W.G. (1983). Reward learning in normal and mutant Drosophila. Proc. Natl. Acad. Sci. U. S. A. *80*, 1482–1486.

37. Tully, T., and Quinn, W.G. (1985). Classical conditioning and retention in normal and mutant Drosophila melanogaster. J. Comp. Physiol. A *157*, 263–277.

38. Tully, T., Preat, T., Boynton, S.C., and Del Vecchio, M. (1994). Genetic dissection of consolidated memory in Drosophila. Cell *79*, 35–47.

39. Shuai, Y., Hu, Y., Qin, H., Campbell, R.A.A., and Zhong, Y. (2011). Distinct molecular underpinnings of Drosophila olfactory trace conditioning. Proc. Natl. Acad. Sci. U. S. A. *108*, 20201–20206.

40. Liu, C., Plaçais, P.-Y., Yamagata, N., Pfeiffer, B.D., Aso, Y., Friedrich, A.B., Siwanowicz, I., Rubin, G.M., Preat, T., and Tanimoto, H. (2012). A subset of dopamine neurons signals reward for odour memory in Drosophila. Nature *488*, 512–516.

41. Vogt, K., Schnaitmann, C., Dylla, K.V., Knapek, S., Aso, Y., Rubin, G.M., and Tanimoto, H. (2014). Shared mushroom body circuits underlie visual and olfactory memories in Drosophila. Elife *3*, e02395.

42. Jacob, P.F., and Waddell, S. (2020). Spaced Training Forms Complementary Long-Term Memories of Opposite Valence in Drosophila. Neuron.

43. Aso, Y., Siwanowicz, I., Bräcker, L., Ito, K., Kitamoto, T., and Tanimoto, H. (2010). Specific dopaminergic neurons for the formation of labile aversive memory. Curr. Biol. *20*, 1445–1451.

44. Campbell, R.A.A., Honegger, K.S., Qin, H., Li, W., Demir, E., and Turner, G.C. (2013). Imaging a population code for odor identity in the Drosophila mushroom body. J. Neurosci. *33*, 10568–10581.

45. Ichinose, T., Aso, Y., Yamagata, N., Abe, A., Rubin, G.M., and Tanimoto, H. (2015). Reward signal in a recurrent circuit drives appetitive long-term memory formation. Elife *4*, e10719.

46. Aso, Y., and Rubin, G.M. (2016). Dopaminergic neurons write and update memories with cell-type-specific rules. Elife *5*, 1–15.

47. Felsenberg, J., Barnstedt, O., Cognigni, P., Lin, S., Waddell, S., and Building, T. (2017). Re-evaluation of learned information in Drosophila. Nature Publishing Group.

48. Felsenberg, J., Jacob, P.F., Walker, T., Barnstedt, O., Edmondson-Stait, A.J., Pleijzier, M.W., Otto, N., Schlegel, P., Sharifi, N., Perisse, E., et al. (2018). Integration of Parallel Opposing Memories Underlies Memory Extinction. Cell.

49. Sayin, S., De Backer, J.-F., Siju, K.P., Wosniack, M.E., Lewis, L.P., Frisch, L.-M., Gansen, B., Schlegel, P., Edmondson-Stait, A., Sharifi, N., et al. (2019). A Neural Circuit Arbitrates between Persistence and Withdrawal in Hungry Drosophila. Neuron.

50. Sabandal, J.M., Berry, J.A., and Davis, R.L. (2021). Dopamine-based mechanism for transient forgetting. Nature *591*, 426–430.

51. Claridge-chang, A., Roorda, R.D., Vrontou, E., Sjulson, L., Li, H., and Hirsh, J. (2009). Writing Memories with Light-Addressable Reinforcement Circuitry. 405–415.

52. Buchanan, S.M., Kain, J.S., and de Bivort, B.L. (2015). Neuronal control of locomotor handedness in Drosophila. Proc. Natl. Acad. Sci. U. S. A. *112*, 6700–6705.

53. Seidenbecher, S.E., Sanders, J.I., von Philipsborn, A.C., and Kvitsiani, D. (2020). Reward foraging task and model-based analysis reveal how fruit flies learn value of available options. PLoS One *15*, e0239616.

54. Wiggin, T.D., Hsiao, Y., Liu, J.B., Huber, R., and Griffith, L.C. (2021). Rest Is Required to Learn an Appetitively-Reinforced Operant Task in Drosophila. Front. Behav. Neurosci. *15*, 681593.

55. Smith, M.A.-Y., Honegger, K.S., Turner, G., and de Bivort, B. (2022). Idiosyncratic learning performance in flies. Biol. Lett. *18*, 20210424.

56. Herrnstein, R.J. (1997). The Matching Law: Papers in Psychology and Economics (Harvard University Press).

57. Vulkan, N. (2000). An Economist's Perspective on Probability Matching. J. Econ. Surv. *14*, 101–118.

58. Iigaya, K., Ahmadian, Y., Sugrue, L.P., Corrado, G.S., Loewenstein, Y., Newsome, W.T., and Fusi, S. (2019). Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales. Nat. Commun. *10*, 1466.

59. Loewenstein, Y., and Seung, H.S. (2006). Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. Proc. Natl. Acad. Sci. U. S. A. *103*, 15224–15229.

60. Adel, M., and Griffith, L.C. (2021). The Role of Dopamine in Associative Learning in Drosophila: An Updated Unified Model. Neurosci. Bull. *37*, 831–852.

61. Jiao, Y., Moon, S.J., Wang, X., Ren, Q., and Montell, C. (2008). Gr64f is required in combination with other gustatory receptors for sugar detection in Drosophila. Curr. Biol. *18*, 1797–1801.

62. Fujii, S., Yavuz, A., Slone, J., Jagge, C., Song, X., and Amrein, H. (2015). Drosophila sugar receptors in sweet taste perception, olfaction, and internal nutrient sensing. Curr. Biol. *25*, 621–627.

63. Haberkern, H., Basnak, M.A., Ahanonu, B., Schauder, D., Cohen, J.D., Bolstad, M., Bruns, C., and Jayaraman, V. (2019). Visually Guided Behavior and Optogenetically Induced Learning in Head-Fixed Flies Exploring a Virtual Landscape. Curr. Biol. *29*, 1647–1659.e8.

64. Loewenstein, Y. (2008). Robustness of learning that is based on covariance-driven synaptic plasticity. PLoS Comput. Biol. *4*, e1000007.

65. Turner, G.C., Bazhenov, M., and Laurent, G. (2008). Olfactory Representations by Drosophila Mushroom Body Neurons. J. Neurophysiol. *99*, 734–746.

66. Honegger, K.S., Campbell, R.A.A., and Turner, G.C. (2011). Cellular-resolution population imaging reveals robust sparse coding in the Drosophila mushroom body. J. Neurosci. *31*, 11772–11785.

67. Caron, S.J.C., Ruta, V., Abbott, L.F., and Axel, R. (2013). Random convergence of olfactory inputs in the Drosophila mushroom body. Nature *497*, 113–117.

68. Aso, Y., Hattori, D., Yu, Y., Johnston, R.M., Nirmala, A., Ngo, T.-T.B., Dionne, H., Abbott, L.F., Axel, R., Tanimoto, H., et al. (2014). The neuronal architecture of the mushroom body provides a logic for associative learning. Elife *3:e04577*, 1–47.

69. Séjourné, J., Plaçais, P.-Y., Aso, Y., Siwanowicz, I., Trannoy, S., Thoma, V., Tedjakumala, S.R., Rubin, G.M., Tchénio, P., Ito, K., et al. (2011). Mushroom body efferent neurons responsible for aversive olfactory memory retrieval in Drosophila. Nat. Neurosci. *14*, 903–910.

70. Cohn, R., Morantte, I., and Ruta, V. (2015). Coordinated and Compartmentalized Neuromodulation Shapes Sensory Processing in Drosophila. Cell *163*, 1742–1755.

71. Perisse, E., Owald, D., Barnstedt, O., Talbot, C.B., Huetteroth, W., and Waddell, S. (2016). Aversive Learning and Appetitive Motivation Toggle Feed-Forward Inhibition in the Drosophila Mushroom Body. Neuron *90*, 1086–1099.

72. Berry, J.A., Phan, A., and Davis, R.L. (2018). Dopamine Neurons Mediate Learning and Forgetting through Bidirectional Modulation of a Memory Trace. Cell Rep. *25*, 651–662.e5.

73. Handler, A., Graham, T.G.W., Cohn, R., Morantte, I., Siliciano, A.F., Zeng, J., Li, Y., and Ruta, V. (2019). Distinct Dopamine Receptor Pathways Underlie the Temporal Sensitivity of Associative Learning. Cell *178*, 60–75.e19.

74. Jiang, L., and Litwin-Kumar, A. (2019). Models of heterogeneous dopamine signaling in an insect learning and memory center. bioRxiv, 737064.

75. Otto, N., Pleijzier, M.W., Morgan, I.C., Edmondson-Stait, A.J., Heinz, K.J., Stark, I., Dempsey, G., Ito, M., Kapoor, I., Hsu, J., et al. (2020). Input Connectivity Reveals Additional Heterogeneity of Dopaminergic Reinforcement in Drosophila. Curr. Biol. *30*, 3200–3211.e8.

76. Li, F., Lindsey, J.W., Marin, E.C., Otto, N., Dreher, M., Dempsey, G., Stark, I., Bates, A.S., Pleijzier, M.W., Schlegel, P., et al. (2020). The connectome of the adult Drosophila mushroom body provides insights into function. Elife *9*.

77. Eschbach, C., Fushiki, A., Winding, M., Schneider-Mizell, C.M., Shao, M., Arruda, R., Eichler, K., Valdes-Aleman, J., Ohyama, T., Thum, A.S., et al. (2020). Recurrent architecture for adaptive regulation of learning in the insect brain. Nat. Neurosci. *23*, 544–555.

78. Riemensperger, T., Völler, T., Stock, P., Buchner, E., and Fiala, A. (2005). Punishment

prediction by dopaminergic neurons in Drosophila. Curr. Biol. *15*, 1953–1960.

79. Dylla, K.V., Raiser, G., Galizia, C.G., and Szyszka, P. (2017). Trace Conditioning in Drosophila Induces Associative Plasticity in Mushroom Body Kenyon Cells and Dopaminergic Neurons. Front. Neural Circuits *11*, 42.

80. König, C., Khalili, A., Niewalda, T., Gao, S., and Gerber, B. (2019). An optogenetic analogue of second-order reinforcement in Drosophila. Biol. Lett. *15*, 20190084.

81. Gkanias, E., McCurdy, L.Y., Nitabach, M.N., and Webb, B. (2022). The incentive circuit: memory dynamics in the mushroom body of Drosophila melanogaster. bioRxiv, 2021.06.11.448104.

82. Zheng, Z., Lauritzen, J.S., Perlman, E., Robinson, C.G., Nichols, M., Milkie, D., Torrens, O., Price, J., Fisher, C.B., Sharifi, N., et al. (2018). A Complete Electron Microscopy Volume of the Brain of Adult Drosophila melanogaster. Cell *174*, 730–743.e22.

83. Scheffer, L.K., Xu, C.S., Januszewski, M., Lu, Z., Takemura, S.-Y., Hayworth, K.J., Huang, G.B., Shinomiya, K., Maitlin-Shepard, J., Berg, S., et al. (2020). A connectome and analysis of the adult Drosophila central brain. Elife *9*.

Figure 1

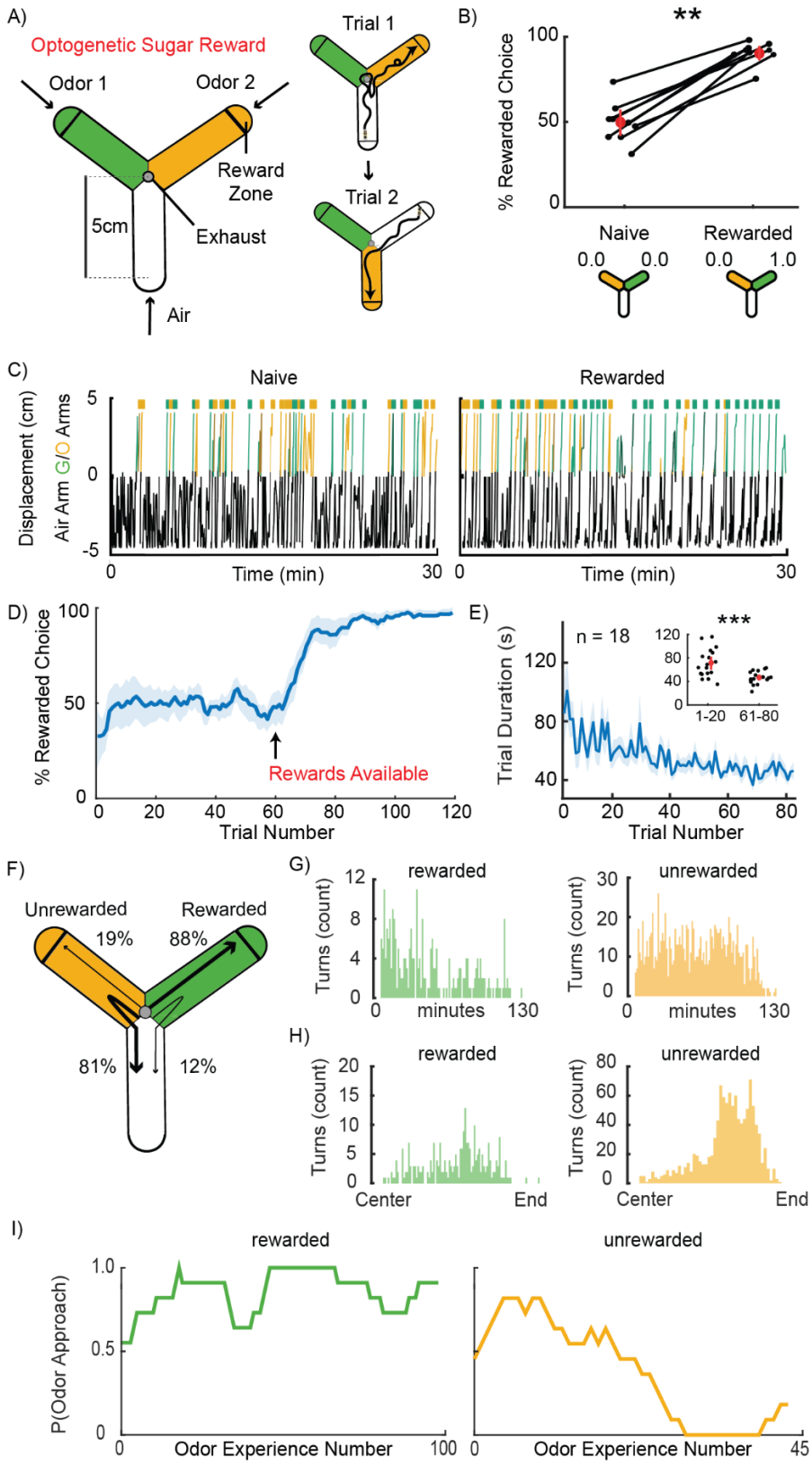

A)
Optogenetic Sugar Reward
Odor 1
Odor 2
Reward Zone
5cm
Exhaust
Air
Trial 1
Trial 2
B)
**
% Rewarded Choice
100
50
0
Naive
Rewarded
0.0
0.0
0.0
1.0
C)
Naive
Rewarded
Displacement (cm)
Air Arm G/O Arms
5
0
-5
0
Time (min)
30
D)
% Rewarded Choice
Rewards Available
Trial Number
E)
n = 18
***
Trial Duration (s)
1-20
61-80
Trial Number
F)
Unrewarded
Rewarded
19%
88%
81%
12%
G)
rewarded
unrewarded
Turns (count)
minutes
130
H)
rewarded
unrewarded
Turns (count)
Center
End
I)
rewarded
unrewarded
P(Odor Approach)
Odor Experience Number
100
45

Figure 1: Individual flies show signatures of learning in a novel 2-AFC task designed to probe foraging strategies.

A Schematic of Y-arena for single fly learning experiments *(left)*. Airflow travels from tips of each arm to an outlet in the center. Reward zones are demarcated by lines on the schematic, but are not visible to the fly. A choice is considered to have been made when a fly crosses from an air arm into the reward zone of an odorized arm, triggering delivery of an optogenetic reward by activating sugar sensing neurons with a 500 ms pulse of red light. The next trial then commences as the choice arm switches to air and the two odors (represented as green and orange) are randomly reassigned to the other two arms *(right)*.

B Percentage rewarded choices calculated over two blocks of 60 trials in the Y arena, the first with no reward available (naive), the second with one odor rewarded every time it is chosen - 100:0 (rewarded). Flies prefer the rewarded option in the rewarded block compared to naive (Wilcoxon signed-rank test: $p = 0.0039$, $n = 9$).

C An example trajectory of a fly in the Y before *(left)* and after *(right)* MCH (green) is paired with reward. The fly's distance down the air arm is represented as negative values and plotted in black, while distances down an odorized arm are represented as positive values and shown in green or orange to indicate the different odors. An odor choice is represented by a colored raster tick, at which point the arena resets and that arm switches to air, so the fly's position jumps to the tip of the air arm. The example fly initially shows no preference during the naive block and this preference switches to the green odor in the rewarded block. The trajectories also become less circuitous and choices more regular.

D The instantaneous percentage of choices made towards the rewarded option (percent rewarded choice) is calculated in a 10 odor-experience window and plotted against number of trials ($n = 9$ flies, mean+- SEM). As in B, flies learn to prefer the rewarded odor. The first 60 trials are part of the naive block. From the 61st trial onwards rewards are available.

E The time taken to make a choice decreases as flies become more familiar with the task (mean +- SEM). *Inset:* Average trial time for the first 20 trials is longer than the average trial for the last 20 trials (Wilcoxon signed-rank test: $p = 6.292*10^{-4}$, $n=18$ flies).

F Percentage of odor arm entries that lead to an odor approach (entry to reward zone) or a reversal (exit to another arm) in a 100:0 task. Flies show increased turning away from the unrewarded odor (Fig. 1F; Mann-Whitney rank-sum test: $p = 3.2278 * 10^{-7}$, $n = 18$), and decreased turning away from the rewarded odor (Mann-Whitney rank-sum test: $p= 3.2166 * 10^{-7}$, $n = 18$).

G Histogram showing the number of reversals over time for rewarded and unrewarded odor choices. Reversals decrease over time for the rewarded odor (Wilcoxon signed rank test: $p =0.0171$, $n = 18$, comparing bins 15-65 with 66-115 minutes- to exclude large time values when many flies had already finished the task), but not the unrewarded one (Wilcoxon signed rank test: $p = 0.1839$, $n = 18$, same as rewarded odor).

H Histogram showing the number of reversals as a function of distance along the odorized arm. The majority of turns occur partway up the arm for both rewarded and unrewarded odors.

I The instantaneous probability of approaching an odor (same as 1 - probability of reversal) is calculated in 6 odor-experience window and plotted against the number of odor experiences for an example fly. A single odor experience is defined as an entry into one of the odorized arms.

Figure 2

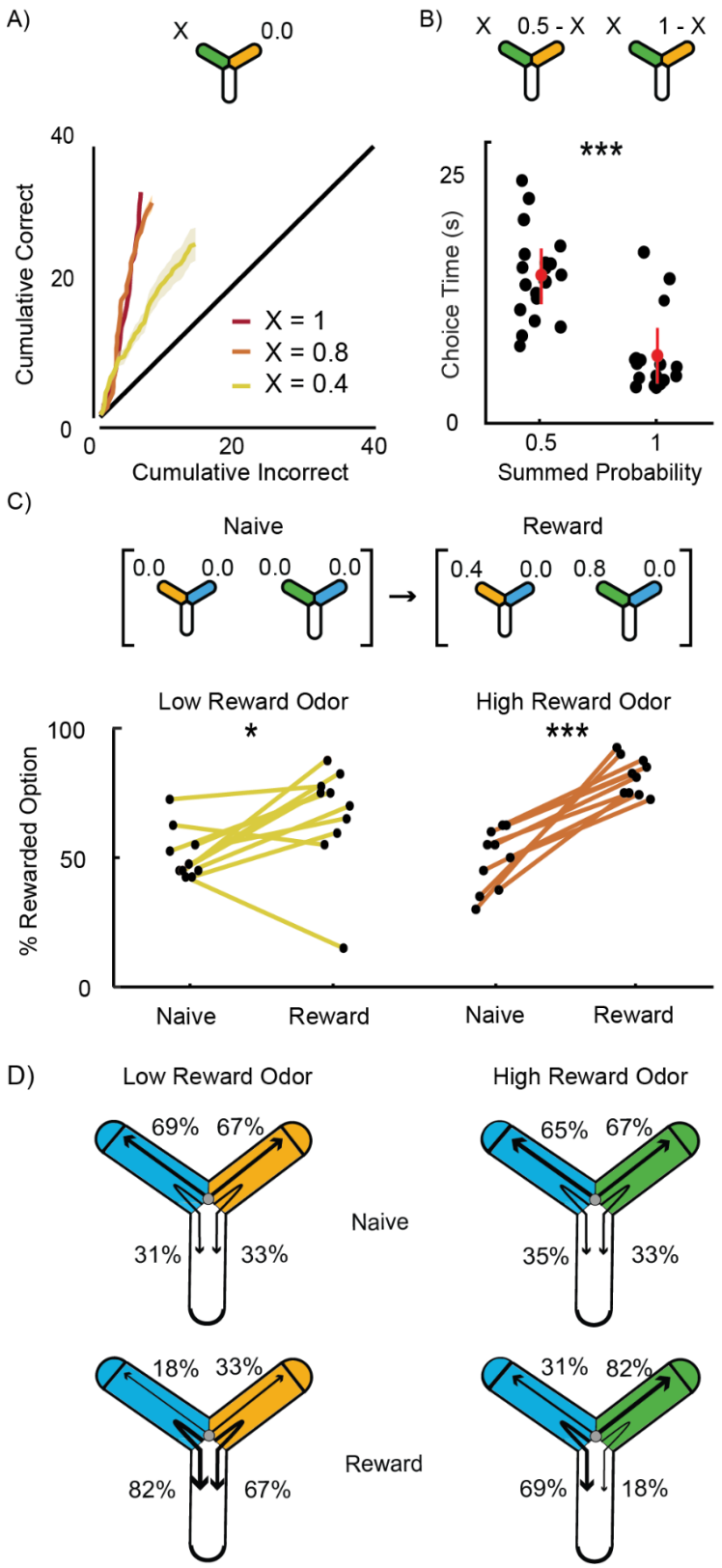

A)
X
0.0
Cumulative Correct
Cumulative Incorrect
X = 1
X = 0.8
X = 0.4
B)
X
0.5 - X
X
1 - X
***
Choice Time (s)
Summed Probability
C)
Naive
Reward
0.0
0.0
0.0
0.0
0.4
0.0
0.8
0.0
Low Reward Odor
High Reward Odor
*
***
% Rewarded Option
Naive
Reward
Naive
Reward
D)
Low Reward Odor
High Reward Odor
69%
67%
31%
33%
Naive
65%
67%
35%
33%
18%
33%
82%
67%
Reward
31%
82%
69%
18%

Figure 2: Flies can learn multiple probabilistic stimulus-reward associations.

A Flies learn to associate odors with probabilistic rewards. Cumulative correct and error choices, across 40 trials, for three different reward probabilities : 1, 0.8, 0.4 *(top)*. Flies show a preference for the rewarded odor in all cases, indicated by the slope of all three curves being greater than 1 (Mann-Whitney rank-sum test: probability = 1, $p = 4.4500*10^{-8}$, n = 18; probability= 0.8, $p = 5.8927*10^{-5}$, n = 10; probability = 0.4, p = 0.0014, n = 10).

B The average choice time, measured from first exit of the air arm till entry into the reward zone, for two different total reward probabilities. Average across 80 trials, in two conditions where the summed probability of receiving reward from both options is 0.5 or 1 *(top)*. Average choice time decreases as the summed probability increases (Mann-Whitney rank-sum test: $p = 4.7653*10^{-4}$, n = 18 for summed probability = 1; n = 20 for summed probability = 0.5).

C *Top:* Schematic of the protocol for training flies with two simultaneous probabilistic cue-reward contingencies. Two different odor choices are interleaved throughout a naive, unrewarded block of 80 total trials, and a reward block, where options were rewarded with probability of 0.4 or 0.8, as indicated. The blue arm indicates that the unrewarded odor is always the same. *Bottom:* Performance (percentage of choices in which the option that may produce reward was chosen) on the low and high reward choices, showing that individual flies learn both associations (Wilcoxon signed-rank test: $p = 1.7962*10^{-4}$ for high rewarding odor; n = 10 p = 0.0124 for low rewarding odor, n = 10).

D The average percentages of turning away from and walking into each odor - high rewarded (green), low rewarded (orange) and unrewarded (blue) are graphically represented in a schematic of the Y arena (n=10 flies).

Figure 3

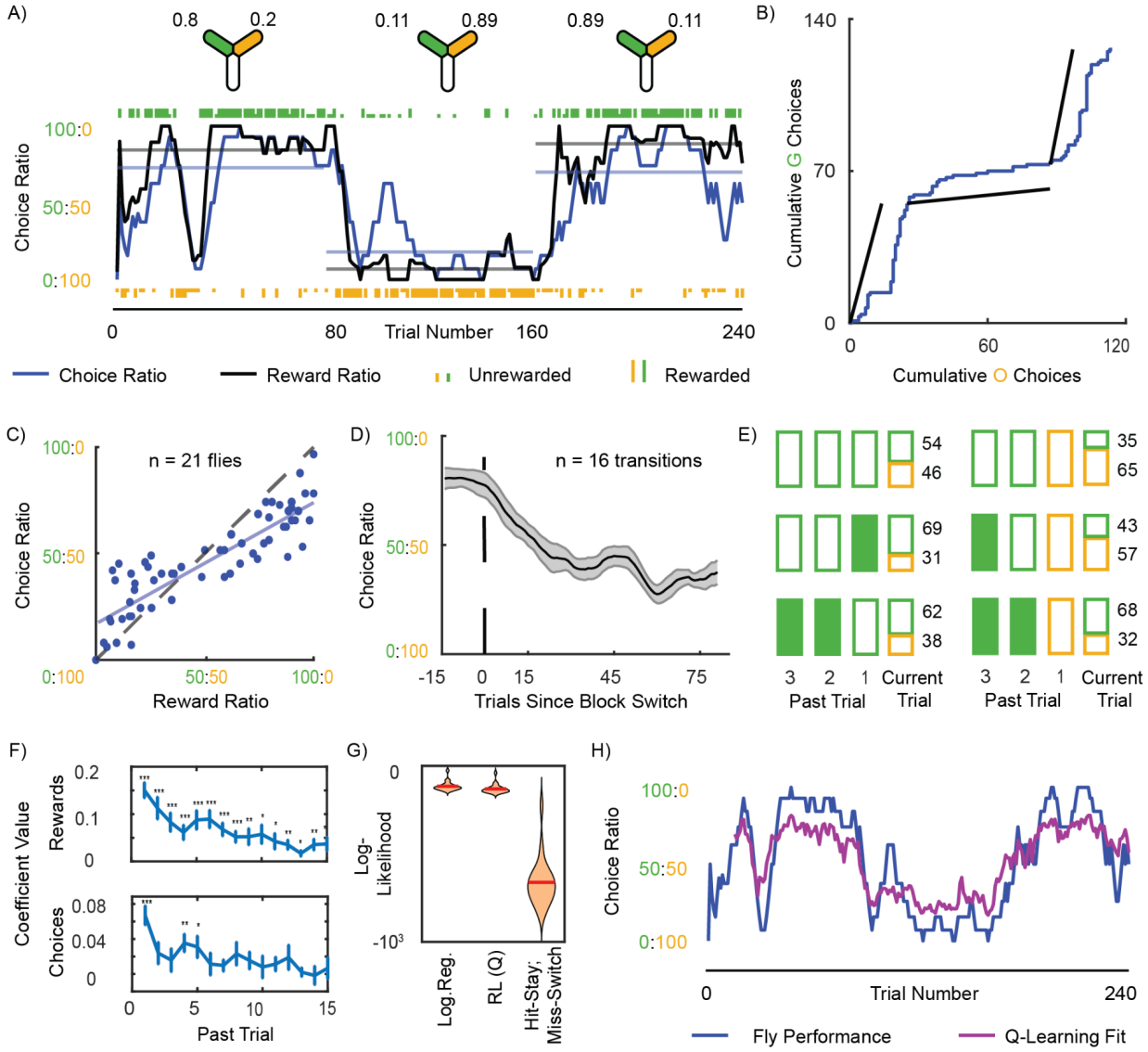

A)
0.8
0.2
0.11
0.89
0.89
0.11
Choice Ratio
100:0
50:50
0:100
0
80
Trial Number
160
240
Choice Ratio
Reward Ratio
Unrewarded
Rewarded
B)
140
70
0
Cumulative G Choices
0
60
120
Cumulative O Choices
C)
n = 21 flies
Choice Ratio
100:0
50:50
0:100
Reward Ratio
D)
n = 16 transitions
Choice Ratio
100:0
50:50
0:100
-15
0
15
45
75
Trials Since Block Switch
E)
54
46
69
31
62
38
35
65
43
57
68
32
3
2
1
Current Trial
Past Trial
F)
Coefficient Value
Rewards
0.2
0.1
0
Choices
0.08
0.04
0
0
5
10
15
Past Trial
G)
Log-Likelihood
0
-10³
Log.Reg.
RL (Q)
Hit-Stay; Miss-Switch
H)
Choice Ratio
100:0
50:50
0:100
0
Trial Number
240
Fly Performance
Q-Learning Fit

Figure 3: Flies perform operant matching.

A Matching of instantaneous choice ratio (blue) and reward ratio (black) in an example fly. Reward probabilities for trial blocks are indicated by the Y arena icons above *(top)*. Individual odor choices are denoted by rasters, tall rasters indicating a trial where reward was received, short indicating no reward. Curves show 10-trial averaged choice ratio and reward ratio, and horizontal lines the corresponding averages over the 80-trial blocks.

B Cumulative choices of the same example fly from A. The black line indicates the block-averaged reward ratio as it changes over the three successive blocks; the blue line indicates the choice ratio across those blocks. The parallel slopes of the two lines indicate matching.

C Block-averaged choice ratio is approximately equal to reward ratio, following the matching law, but with a small amount of undermatching (n=60 blocks from n=21 flies).

D Change in instantaneous choice ratio around block changes. Choice ratio adjusts to plateau levels in ~15 trials.

E An analysis of choices following particular histories of experience. Here, choices made by flies over three consecutive past trials are represented by boxes of different colors. Colors represent different odors chosen, and rewarded choices are represented by filled boxes. Probabilities with which flies choose the green and orange odor on the current trial conditional on this history are illustrated with associated values. Only 6 out of 64 possible sequences are illustrated here. These results hint that present choice probabilities are affected by rewards from at least three trials in the past.

F Coefficients from logistic regression performed on fly choice behavior to determine the influence of 15 past rewards *(top)* and choices *(bottom)* on a fly's present choice (Student's t-test: *** - $p < 0.001$, ** - $p < 0.01$, * - $p<0.05$, n = estimates from 21 flies).

G Model fit (log-likelihood) quality for logistic regression, Q-learning and stay-switch models, showing that models with a longer choice and reward history do better at predicting behavior.

H Q-learning model fit on behavior from the example fly from panel A. The fit is plotted from the 15th trial onwards to avoid edge effects.

Figure 4

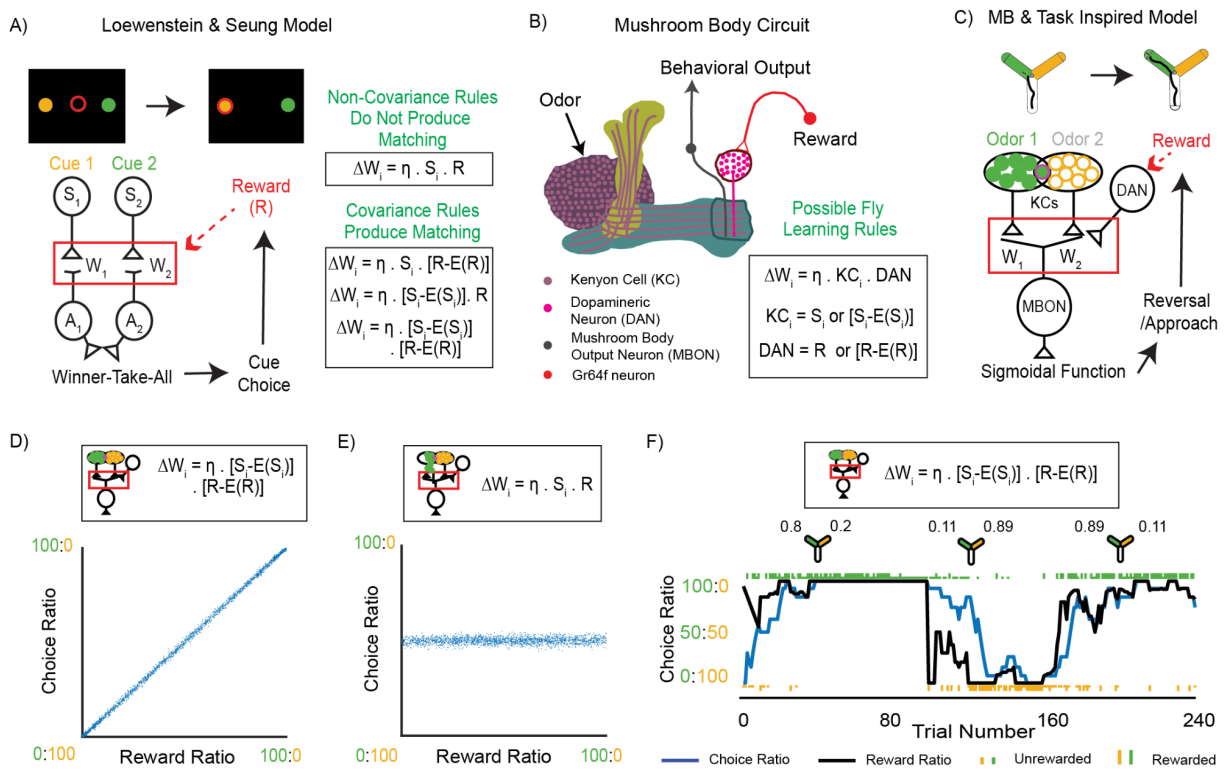

A)
Loewenstein & Seung Model
Cue 1
Cue 2
Reward (R)
Winner-Take-All
Cue Choice
Non-Covariance Rules Do Not Produce Matching
ΔW_i = η . S_i . R
Covariance Rules Produce Matching
ΔW_i = η . S_i . [R-E(R)]
ΔW_i = η . [S_i-E(S_i)]. R
ΔW_i = η . [S_i-E(S_i)] . [R-E(R)]
B)
Mushroom Body Circuit
Behavioral Output
Odor
Reward
Possible Fly Learning Rules
Kenyon Cell (KC)
Dopamineric Neuron (DAN)
Mushroom Body Output Neuron (MBON)
Gr64f neuron
ΔW_i = η . KC_i . DAN
KC_i = S_i or [S_i-E(S_i)]
DAN = R or [R-E(R)]
C)
MB & Task Inspired Model
Odor 1
Odor 2
Reward
KCs
DAN
MBON
Reversal /Approach
Sigmoidal Function
D)
ΔW_i = η . [S_i-E(S_i)] . [R-E(R)]
Choice Ratio
Reward Ratio
100:0
0:100
E)
ΔW_i = η . S_i . R
Choice Ratio
Reward Ratio
100:0
0:100
F)
ΔW_i = η . [S_i-E(S_i)] . [R-E(R)]
0.8
0.2
0.11
0.89
0.89
0.11
Choice Ratio
100:0
50:50
0:100
0
80
160
240
Trial Number
Choice Ratio
Reward Ratio
Unrewarded
Rewarded

Figure 4: Covariance-based learning rules are necessary for operant matching in a mushroom body-like model.

A A description of the circuit and task developed by Loewenstein and Seung to study the requirements of matching behavior. The neural network consists of two sensory neurons S_1 and S_2 that synapse into two motor neurons A_1 and A_2 via synapses with weights W_1 and W_2. Choices are determined via a winner-take-all computation downstream of motor neurons. Once choices are made, weights are updated according to one of the plasticity rules shown in the boxes (right). Here, S_i is the activity of the ith sensory neuron; R is a measure of the presence or absence of reward; $E(S_i)$ is the mean or expectation of the activity of S_i and E(R) is the expectation of reward.

B A schematic representing the MB. Kenyon Cells (KCs) - pink circles with axons projecting into the lobes of the MB (yellow and blue). Protocerebral anterior medial (PAM) cluster dopaminergic neurons (DANs) - red circles with axons of one example DAN shown projecting to a particular MB compartment. MBON - gray with axon projecting to the same MB compartment. Gr64f receptor expressing sugar sensory neuron - green - providing reward information to the DANs. Inset : a schematic depicting the reward dependent synaptic plasticity rule at the KC-MBON synapse .

C A schematic adapting the circuit and task developed by L&S to more closely resemble the MB and the features of our olfactory task. Sensory inputs are now thought of as populations of KCs with overlap between representations - depicted as multiple circles colored to match the odor they respond to. Weights between all input and MBON are modified in the same way - represented by a single synapse between the multiple inputs and the MBON. In this version of the task, agents only experience one odor at a time and MBON output determines whether the agent turns away and experiences another input or walking forward and enters the reward zone. Reward information is provided to this circuit via DAN activity which either represents simply reward (R) or reward minus reward expectation (R-E[R]).

D Block-averaged choice ratio produced by the model described in (C) plotted against reward ratio. Simulation uses a covariance-based learning rule $[S_i-E(S_i)]*[R-E(R)]$. The model exhibits a matching behavior (slope is 1).

E Same as (C), but simulated with a non-covariance learning rule where updates are based simply on R. The model produces behavior that does not show matching (slope < 1).

F An example simulation showing the performance in a 3 block task of a model incorporating a covariance-based rule $[S_i-E(S_i)]*[R-E(R)]$. Task reward contingencies are the same as shown for the example fly in Fig. 3A.

Figure 5

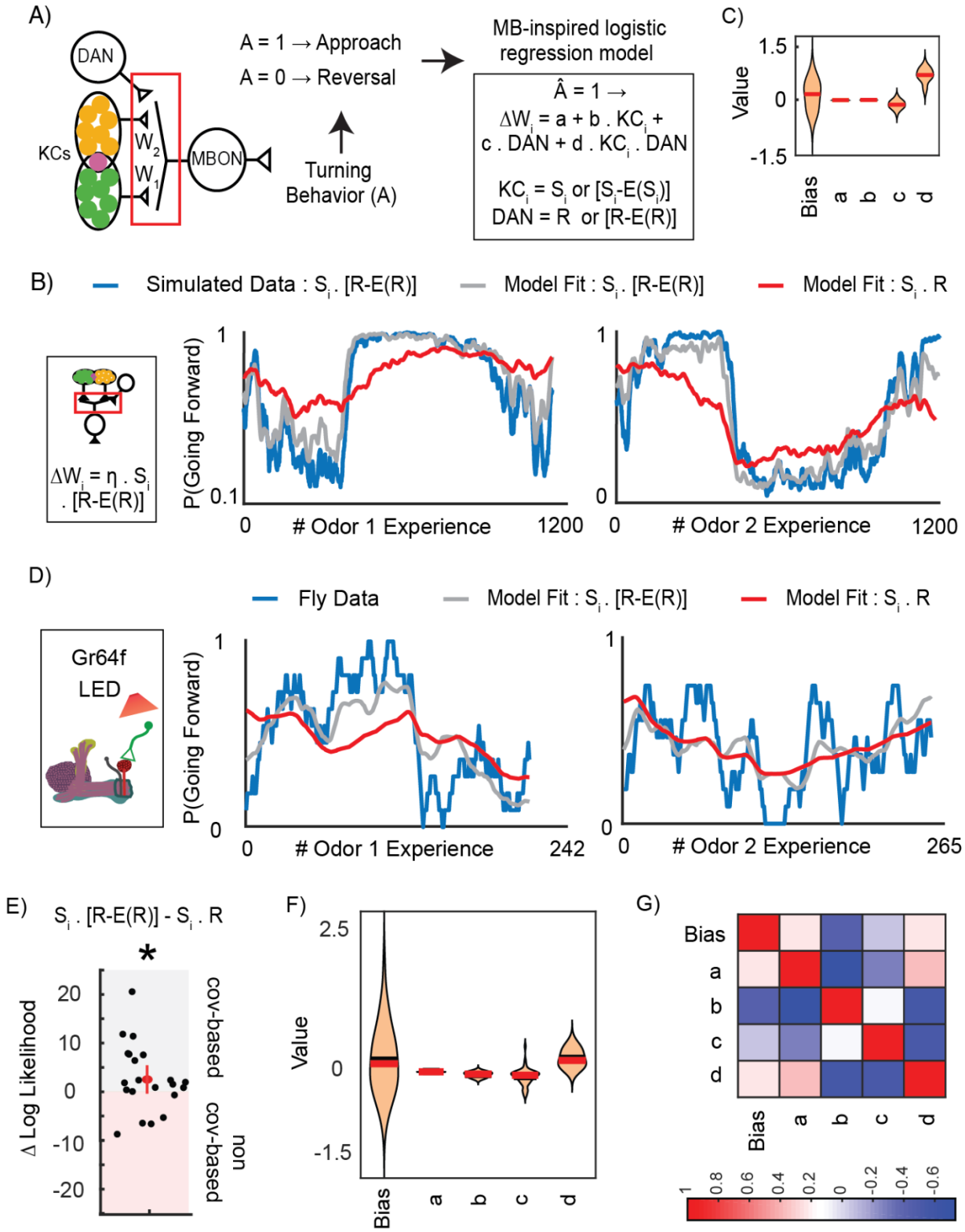

A)
DAN
KCs
W2
W1
MBON
A = 1 → Approach
A = 0 → Reversal
Turning Behavior (A)
MB-inspired logistic regression model
Â = 1 →
ΔWi = a + b . KCi + c . DAN + d . KCi . DAN
KCi = Si or [Si-E(Si)]
DAN = R or [R-E(R)]
C)
Value
1.5
0
-1.5
Bias
a
b
c
d
B)
Simulated Data : Si . [R-E(R)]
Model Fit : Si . [R-E(R)]
Model Fit : Si . R
ΔWi = η . Si . [R-E(R)]
P(Going Forward)
1
0.1
0
Odor 1 Experience
1200
Odor 2 Experience
D)
Fly Data
Model Fit : Si . [R-E(R)]
Model Fit : Si . R
Gr64f LED
242
265
E)
Si . [R-E(R)] - Si . R
*
Δ Log Likelihood
20
10
0
-10
-20
cov-based
non cov-based
F)
2.5
-1.5
G)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6

Figure 5: Regression analysis explains individual choice behavior better when incorporating covariance-based learning rules.

A A schematic showing the details and logic of the MB-inspired regression model.

B Example approach/reversal behavior (blue) simulated using an S_i*[R-E(R)] rule, and fit using an MB-inspired regression model (A) that incorporates either the same rule (grey), or an S*R rule (red). The predictions resulting from the model using the covariance-based rule is a better fit for the simulated data.

C Regression coefficients assigned to each term of the learning rule when the MB-inspired regression model using an S_i*[R-E(R)] rule was fit to the simulated behavior. Note the non-zero weight of the c and d terms.

D Example fly data (blue) fit using an MB-inspired regression model (A) that incorporates either an S_i*[R-E(R)] rule (gray), or an S*R rule (red).

E Change in log-likelihood (LL) metric, computed by subtracting the LL of the non-covariance-based model from a covariance-based rule, plotted for each fly (n = 21) . On average, fly behavior was better predicted by the covariance-based model (Wilcoxon signed-rank test: p = 0.0386). Individual flies that were better fit by the covariance-based model have a positive value on this plot (gray region) while flies better fit by the non covariance-based model have a negative value (red region).

F Regression coefficients assigned to each term of the learning rule when the MB-inspired regression model using an S_i*[R-E(R)] rule was fit to the flies behavior. Overall profile of weights is similar to (C).

G Correlation between regression coefficients resulting from MB-inspired regression model using an S_i*[R-E(R)] rule.

Figure 6

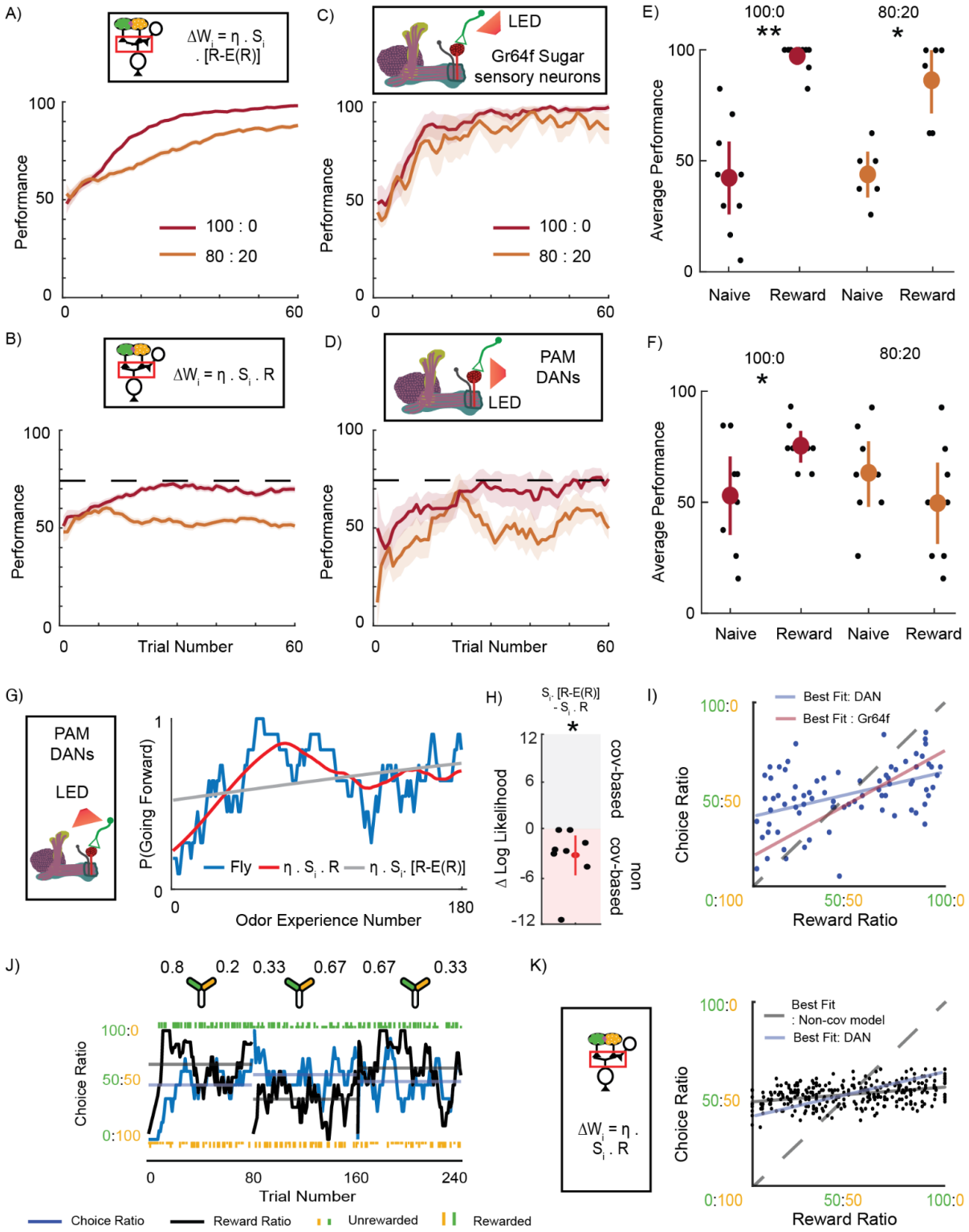

A)
ΔW_i = η . S_i . [R-E(R)]
Performance
100 : 0
80 : 20
B)
ΔW_i = η . S_i . R
Trial Number
C)
LED
Gr64f Sugar sensory neurons
D)
PAM DANs
LED
E)
100:0
80:20
Average Performance
Naive
Reward
F)
G)
PAM DANs
LED
P(Going Forward)
Fly
η . S_i . R
η . S_i. [R-E(R)]
Odor Experience Number
H)
S_i. [R-E(R)] - S_i . R
Δ Log Likelihood
cov-based
non cov-based
I)
Best Fit: DAN
Best Fit : Gr64f
Choice Ratio
Reward Ratio
J)
0.8 0.2 0.33 0.67 0.67 0.33
Choice Ratio
Trial Number
Choice Ratio
Reward Ratio
Unrewarded
Rewarded
K)
ΔW_i = η . S_i . R
Best Fit : Non-cov model
Best Fit: DAN
Choice Ratio
Reward Ratio

Figure 6: Bypassing reward expectation with direct DAN activation indicates that DANs incorporate reward expectation.

A Simulated instantaneous performance plotted as a function of trial number (defined as the percentage of choices towards the option with higher pre-defined reward probability in a 10 trial window) of an agent using an S_i*R-E[R] learning rule in 80:20 (orange) and 100:0 (red) reward conditions.

B As (A), except for an agent using a non covariance rule, simply R.

C Flies' instantaneous performance when providing optogenetic reward using the sugar sensing neuron driver Gr64F-Gal4 to drive UAS CSChrimson. Again reward ratios are 80:20 (n = 6 flies) and 100:0 (n = 9 flies).

D As in C except reward provided via the PAM DANs using R58E02-Gal4 to drive UAS-CSChrimson (n=8 flies in both 80:20 and 100:0).

E Performance scores of individual flies from C showing significant learning at both 100:0 and 80:20 paradigms (Wilcoxon signed-rank test: 100:0, p = 0.0039; 80:20, p = 0.0312)

F Performance scores of individual flies from D. Flies showed a significant preference towards the rewarded odor in 100:0 but not 80:20 (100:0, p = 0.0391; 80:20, p = 0.1875).

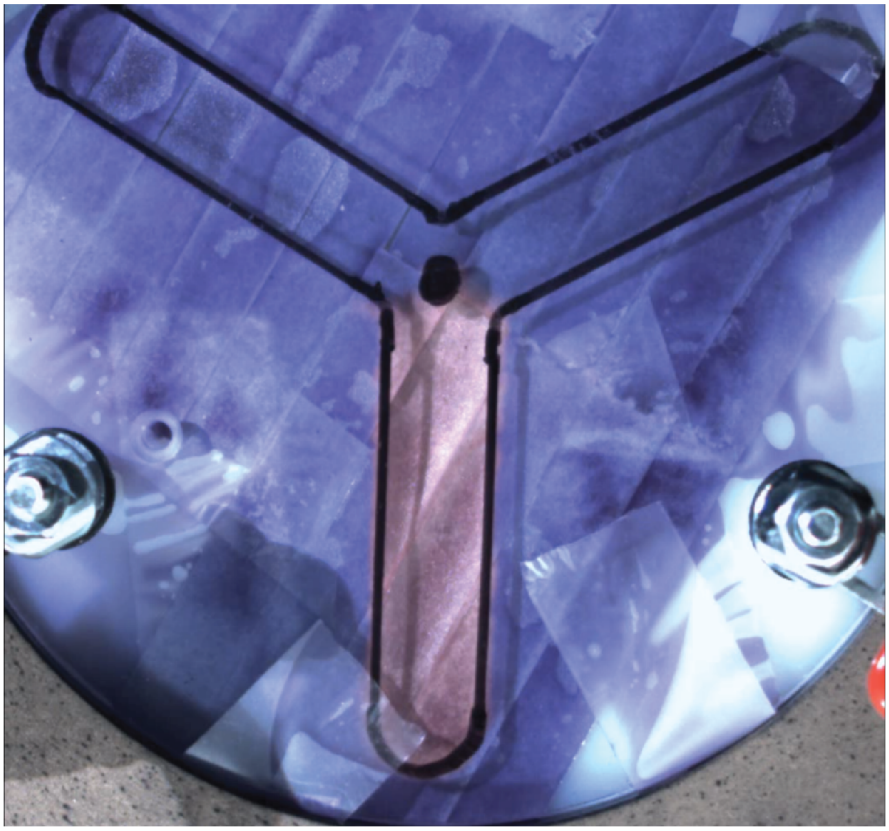
G Block-averaged choice ratios and reward ratios are plotted against each other (n = 26 flies * 3 blocks) when DAN activation is used as reward. The best fit line for the DAN reward data is in blue, and best fit line for Gr64f sugar sensory reward (Fig. 3C) is in red. Choice ratio is not equal to reward ratio suggesting that these flies are not following the matching law.

H The instantaneous choice ratio of an example fly plotted against trial number in a paradigm with dynamic reward contingencies *(bottom)* as in Fig. 3A. The reward contingencies in each block for this example fly are depicted with the help of schematics of Y arenas mapping reward probability to odor *(top)*. Individual choices made by the fly are indicated by long and short rasters for rewarded and unrewarded choices respectively with colors representing the chosen odor. Blue curve: 10 trial window-averaged instantaneous choice ratio. Black curve: instantaneous reward ratio. Blue and black curves overlap as expected by matching. Transparent blue lines: block-averaged choice ratio. Gray lines: block-averaged reward ratio.

I Block-averaged choice ratios and reward ratios (n = 50 simulations * 3, 80 trial blocks) from data simulated using a non-covariance-based rule are plotted against each other. The best fit line for the simulated data is in black, and the best fit line for DAN reward (same as panel G) is in blue.

Supplementary Figure 1

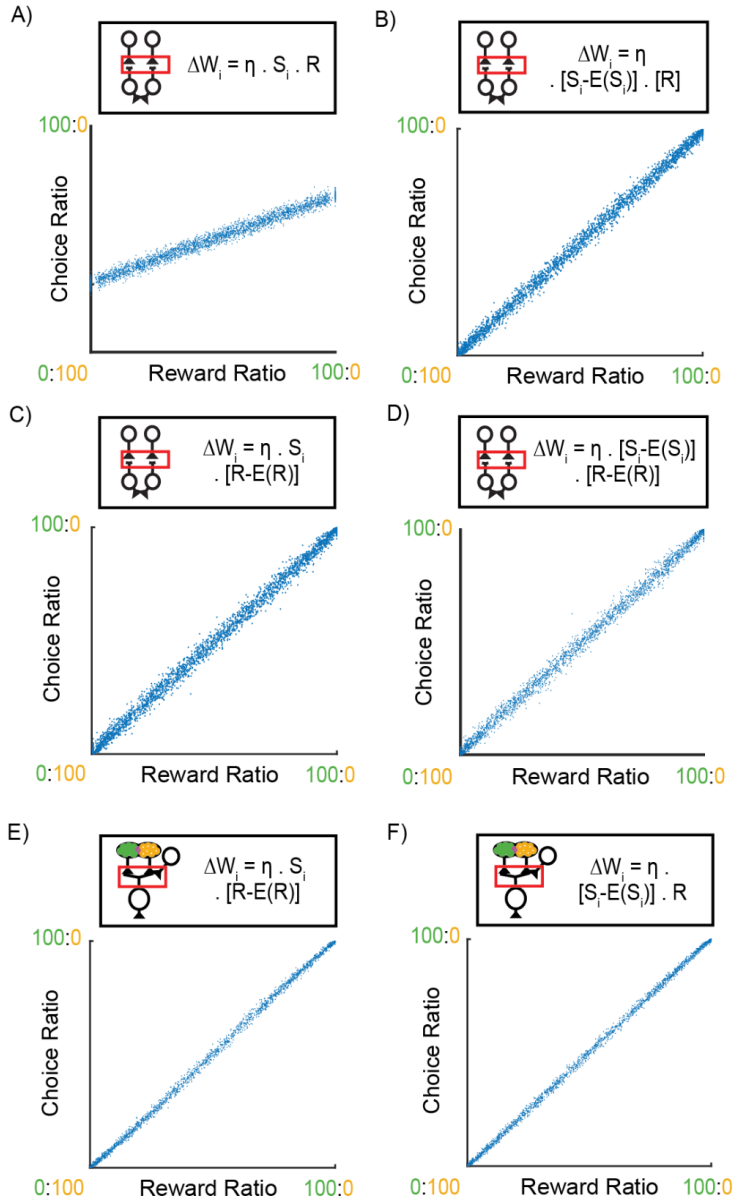
A)



Supplementary Figure 1: Visualizing odor boundaries at the center of the Y-arena

A A test of the crispness of odor boundaries in the Y arena. Moist blue litmus paper was placed in the Y arena while the arm at the bottom was filled with with carbon dioxide. This caused the color of the litmus paper to change providing an estimate of how odor boundaries are formed at the center of the Y.

Supplementary Figure 2



Supplementary Figure 2: Covariance-based learning rules are necessary for operant matching

A Block-averaged choice ratios versus reward ratios (n = 2000 simulations) from data simulated using a non-covariance rule in the model described in 4A are plotted against each other.

B Block-averaged choice ratios versus reward ratios (n = 2000 simulations) from data simulated using a covariance-based rule $[S_i-E(S_i)]*R$ in the model described in 4A are plotted against each other.

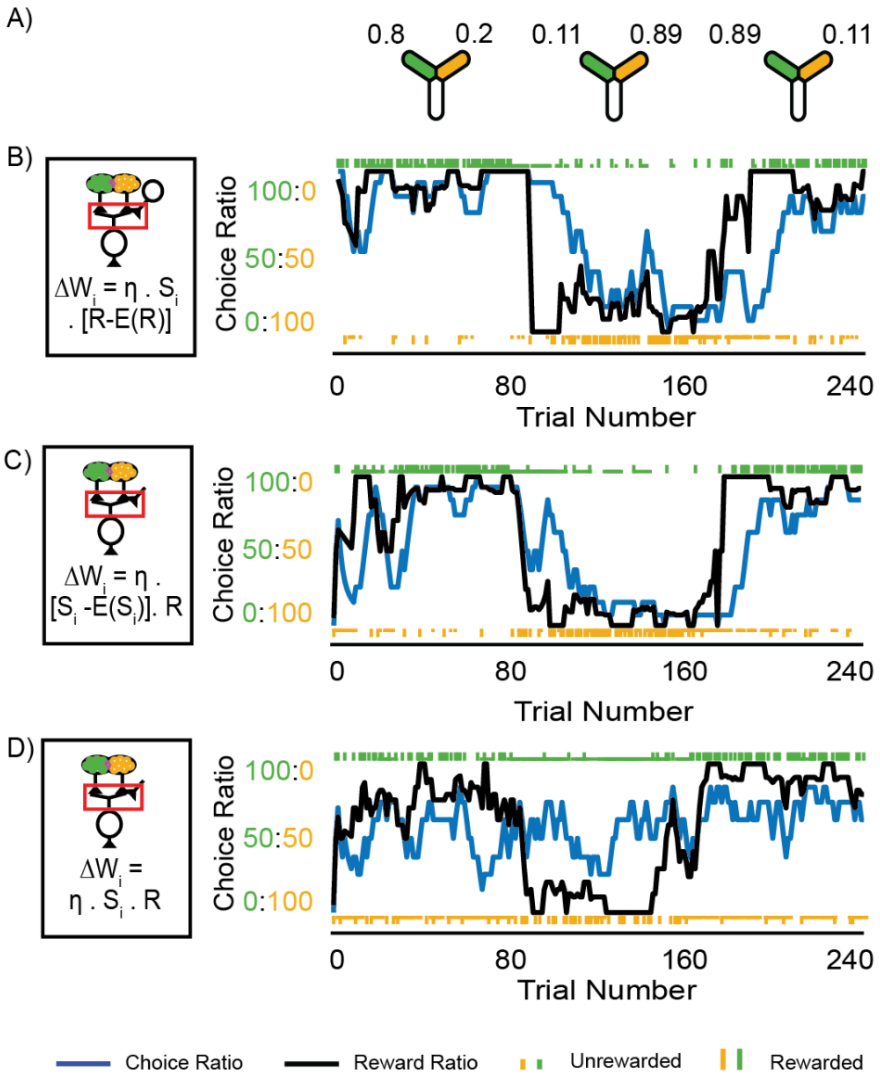
C Block-averaged choice ratios versus reward ratios (n = 2000 simulations) from data simulated using a covariance-based rule $S_i*[R-E(R)]$ in the model described in 4A are plotted against each other.

D Block-averaged choice ratios versus reward ratios (n = 2000 simulations) from data simulated using a covariance-based rule $[S_i-E(S_i)]*[R-E(R)]$ in the model described in 4A are plotted against each other.

E Block-averaged choice ratios versus reward ratios (n = 2000 simulations) from data simulated using a covariance-based rule $[S_i-E(S_i)]*R$ in the model described in 4C are plotted against each other.

F Block-averaged choice ratios versus reward ratios (n = 2000 simulations) from data simulated using a covariance-based rule $S_i*[R-E(R)]$ in the model described in 4C are plotted against each other.

Supplementary Figure 3



Supplementary Figure 3: Models using covariance-based learning rules produce behavior more similar to real fly behavior

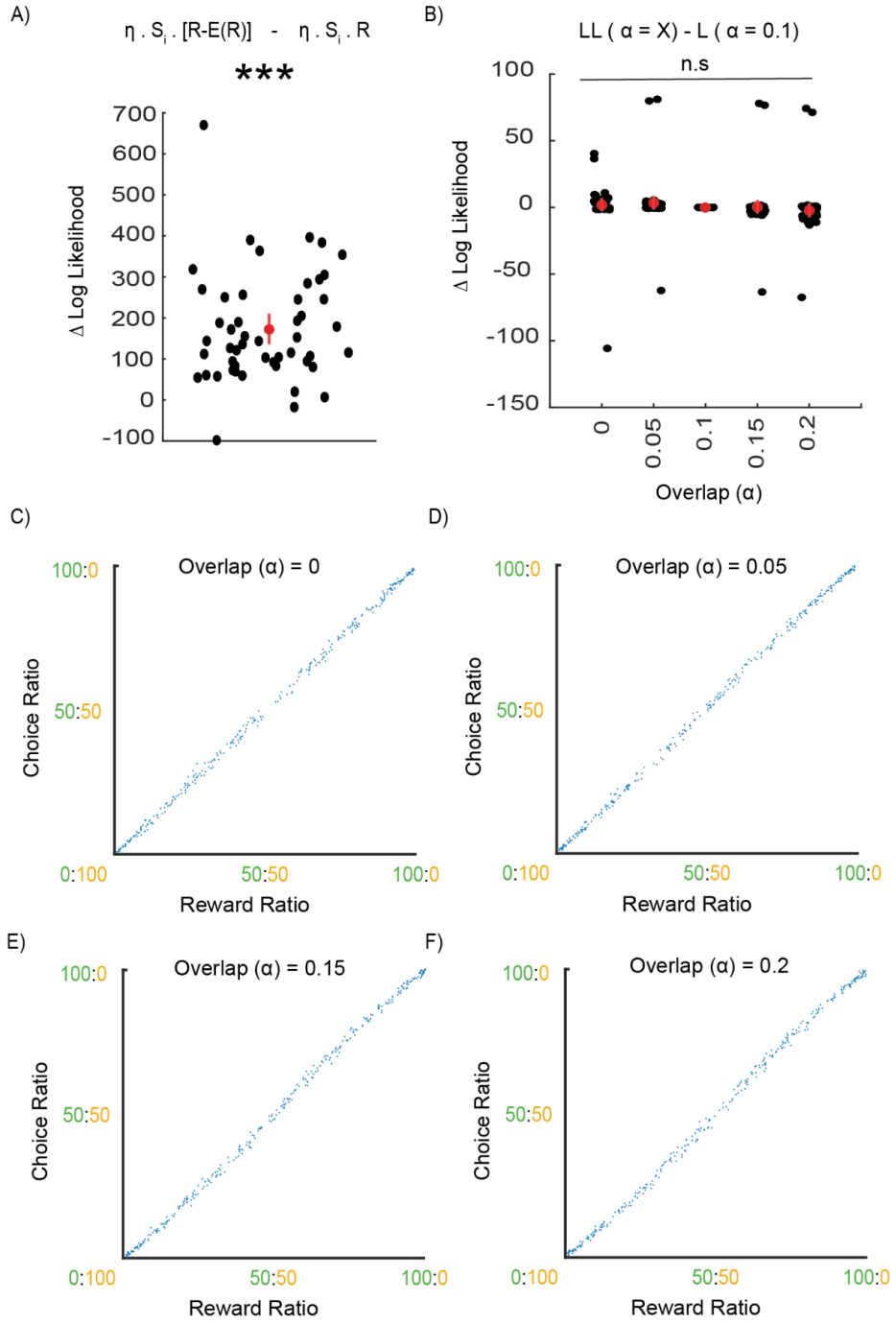
A Reward contingencies for the example task simulated in B, C and D.

B An example simulation showing the performance in a 3 block task of the model described in 4C incorporating a covariance-based rule S_i*[R-E(R)].

C An example simulation showing the performance in a 3 block task of the model described in 4C incorporating a covariance-based rule [S_i-E(S_i)]*R.

D An example simulation showing the performance in a 3 block task of the model described in 4C incorporating the non-covariance rule .

Supplementary Figure 4



Supplementary Figure 4: Extent of sensory overlap does not affect the behavior of our model

A Change in log-likelihood (LL) metric, computed by subtracting the LL of the non-covariance-based regression model from a covariance-based rule, plotted for each simulation (n = 50). Data was simulated using the covariance-based rule S_i*[R-E(R)], and as expected this rule better fit the simulated behavior than the non covariance-based rule.

B Change in log-likelihood (LL) metric, computed by subtracting the LL of the covariance-based regression model with overlap = 0.1, from that of regression models with covariance-based rules and different amounts of overlap (0,0.05,0.15,0.2). Points are plotted for each simulation (n = 50). Simulations were run with overlap α = 0.1.

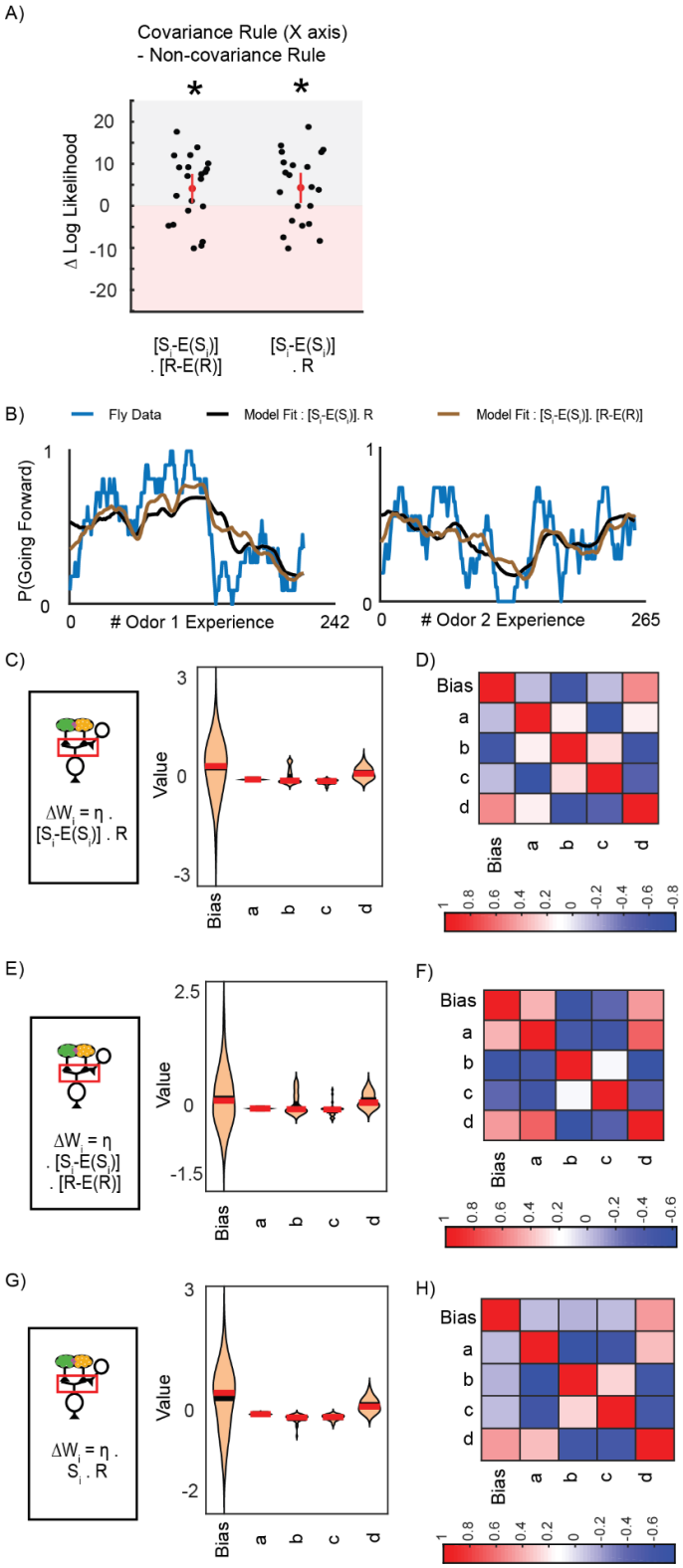
C Block-averaged choice ratios versus reward ratios (n = 300 simulations) from data simulated using a covariance-based rule S_i*[R-E(R)] in the model described in 4C are plotted against each other. Sensory overlap α = 0.

D Block-averaged choice ratios versus reward ratios (n = 300 simulations) from data simulated using a covariance-based rule S_i*[R-E(R)] in the model described in 4C are plotted against each other. Sensory overlap α = 0.05.

E Block-averaged choice ratios versus reward ratios (n = 300 simulations) from data simulated using a covariance-based rule S_i*[R-E(R)] in the model described in 4C are plotted against each other. Sensory overlap α = 0.15.

F Block-averaged choice ratios versus reward ratios (n = 300 simulations) from data simulated using a covariance-based rule S_i*[R-E(R)] in the model described in 4C are plotted against each other. Sensory overlap α = 0.2.

Supplementary Figure 5



Supplementary Figure 5: Covariance-based learning rules are better predictors on individual choice behavior

A Change in log-likelihood (LL) metric, computed by subtracting the LL of the non-covariance-based model from two models with covariance-based rules, plotted for each fly (n = 21) .

B Example fly data (blue) fit using an MB-inspired regression model (Fig. 5A) that incorporates either an $[S_i-E(S_i)]*[R-E(R)]$ rule (brown), or an $[S_i-E(S_i)]*R$ rule (black).

C Regression coefficients assigned to each term of the learning rule when the MB-inspired regression model using an $[S_i-E(S_i)]*R$ rule was fit to the flies behavior.

D Correlation between regression coefficients resulting from MB-inspired regression model using an $[S_i-E(S_i)]*R$ rule.

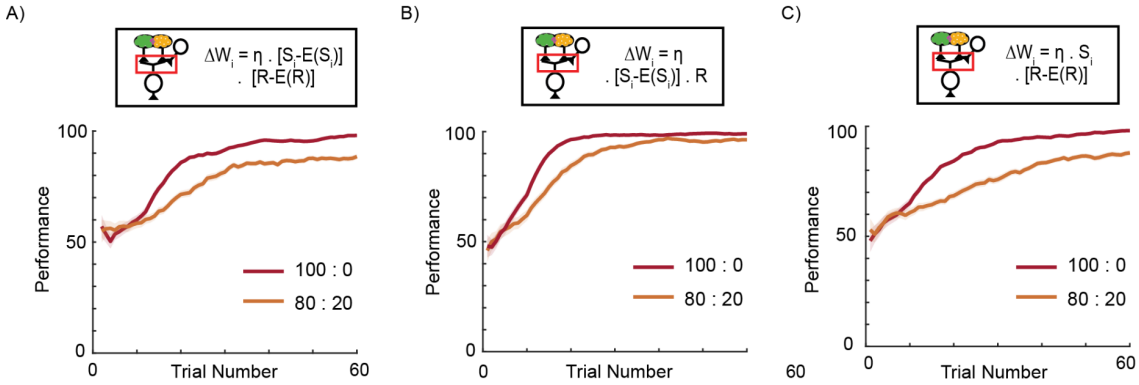
E Regression coefficients assigned to each term of the learning rule when the MB-inspired regression model using an $[S_i-E(S_i)]*[R-E(R)]$ rule was fit to the flies behavior.

F Correlation between regression coefficients resulting from MB-inspired regression model using an $[S_i-E(S_i)]*[R-E(R)]$ rule.

G Regression coefficients assigned to each term of the learning rule when the MB-inspired regression model using an S_i*R rule was fit to the flies behavior.

H Correlation between regression coefficients resulting from MB-inspired regression model using an S_i*R rule.

Supplementary Figure 6



Supplementary Figure 6: Covariance-based learning rules produce similar behavior in 100:0 and 80:20 tasks

A Simulated instantaneous performance plotted as a function of trial number (defined as the percentage of choices towards the option with higher pre-defined reward probability in a 10 trial window) of an agent using an $[S_i-E(S_i)]*[R-E(R)]$ learning rule in 80:20 (orange) and 100:0 (red) reward conditions.

B Simulated instantaneous performance plotted as a function of trial number (defined as the percentage of choices towards the option with higher pre-defined reward probability in a 10 trial window) of an agent using an $[S_i-E(S_i)]*R$ learning rule in 80:20 (orange) and 100:0 (red) reward conditions.

C Simulated instantaneous performance plotted as a function of trial number (defined as the percentage of choices towards the option with higher pre-defined reward probability in a 10 trial window) of an agent using an $S_i*R-E[R]$ learning rule in 80:20 (orange) and 100:0 (red) reward conditions.