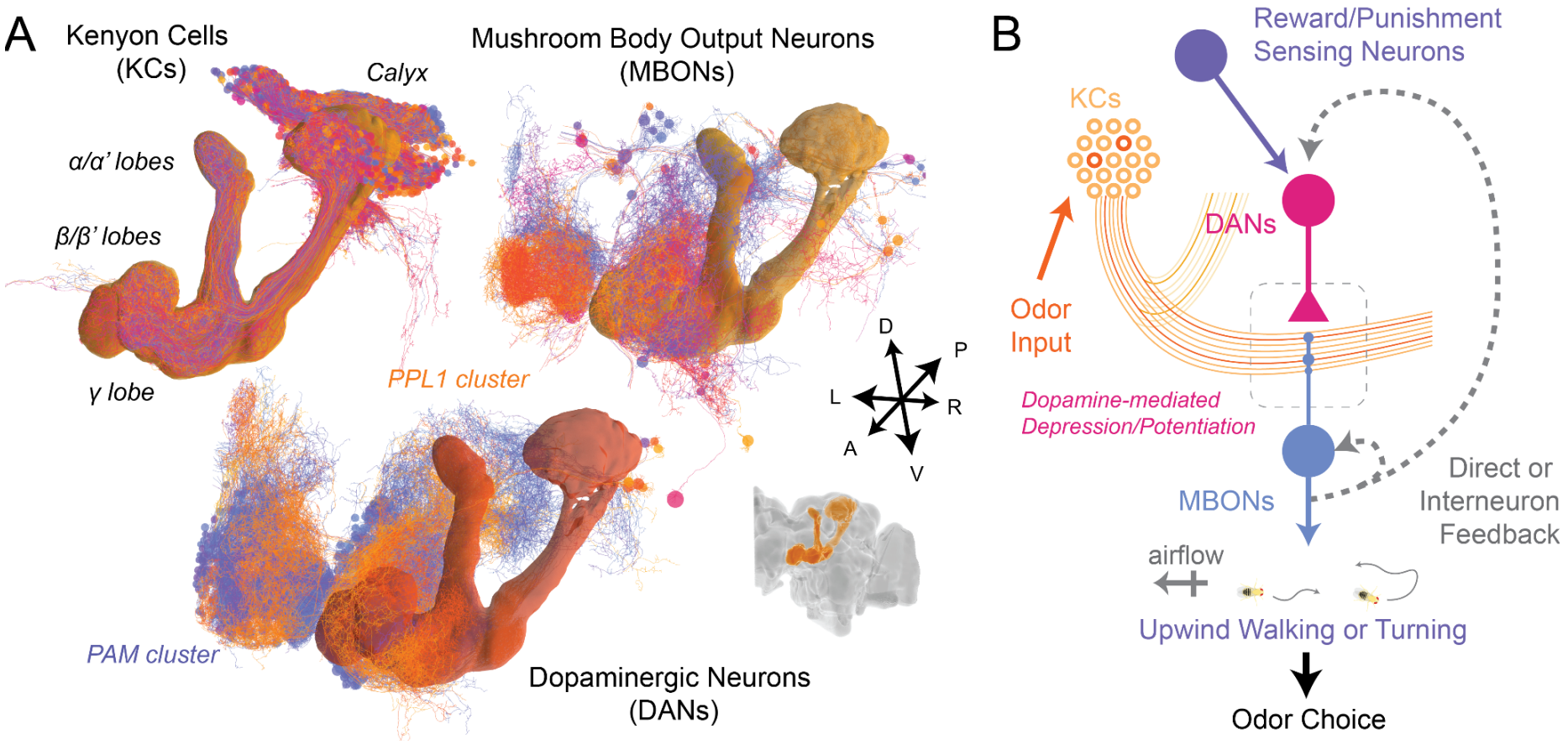


Figure 1. Mushroom Body of the Fruit fly (*Drosophila melanogaster*).

The three main neuronal cell types of the fruit fly mushroom body: i. Kenyon cells receive input from upstream odor circuits at the calyx; ii. Mushroom body output neurons that receive input from the Kenyon Cells; iii. Dopaminergic neurons (typically reward-sensitive PAM cluster and punishment-sensitive PPL1 cluster) that modulate KC-MBON synapses.

(A) 3D neuron reconstruction of mushroom body neurons from Hemibrain v1.2.1 electron microscopy dataset rendered using navis 1.3.1 and plotly.

(B) A simplified circuit schematic for the fruit fly mushroom body. Yellow: Inactive KCs; Orange: Odor-activated KCs; Purple: Reward/Punishment sensing neurons; Pink: Reward/Punishment sensitive dopaminergic neurons; Blue: Aversive/Appetitive MBONs; Grey: Direct/Indirect feedback connections. Note that each “neuron” in the schematic represents a population of neurons shown in subfigure A.

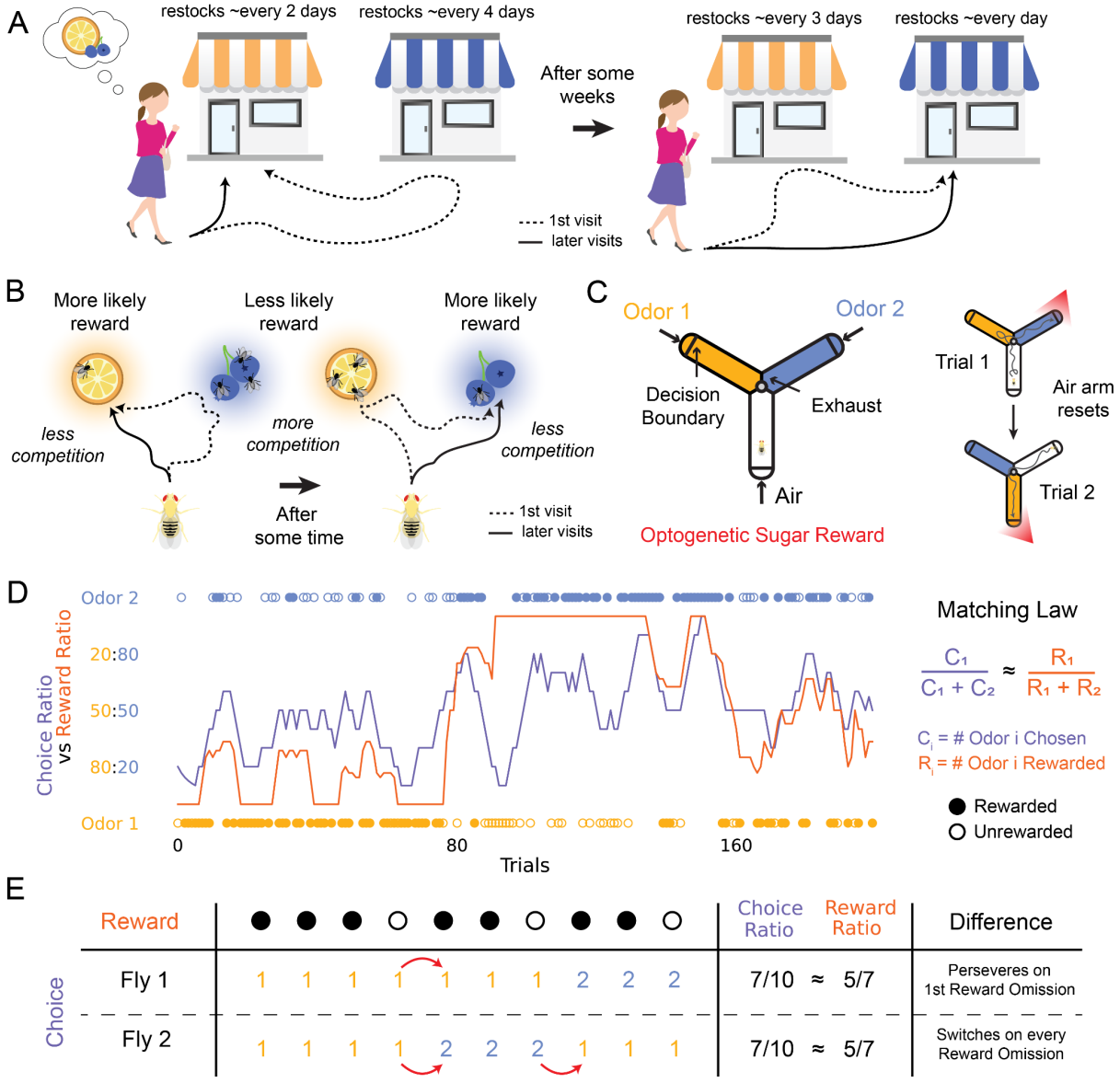


Figure 2. Foraging as a 2AFC Task and the limitations of the Matching Law.

(A) Humans face foraging challenges in daily life. Consider someone looking for fresh fruits but have two comparable options for grocery stores that they can visit, but the two stores restock supplies at a different (unknown) frequency. Therefore, the probability of finding fresh fruits will differ between the two stores and can be estimated by the person after a few visits allowing them to make better decisions about which store to visit. However, the restocking frequency might change after a few weeks, and the person has to update their estimates to make the best choices.

(B) Flies, too, can face dynamically changing reward probabilities during foraging as they might have to compete with other individuals for limited resources. For example,

consider a fly with two possible food sources: lemons and blueberries. A naive fly (dotted line) visits the lemons to find many competitors and receives a reward with low probability. Then, on finding the blueberries learns that the blueberries have fewer competitors and more probability of reward (solid line). As more flies do the same, the distribution of competitors changes, and the fly must learn to switch to the lemons for more reward.

(C) The decision-making process underlying foraging can be replicated in an artificial Y-maze with two odorized and one clean-air arm. Each trial is completed when the decision boundary on an odorized arm is crossed. The relative orientations of the odor arms are randomized to ensure flies do not learn directional associations. A probabilistic reward is delivered through optogenetic activation of sugar-sensing neurons.

(D) Flies show operant matching behavior. Operant matching law is an optimal strategy for foraging where the choices closely follow the same ratio as the rewards received for the different choices (right). Figure reproduced with data from Rajagopalan et al., 2022, with permission. Orange and Blue dots in the reward schedule represent choosing Odor 1 and 2, respectively. Filled and empty dots represent the rewarded choice and unrewarded choices, respectively. The lines represent the reward and choice ratios calculated for 10 trials till the current trial (including the current trial).

(E) A toy example of the limitation of the matching law. See main text. Column 2 provides the reward and choice sequence between odor 1 (orange) and odor 2 (blue). Column 3 shows the estimate of choice and reward ratios. Red arrows highlight transitions in chosen odors.

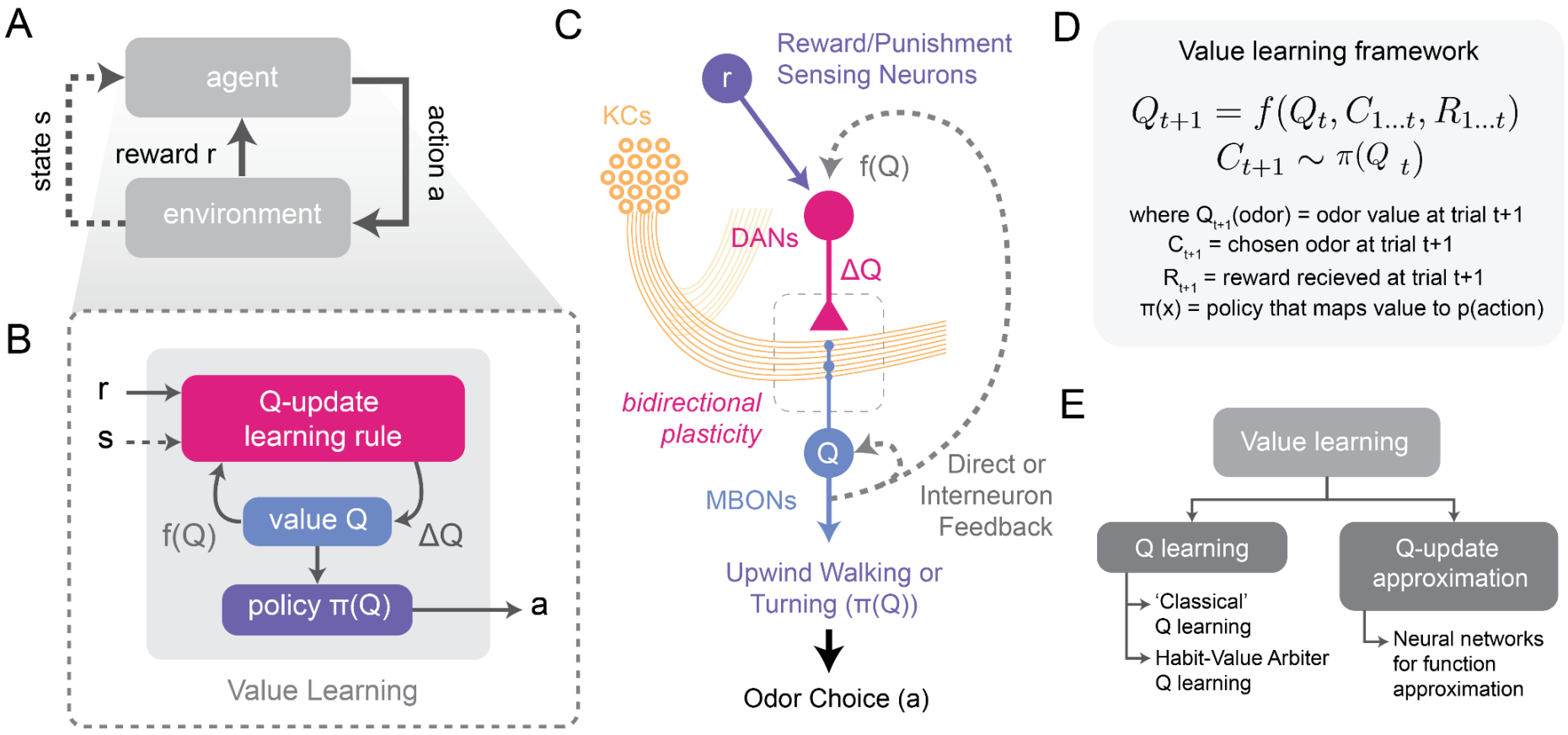


Figure 3. Reinforcement Learning in the Fly Brain through Value Learning.

(A) The Reinforcement Learning (RL) Framework. The agent receives information from the environment in the form of the outcomes for past actions (reward r; can be +ve or -ve) and the world's current condition (state s; can be a high dimensional input).

Using this information, the agent chooses the best action (a) to perform in order to maximize its reward. In turn, the environment receives the action, updates the state, and gives the appropriate reward to the agent.

(B) Value learning is a type of RL framework that involves three major elements: i. Value (Q) - a measure of how much reward an animal expects given the state and action; ii. Policy (π(Q)) - a function that transforms the value to a probability of taking any action and determines the action taken by the animal; iii. Q-update Reinforcement Learning Algorithm that updates the value of the state and action, given the information from the environment.

(C) Mapping action value learning to the fruit fly MB. MBON activity during odor exposure encodes stimulus valence and, therefore, can represent the action value of choosing an odor (Q). The KC → MBON synaptic weights are updated bidirectionally by DANs, allowing value updation (δQ). DANs receive reward/punishment signals (R) from sugar/bitter/shock-sensing neurons. MB-intrinsic and MB-extrinsic interneuronal circuits can provide complex feedback (F(Q)) from the MBONs to the DANs. Thus, DANs can integrate reward signals and feedback to implement complex learning rules. The downstream circuitry then transforms the value code to behavioral patterns such as upwind walking and turning that result in the choice outcome, i.e., the policy (π(Q)).

(D) Functional form of the value learning framework for odor preference. The first equation represents how past choices and reward history is integrated with the past value to get the new updated value. The second equation represents how the policy transforms the value into choice distribution.

(E) Variations of value learning. Q-Learning is the most common form of value learning. Over the last two decades, many variations of Q-learning have been developed to explain animal behavior. We divide them into two categories: i. classical Q-learning; ii. habit-value arbiter Q-learning. We also develop a novel modeling framework to infer learning rules from behavior which we call Q-update approximation.

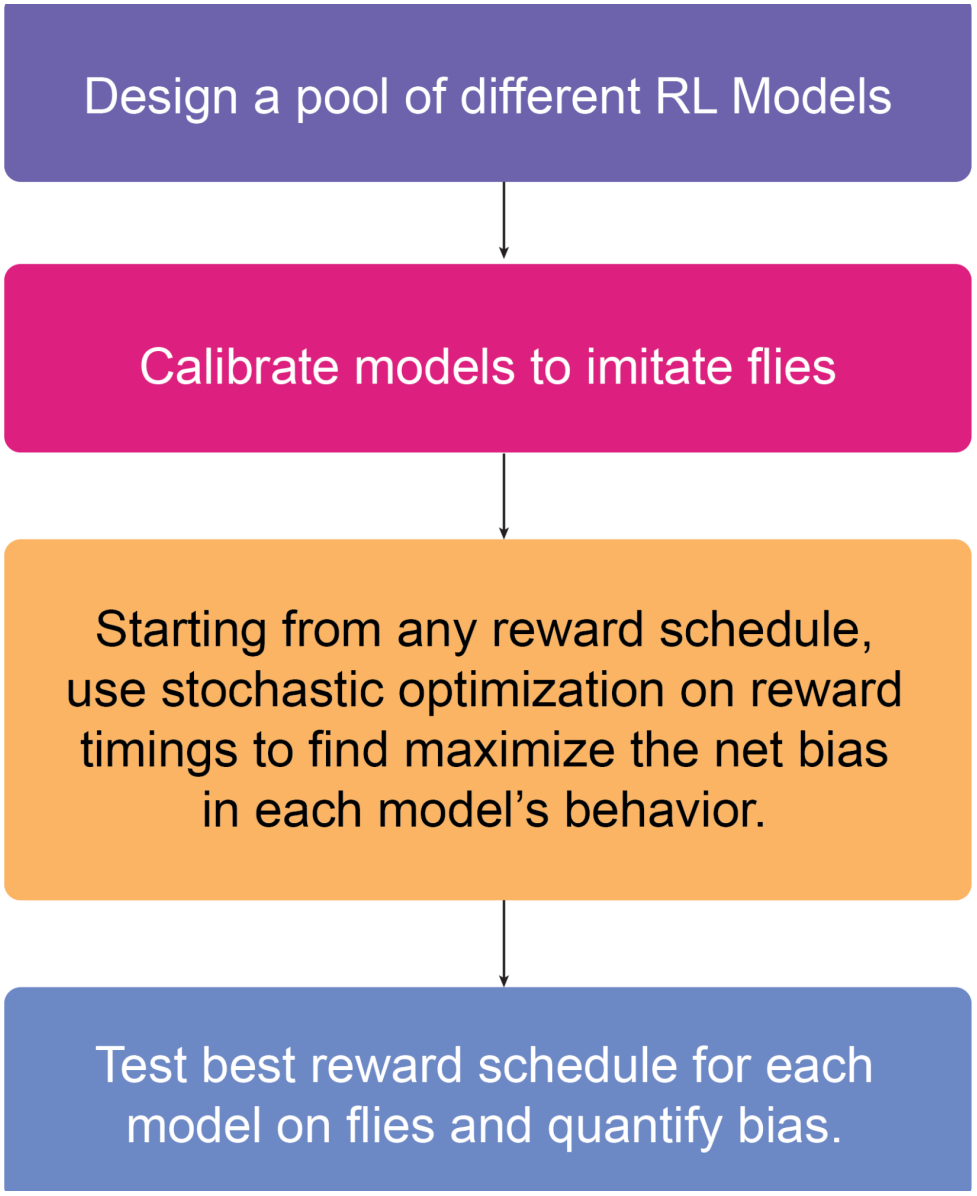


Figure 4. Choice Engineering Paradigm.

A *reward schedule* is a series of choice-reward outcomes for both odors that the fly can choose. The' optimal' reward schedule for choice engineering is the series of odor-reward associations that maximizes the number of choices made towards a preferred side, providing a predictable behavioral perturbation that can be tested on fruit flies. We sample the space of the reward structures using stochastic optimization techniques to find the optimal reward schedule.

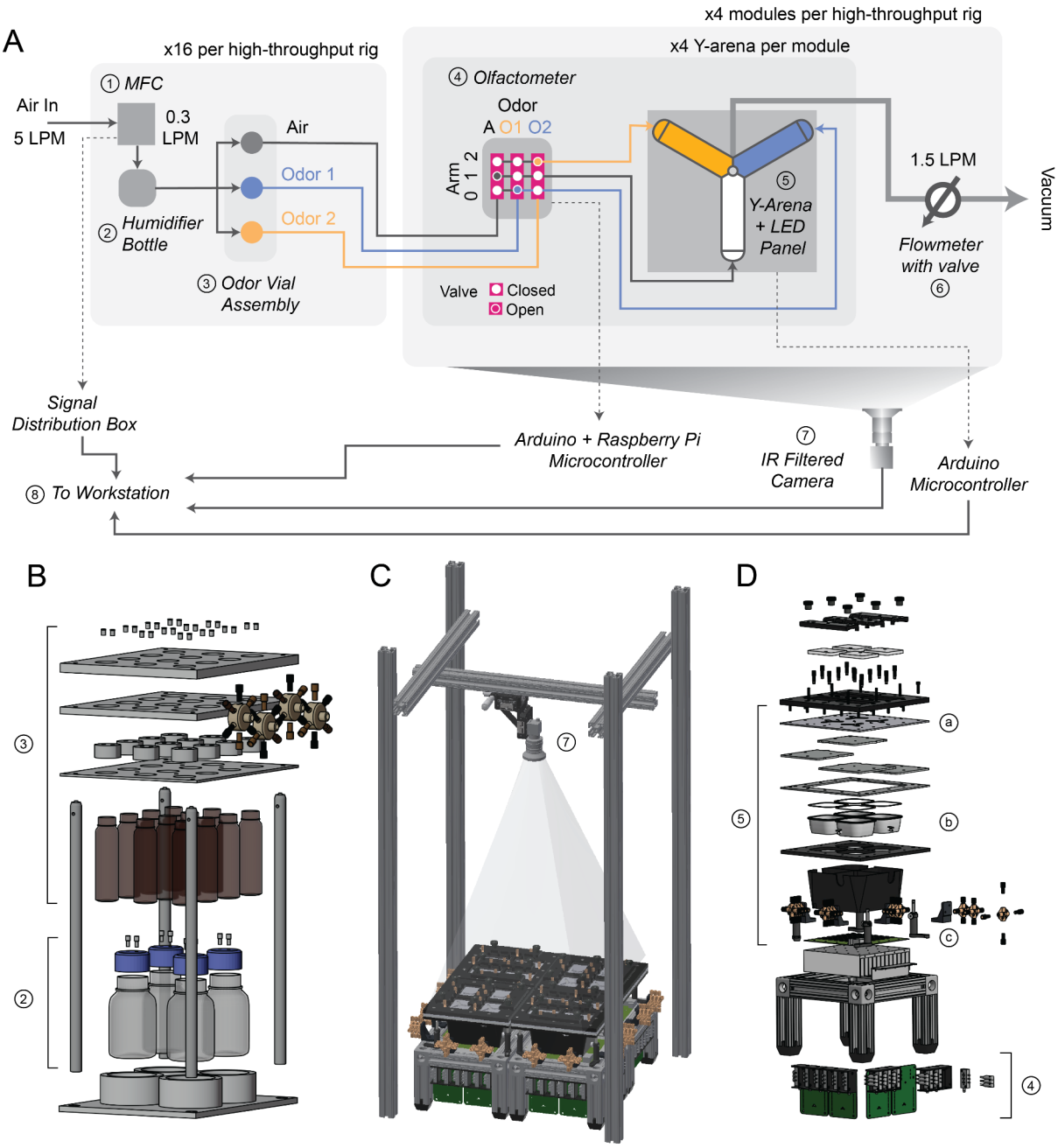


Figure 5. High-level schematic of the 16Y high-throughput Y-maze behavioral rig.

(A) Overall schematic of the 16Y assembly. Four Humidifier bottles (1 per arena) and 12 (3 per arena) were combined as a single ‘Odor Distribution Assembly’. Four Olfactometers and a single LED Panel are combined with four Y-arena chambers to form a single ‘4Y Module’. Four identical 4Y-Modules and Odor Distribution Assemblies combine to form the entire 16Y experimental rig.

(B) Exploded View of an Odor Distribution Assembly. (2) represents the four Humidifier bottles and Holders that supply humidified air for four arenas; (3) represents the Odor Vial Assembly that splits the humidified airstream to create three parallel air streams that can be odorized. Saptarshi Soham Mohanta (Turner Lab, HHMI Janelia Research Campus, Virginia, USA) designed all parts of the module parts.

(C) Overall Arrangement of the 4Y Module and Camera for the 16Y Setup. The camera (7) is placed on a 3D micromanipulator using support beams over the 4Y Modules. The placement is made such that all 16 Arenas are in a common field of view.

(D) Exploded View of the 4Y Module. (4) represents the four olfactometers for the four Y-arenas; (5) represents the four combined Y-Arena chambers (a,b) and LED Panels (c). All module parts were designed at jET (Janelia Experimental Team, Virginia, USA).

Versilon SE 200 1/8” OD x 1/16” ID and 1/4” OD x 1/8” ID (McMaster Carr, Illinois, USA) tubing were used to direct airflow at all points. The entire rig is placed inside a dark, thermally insulated box in a temperature-controlled room. All experiments are done at 24-26°C.

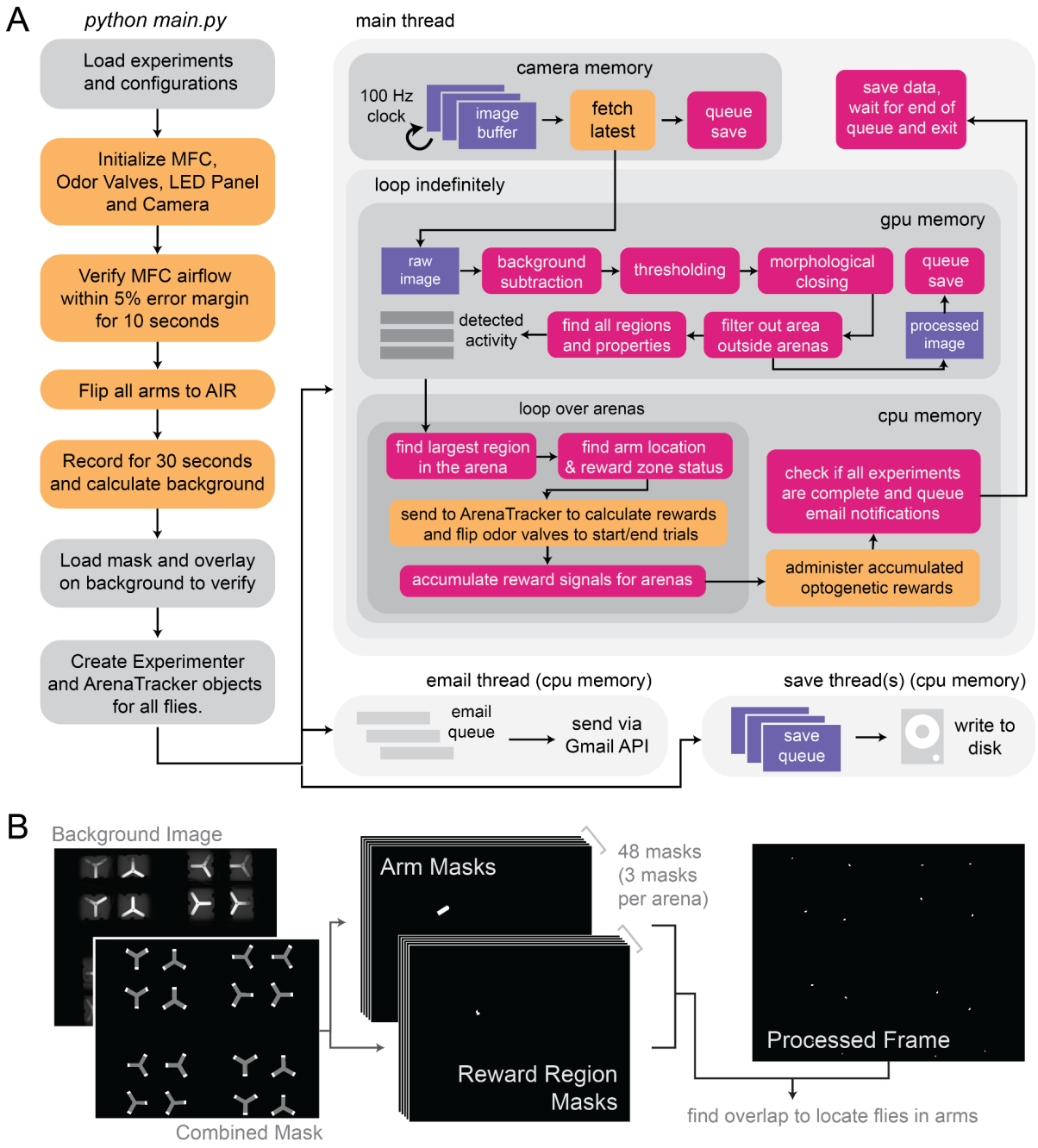


Figure 6. Schematic of closed-loop control for running parallel experiments.

(A) Operational loop for running an experiment using the main.py Python script in the sixteeny Python package. Gray represents preparatory steps, Orange represents hardware interfacing steps, Purple represents image data in the memory, and Pink represents closed-loop control steps. For a summary of the control flow, see the main text.

(B) Schematic of Mask-based Filtering and Localization. To quickly find the current position of every fly in different arms and reward zones, we use a system of 96

masks (48 for reward regions with one per reward zone on each of three arms on 16 arenas & 48 for each of three arms on sixteen arenas). A combined mask (left) can be used to filter activity on the processed frame (right). Looking for overlap between each detected blip and the Arm and Reward Region masks (center) allows us to efficiently identify the location and reward zone status (whether the fly is in a reward zone).

Code available at: https://github.com/neurorishika/TurnerLab_Opto2AFC_16Y

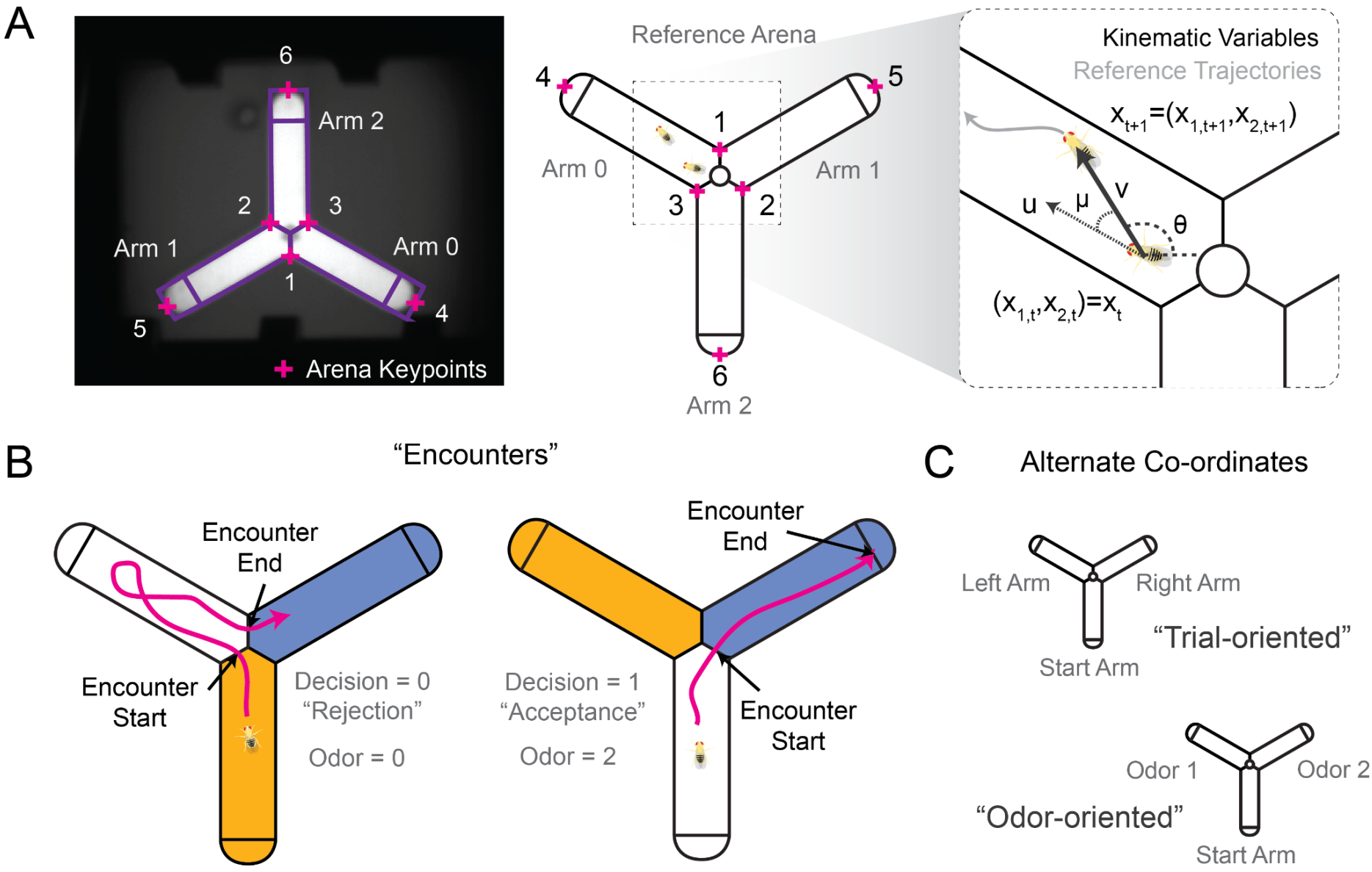

A
6
Arm 2
2
3
Arm 1
Arm 0
1
5
4
Arena Keypoints
Reference Arena
4
5
1
Arm 0
Arm 1
3
2
6
Arm 2
Kinematic Variables
Reference Trajectories
$x_{t+1}=(x_{1,t+1},x_{2,t+1})$
u
μ
v
θ
$(x_{1,t},x_{2,t})=x_t$
B
"Encounters"
Encounter End
Encounter Start
Decision = 0
"Rejection"
Odor = 0
Encounter End
Decision = 1
"Acceptance"
Odor = 2
Encounter Start
C
Alternate Co-ordinates
Left Arm
Right Arm
"Trial-oriented"
Start Arm
Odor 1
Odor 2
"Odor-oriented"
Start Arm

Figure 7. Post-hoc processed variables for high-throughput Y-maze data.

(A) Calculation of a reference coordinate system. Each arena is characterized using six key points: 3 points at the intersection of the arms and the three ends of the arms (left). These key points can be used to generate an affine transform to a reference coordinate system (middle). In the reference coordinate system, different kinematic variables can be calculated from the reference trajectories (right), such as speed (v), the direction of motion (θ), upwind speed (u), upwind motion direction (μ) by using the information about the change in position (x_t). See Table 5 for more details.

(B) *Encounters* boundaries are defined as every time a fly experiences a different odor condition (including air). Boundaries can happen at the end of a trial (right; Encounter “Acceptance”) or if a fly enters and leaves with the trial not being completed (left; Encounter “Rejection”).

(C) Reference coordinate systems can be reoriented to align them with respect to the start arms and odor positions for easier comparison between trials.

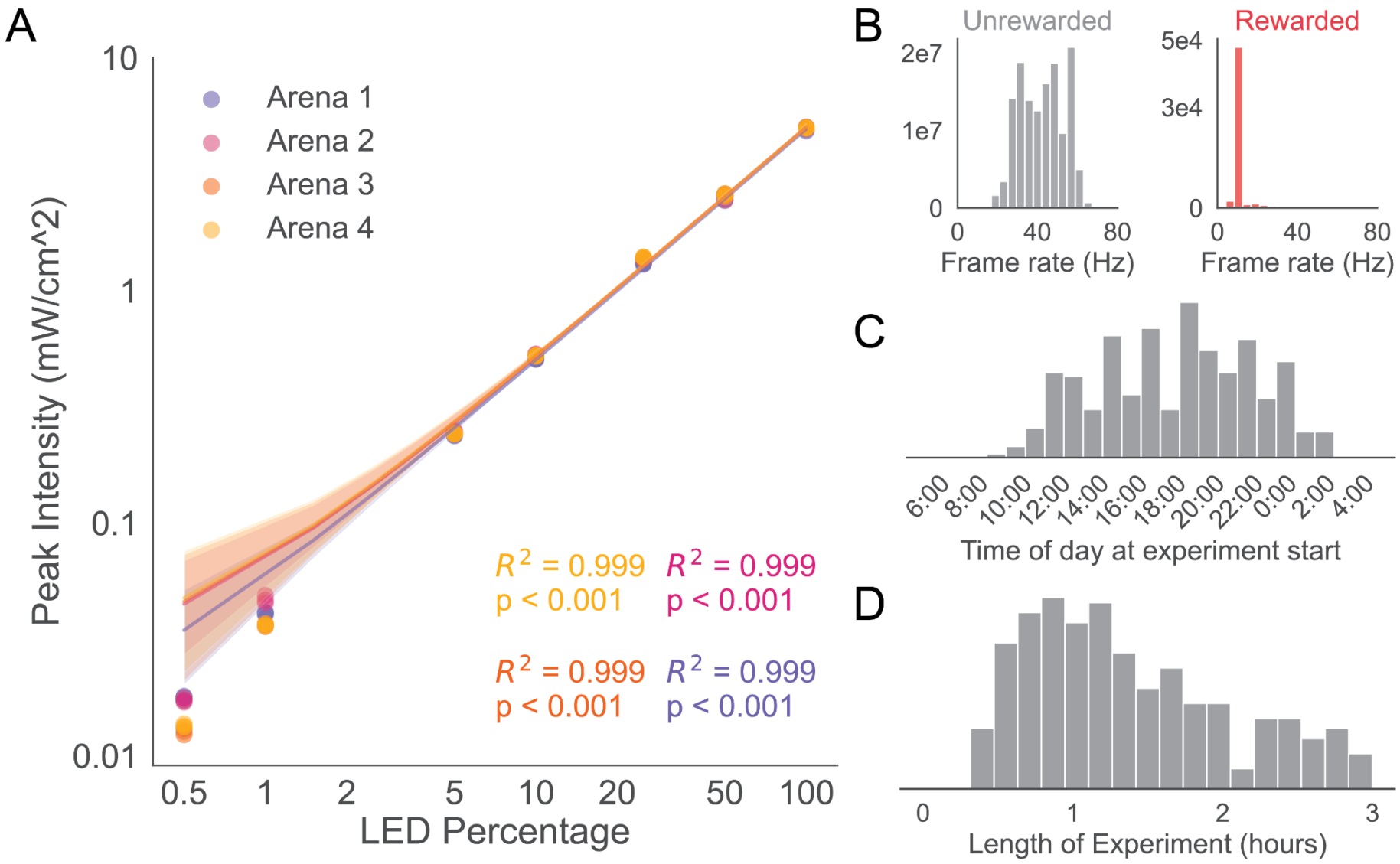

A
10
1
0.1
0.01
Peak Intensity (mW/cm^2)
Arena 1
Arena 2
Arena 3
Arena 4
R^2 = 0.999 R^2 = 0.999
p < 0.001 p < 0.001
R^2 = 0.999 R^2 = 0.999
p < 0.001 p < 0.001
0.5 1 2 5 10 20 50 100
LED Percentage
B
Unrewarded
2e7
1e7
0
0 40 80
Frame rate (Hz)
Rewarded
5e4
3e4
0
0 40 80
Frame rate (Hz)
C
6:00 8:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00 0:00 2:00 4:00
Time of day at experiment start
D
0 1 2 3
Length of Experiment (hours)

Figure 8. LED and Camera Calibration of the 4 Y-arenas used for experiments

(A) LED Power-Intensity log-log calibration curve comparison across 4 Y-arenas fitted with a linear fit (p=4.6e-32, 2.6e-31,1.29e-30, 5.35e-30).

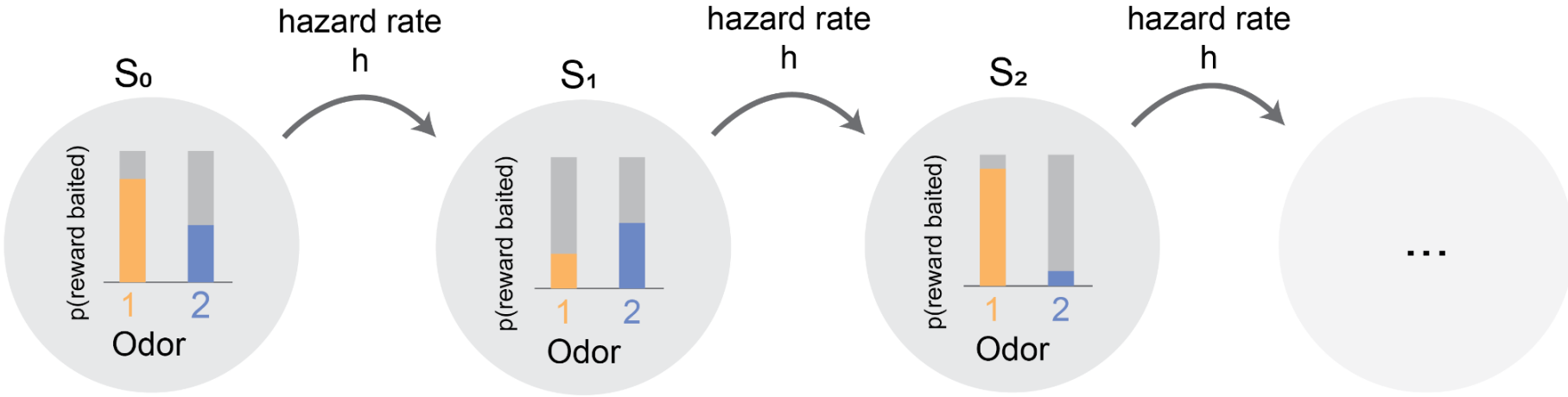
(B) Frame rate statistics for the experiments for rewarded and unrewarded frames.

(C) Histogram of the time of the day when experiments were run.

(D) Histogram of the duration of each experiment.

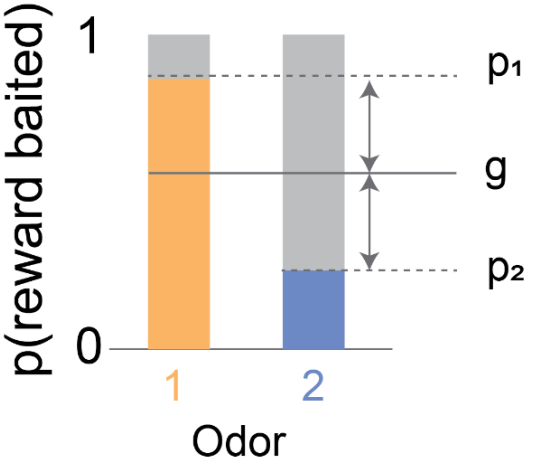
A

randomized non-stationary probabilistic foraging task



p(reward baited | odor 1 chosen, S_i) < p(reward baited | odor 2 chosen, S_i) if i is even
p(reward baited| odor 1 chosen, S_i) > p(reward baited | odor 2 chosen, S_i) if i is odd

B



reward gain g
= (p_1 + p_2)/2

reward contrast c
= p_1/(p_1 + p_2)

C

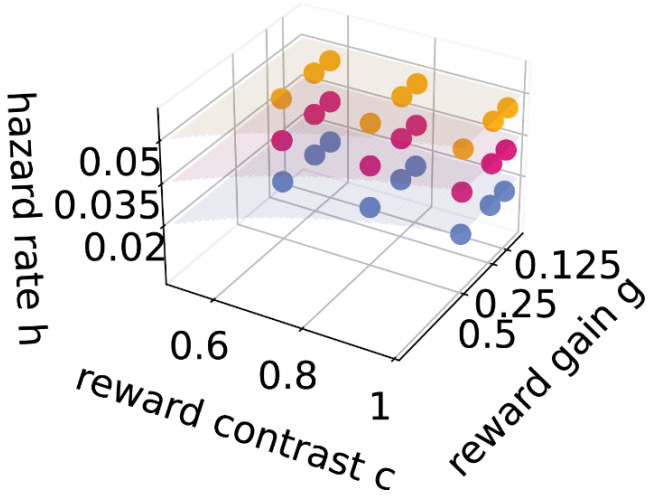


Figure 9. Design of “Variable Block” 2 Alternative Forced Choice (2AFC) experiments.

(A) State transitions in the Variable Block experiments. The state defines the reward-baiting probability for both odors. Markov-state updates happen at the end of each trial. Transitions to the next state (S_i to S_{i+1}) happen with a probability of h (referred to as the hazard rate). Alternatively, the experiment remains in the same state with a probability of 1-h. A *block* is defined as the trials where the state is conserved. Therefore, the length of a block is a geometric distribution. The odor associated with a greater reward baiting probability is always switched between the two states. Further, we rounded off the sampled block lengths to the nearest 5th trial. Within each state, the rewards are baited (see the section on Baiting above).

(B) Each state is characterized by two values: reward gain (g) and reward contrast (c) which scale the average reward rate and separation of value, respectively. The quantities together define the baiting probabilities for both odors.

(C) All experiments are sampled from the space of reward gain, reward contrast, and hazard rate. The hazard rate is kept constant for a session, but the reward gain and contrast are sampled independently for each block of trials. Reward gain is chosen to be either 0.5, 0.25 or 0.125; reward contrast is chosen to either 1.0, 0.8, and 0.6; Hazard rate is chosen to be 0.02, 0.035, 0.05.

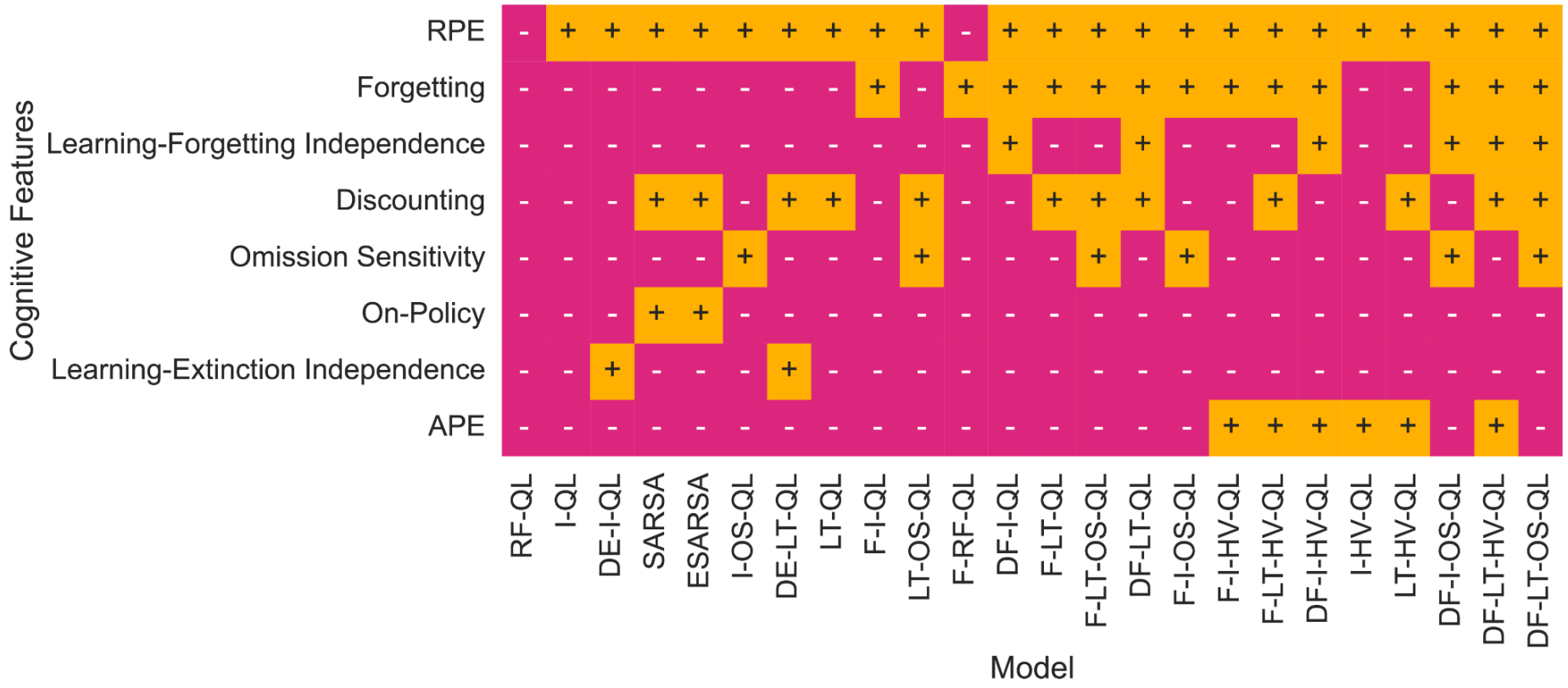


Figure 10. Map of Cognitive Feature to Model Identity. For a description of models and cognitive features take a look at Table 6 and Table 7.

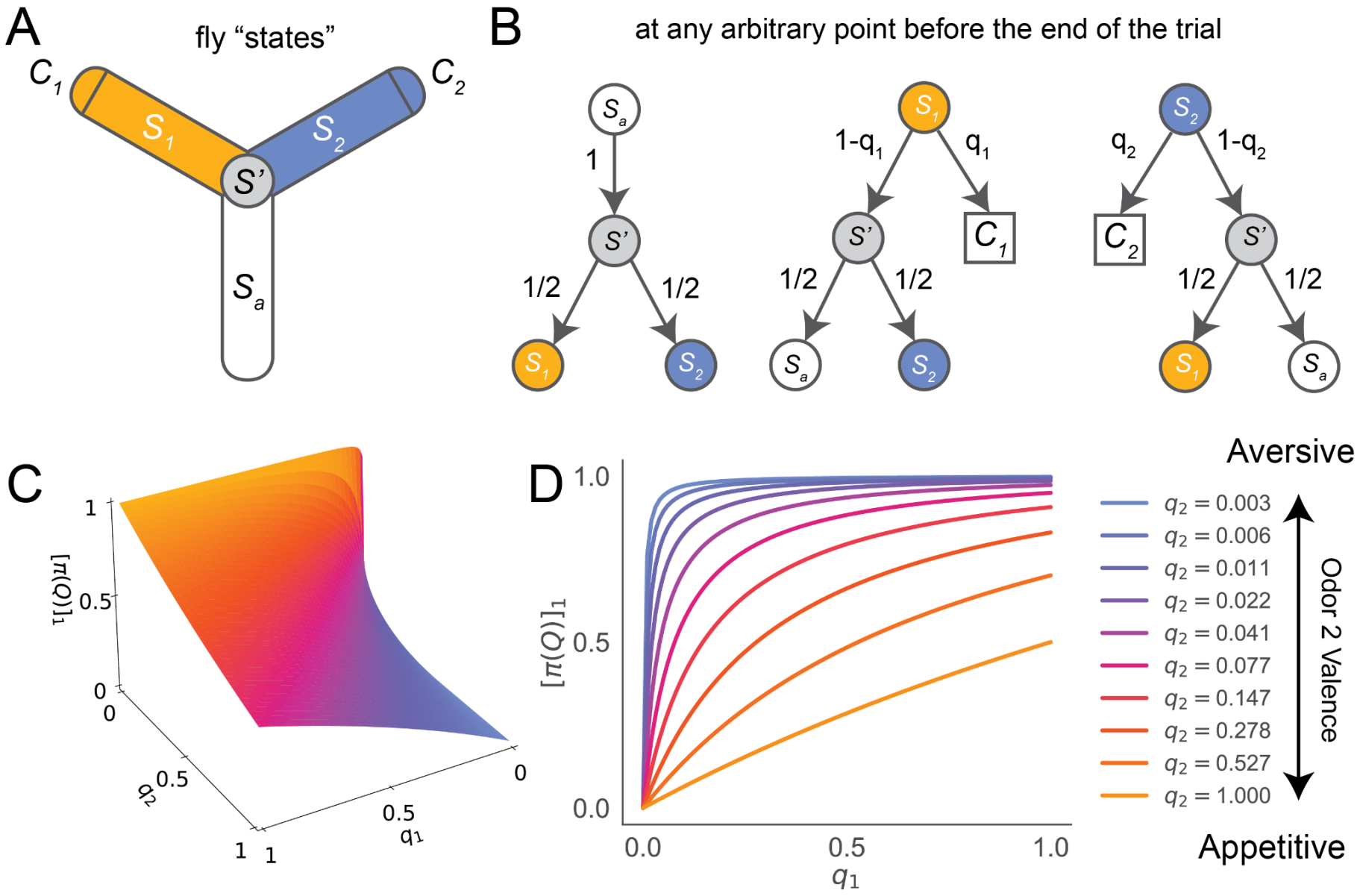

A
fly "states"
C_1
C_2
S_1
S_2
S'
S_a
B
at any arbitrary point before the end of the trial
1
1/2
1/2
1-q_1
q_1
1/2
1/2
q_2
1-q_2
1/2
1/2
C
$[\pi(Q)]_1$
q_2
q_1
D
$[\pi(Q)]_1$
q_1
q_2 = 0.003
q_2 = 0.006
q_2 = 0.011
q_2 = 0.022
q_2 = 0.041
q_2 = 0.077
q_2 = 0.147
q_2 = 0.278
q_2 = 0.527
q_2 = 1.000
Aversive
Odor 2 Valence
Appetitive

Figure 11. Derivation and Characterization of the Accept-Reject Policy.

(A) At any point in time, the fly can be in one of 4 areas of the Y-Maze, which we define as fly "states," i.e., the air arm (S_a), the odor 1 arm (S_1), the odor 2 arm (S_2) or at the decision boundary (S'). Further, a trial only terminates with a choice state where odor 1 is chosen (C_1) or odor 2 is chosen (C_2).

(B) Let q_i be the probability that odor i = 1 or 2 is accepted. Assuming that once the fly reaches the decision boundary, it will necessarily enter one of the two other arms with equal probability, we can define all possible transitions starting from any of the arms. Circles represent the arm in which the fly is. Square represents a choice that leads to the termination of the trial.

(C) A 3D plot of the odor 1 choice probability ($[\pi(Q)]_i$; i = 1) in terms of the acceptance probability of the two odors (q_1 and q_2). Note the non-linear response of the function that allows for both exploratory behavior (choice probability close to 0.5) and greedy behavior (choice probability close to 1) with small changes in acceptance probability.

(D) A cross-section of the policy function at different odor 2 valences. (q<0.5 = Aversive, q>0.5 = Appetitive)

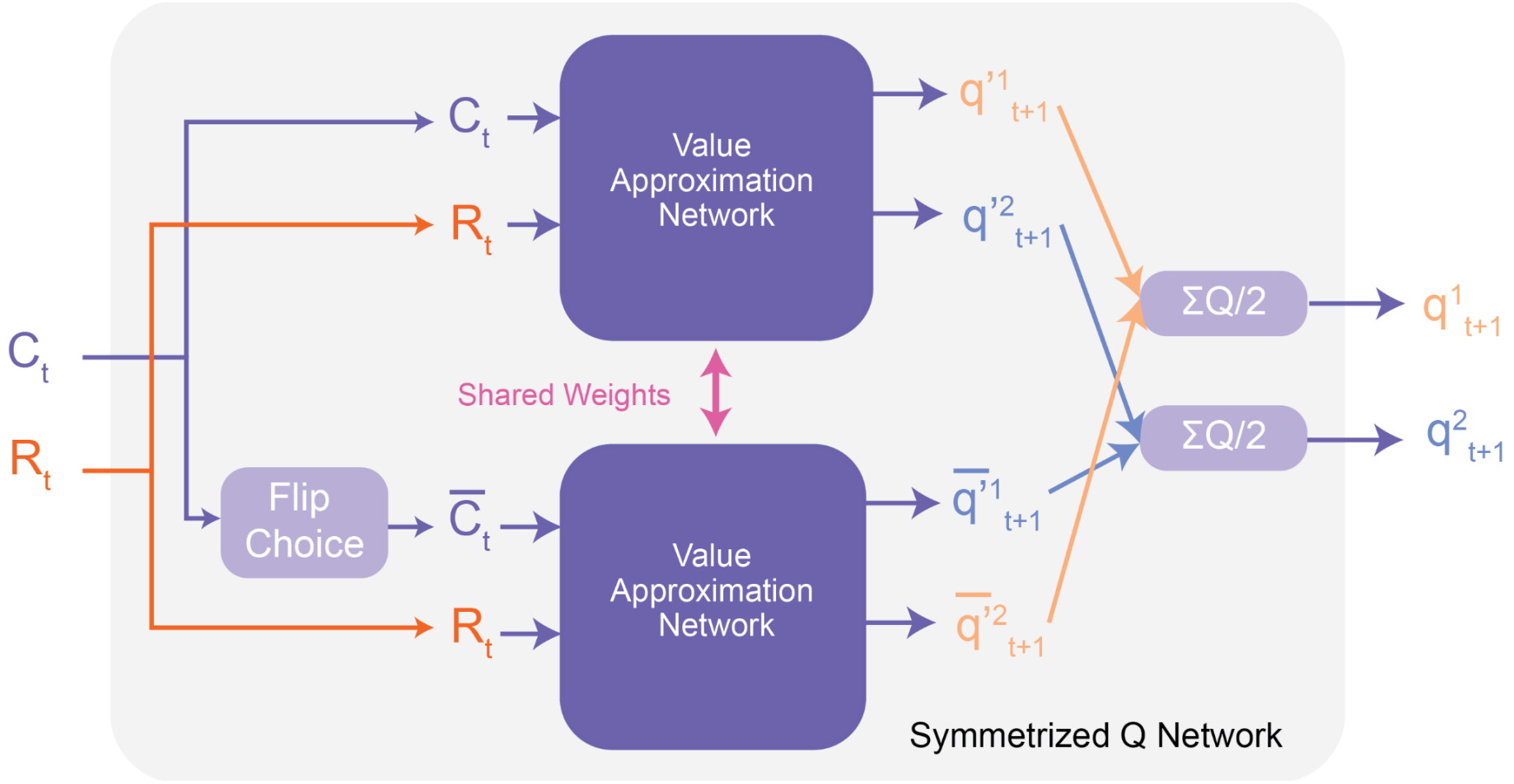


Figure 12. Schematic of q-Network Output Symmetrization (qNOS)

qNOS modifies the architecture of the network in other to ensure that the network's output is always symmetric, i.e., if the identities of the odors were flipped, the predicted acceptance probabilities would also be exactly flipped. We do this by creating a copy of the

choice input, flipping the odor identities, and passing it into a copy of the network. The outputs of the copies of the network (q'^1,q'^2, and $\bar{q}'^1$,$\bar{q}'^2$ respectively) are cross-averaged between the two copies of the network, i.e., q'1 is mixed with $\bar{q}'^2$ to get the final output q^1 and vice versa. While this effectively increases the number of independent neurons in the network, we retain the same number of parameters by coupling networks' activities through shared weights. This symmetrization ensures that learning both possible directions of odor choice-reward association happens simultaneously.

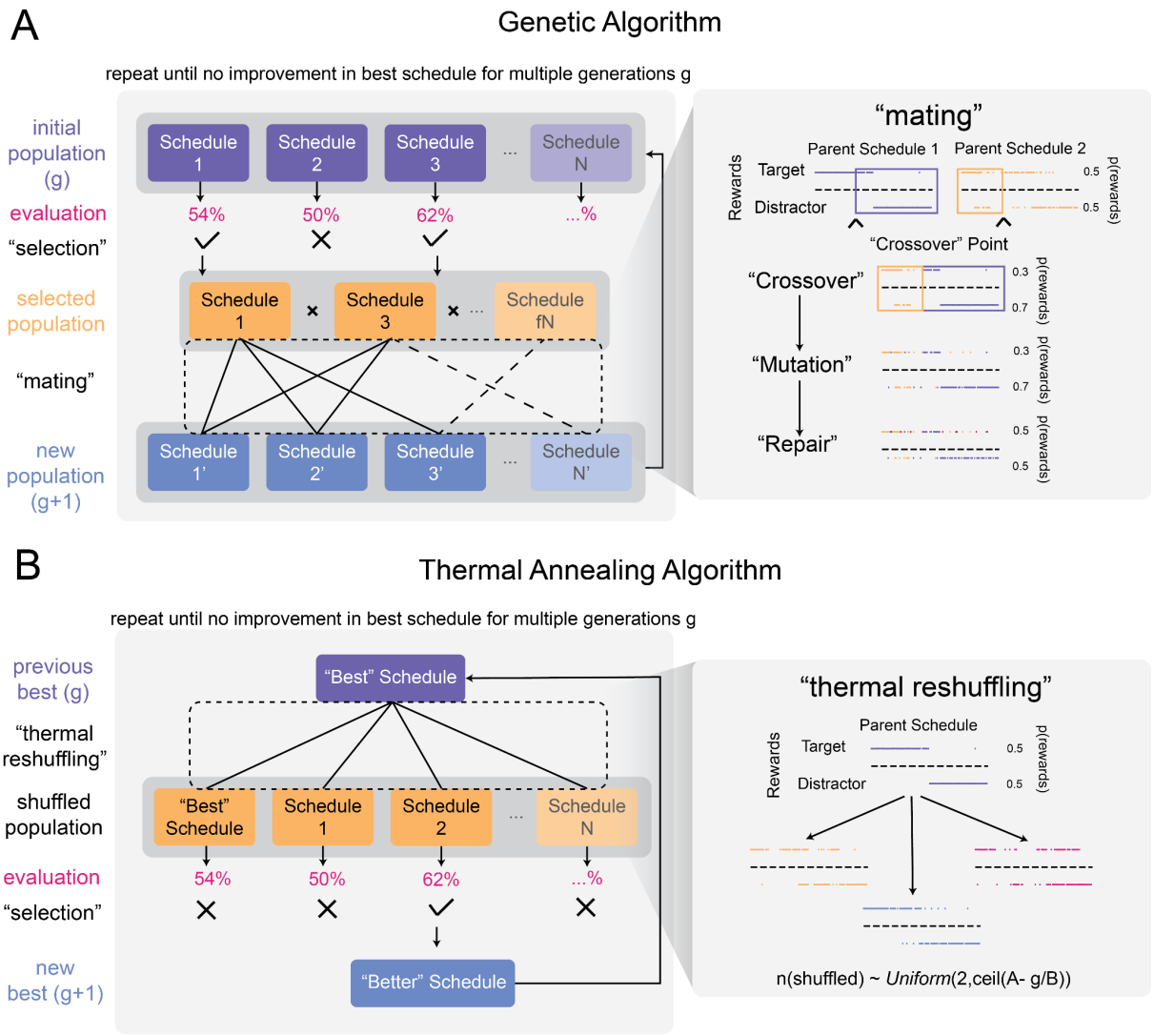


Figure 13. Open-loop choice engineering using stochastic optimization techniques

(A) Genetic Algorithm approach: A population of N = 100 reward schedules is initialized. In every generation, there are three steps: i) Evaluation: 1000 agents of the RL model being tested are simulated for each schedule; ii) Selection: The top f = 20% schedules with the maximum average bias (% choices where the target odor was chosen) are kept, and the rest are discarded. iii) Mating (see inset): From the surviving population, pairs of parents are randomly selected (with replacement) to generate the new population. New children are created from the parents by swapping blocks of rewards between parents defined by n+1 "crossover" points along the session where n ~ Poisson(0.25). New mutations are added by randomly shuffling 5% of the trials for each odor independently. Further, a repair process

randomly removes excess rewards or compensates for reward deficits to ensure the number of rewards remains constant.

(B) Thermal Annealing approach: A single schedule is taken, and its rewards are randomly shuffled to generate a population of 100 new schedules, keeping a copy of the original schedule in the population. The number of shuffles is randomly chosen uniformly between 2 and T where the temperature $T = \lceil A - g/B \rceil$ where A = 100, B = 2, and g is the generation number. We simulate 1000 agents of the model being tested on each schedule, and the ‘best’ option is kept, and the entire process is repeated.

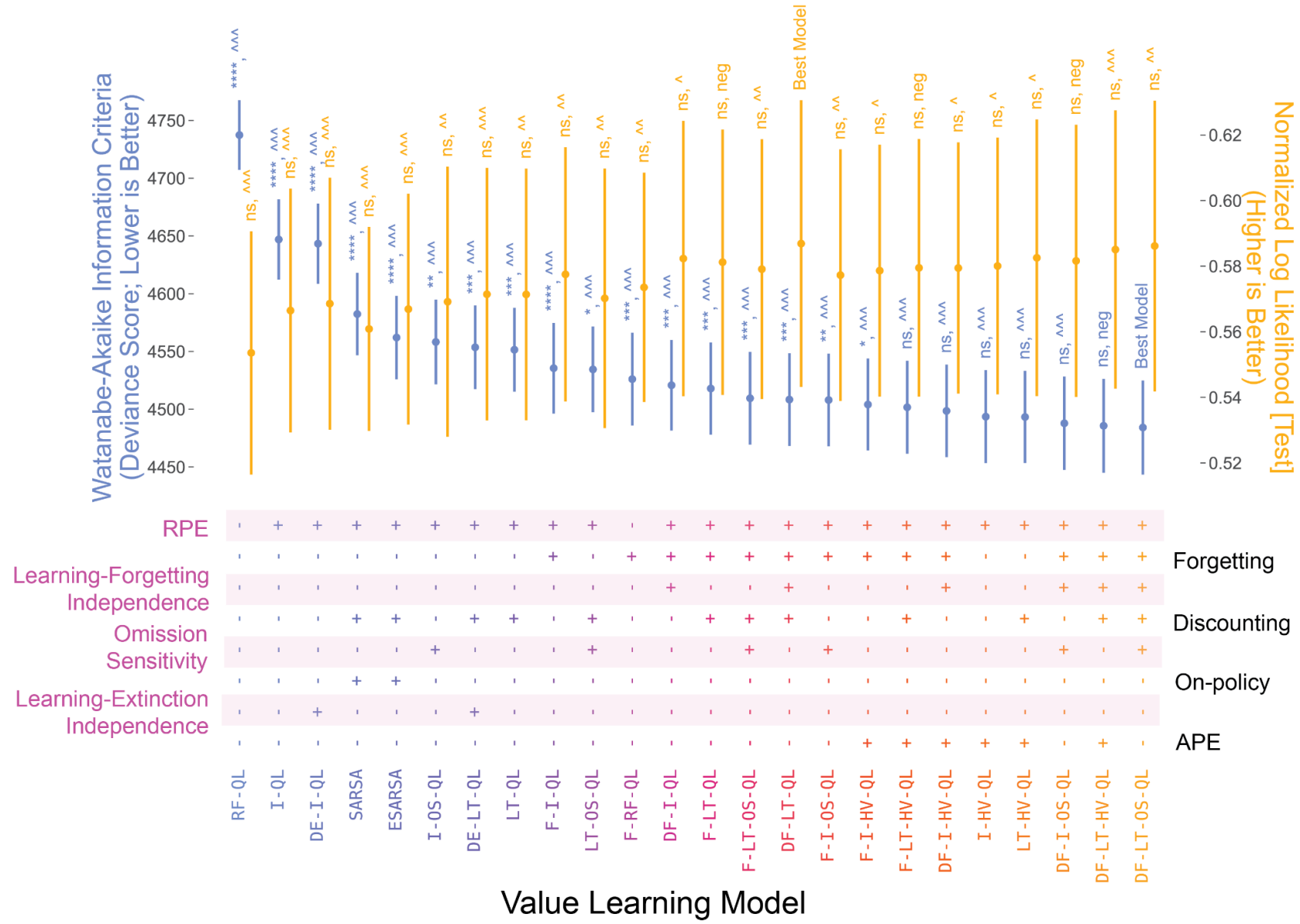

Watanabe-Akaike Information Criteria
(Deviance Score; Lower is Better)
4750
4700
4650
4600
4550
4500
4450
Normalized Log Likelihood [Test]
(Higher is Better)
0.62
0.60
0.58
0.56
0.54
0.52
Best Model
RPE
Learning-Forgetting Independence
Omission Sensitivity
Learning-Extinction Independence
Forgetting
Discounting
On-policy
APE
RF-QL
I-QL
DE-I-QL
SARSA
ESARSA
I-OS-QL
DE-LT-QL
LT-QL
F-I-QL
LT-OS-QL
F-RF-QL
DF-I-QL
F-LT-QL
F-LT-OS-QL
DF-LT-QL
F-I-OS-QL
F-I-HV-QL
F-LT-HV-QL
DF-I-HV-QL
I-HV-QL
LT-HV-QL
DF-I-OS-QL
DF-LT-HV-QL
DF-LT-OS-QL
Value Learning Model

Figure 14. Q-Learning Models of Rajagopalan (2022) "Fixed Block" dataset reveals that including learning-independent forgetting, perseverance, and temporal discounting in the value update improves the model's explanatory power.

The goodness of fit is estimated using the deviance-scaled Watanabe-Akaike Information Criterion (WAIC; blue), which is a bayesian posterior estimate of parameter count adjusted deviance. The difference of each model's WAIC relative to the best model is compared using a two-sided z-test (stars for statistical significance; see methods) and Cohen's d (carets for effect size). Predictive accuracy estimated using Normalized Likelihood [Test] (yellow) is compared relative to the best model using a bootstrap-corrected two-sided paired samples t-test (m=3 flies, n=1000 bootstraps; see methods) (stars for statistical significance) and paired Cohen's d (carets for effect size). The '+' and '-' symbols at the bottom signify which cognitive features (see Figure 10. and Table 7) are included in the model. Error bars show Standard Error for WAIC and Normalized Likelihood [Test]. See Table 13 for statistics, p-values, and effect sizes

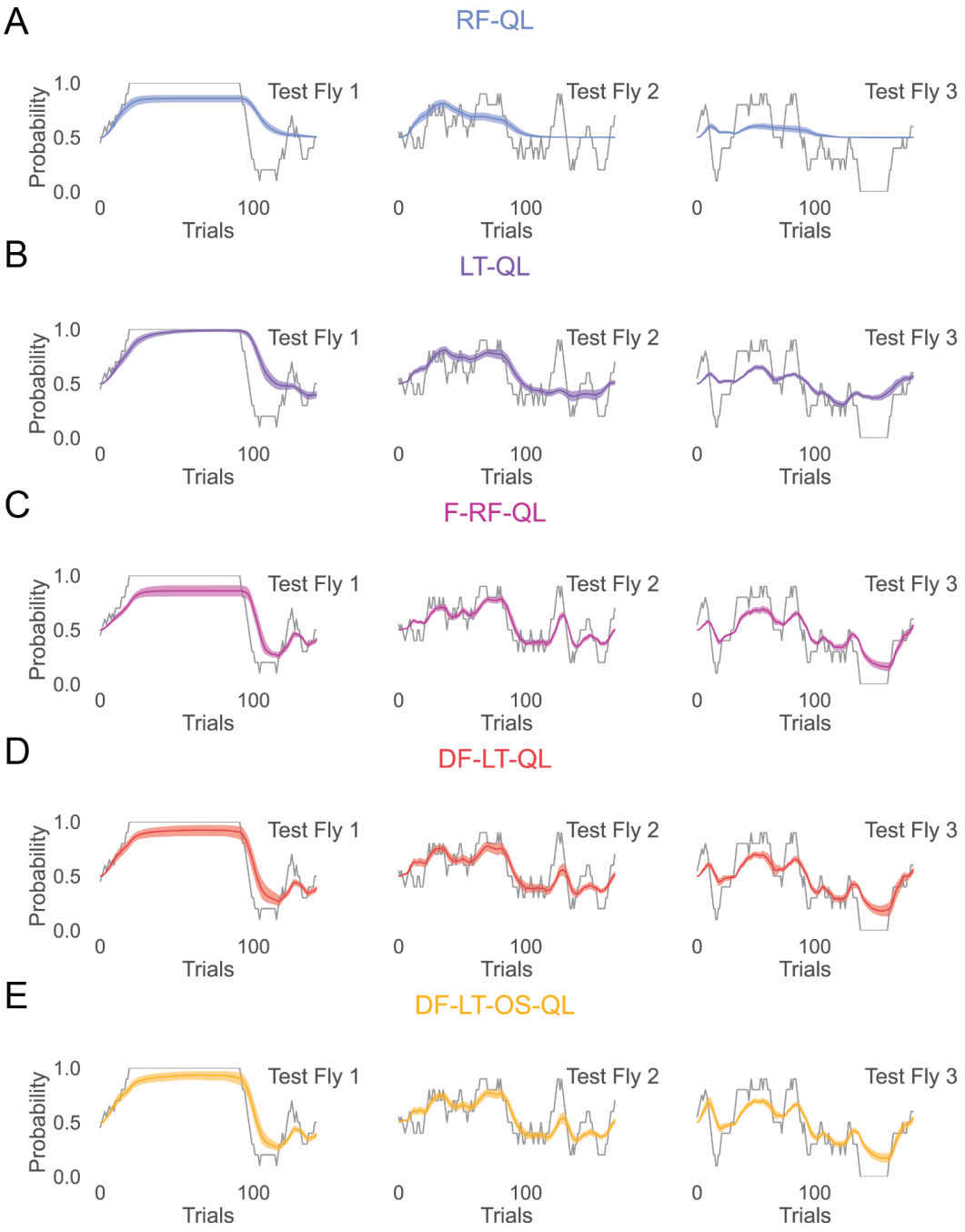


Figure 15. Predicted choice probabilities for different models show diminishing differences with more complex models.

(A–E) Smoothed predicted choice probabilities with 95% confidence interval estimated from 1000 simulations of 5 representative models across the spectrum of model fits for three flies that were not trained on the data overlaid on smoothed choice probabilities estimated from the data with a ten-trial window (see methods).

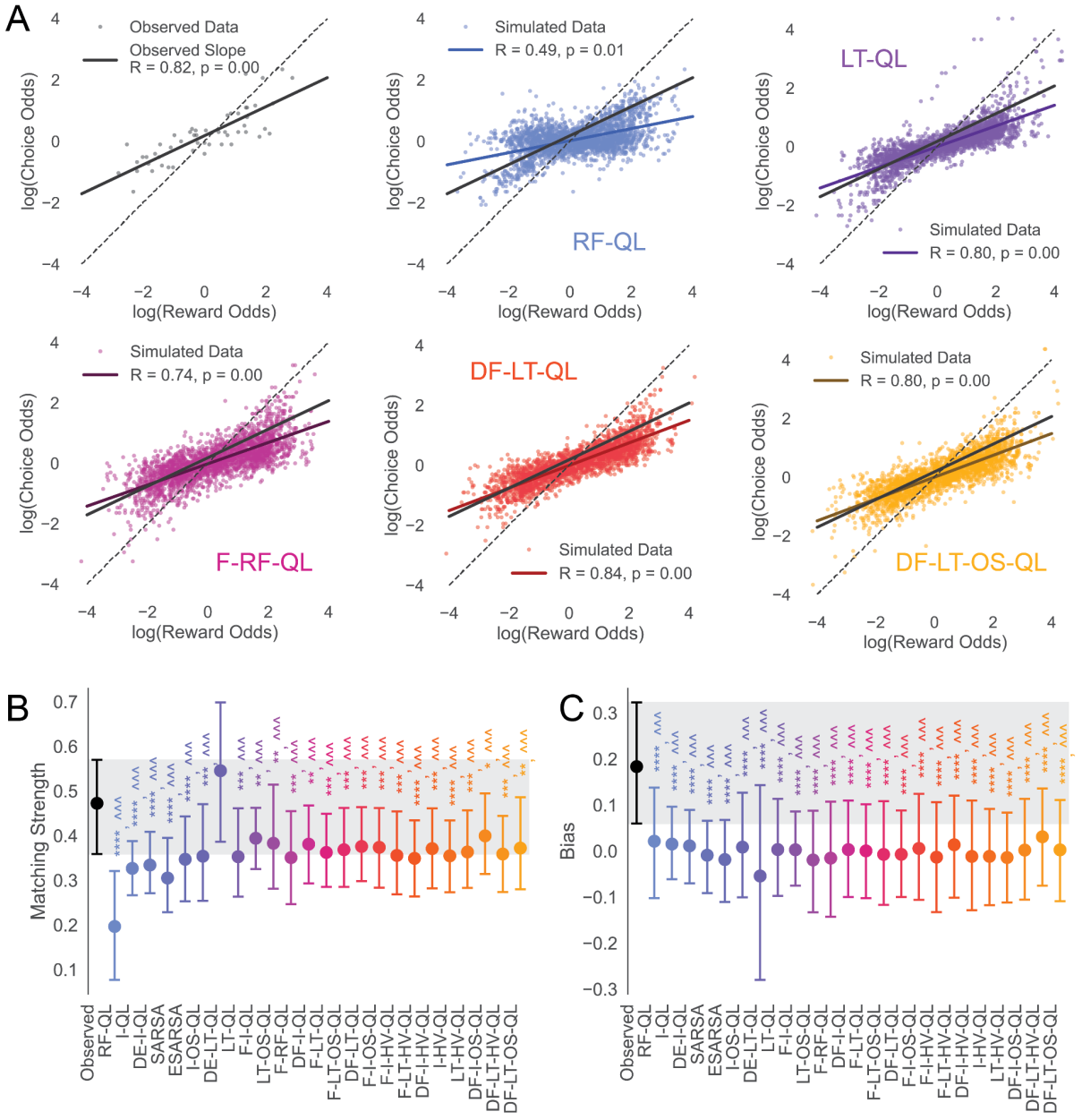


Figure 16. Q-Learning Models preserve the matching behavior observed in behavior.

(A) Generalized matching law observed as a linear function between log(choice odds) vs. log(reward odds) within each block of trials with static baiting probabilities for the experimental data and simulations of 50 repeats of the 18 experiments for the different models. Five representative models along the spectrum of the model fits are visualized. Linear fit, correlation coefficient R, and associated p-value are plotted and reported.

(B–C) Matching strength and bias (see methods) for the data and the model simulations with bootstrapped 95% CI. The grey band represents the 95% confidence interval for the observed data. Model behavior is compared to the experimental data using bootstrap-corrected Mann-Whitney test (m=18 flies, n=1000 simulations, 1000 random bootstraps; see methods) (stars for statistical significance) and Cliff's delta effect size (caret for effect size). See Table 15 for statistics, p-values, and effect sizes.

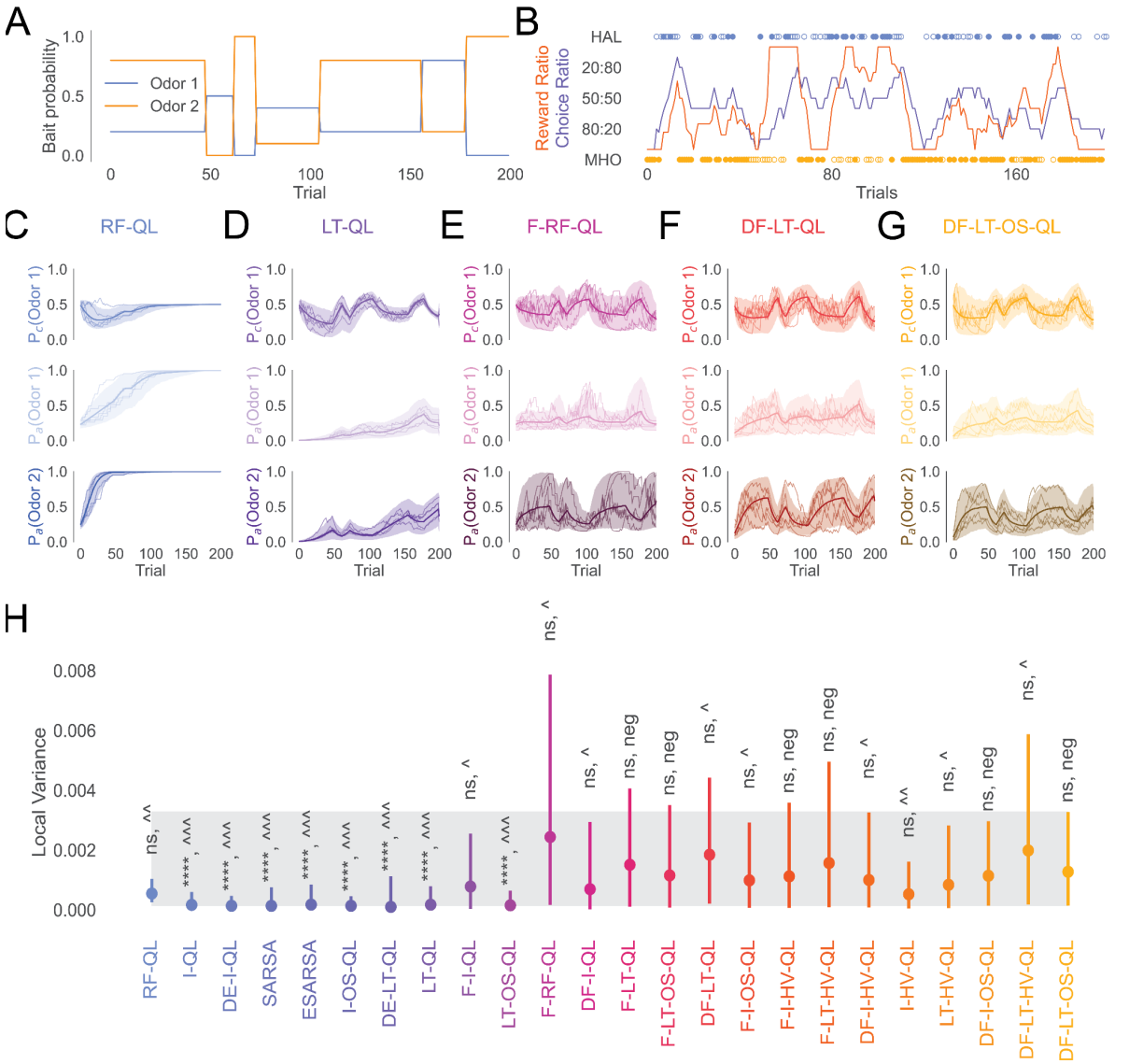


Figure 17. The dynamics of value underlying different models reveal differences in local variance.

(A) An example of a single random “Variable Block” experiment generated by simulating a simple Markov chain (see methods).

(B) Behavioral trajectory of a simulated fly using the best model (DF-LT-OS-QL) on the example Variable Block experiment that shows a matching between running reward ratio and choice ratio.

(C–G) Underlying preference dynamics for five representative models across the spectrum of model fits visualized using i) choice probability P_c(Odor 2) = the

probability of choosing odor 2; ii) acceptance probability P_a(Odor 1 or Odor 2) = the probability of choosing odor 1 or odor 2 (representative of its value) along with its 95% confidence interval (shaded area) calculated with 1000 independent trajectories. Five sample trajectories from the simulated data are shown overlaid on the data.

(H) Quantification of the local variance across a single session for different models. The shaded area represents the 95% confidence interval of the best model. Differences from the best model are quantified using bootstrap-corrected Mann-Whitney U test (m=18 flies, n=1000 simulations; unpaired data was sampled using 1000 bootstraps; see methods) (stars for statistical significance) and Cliff's delta effect size (carets for effect size). See Table 16 for p-values and effect sizes.

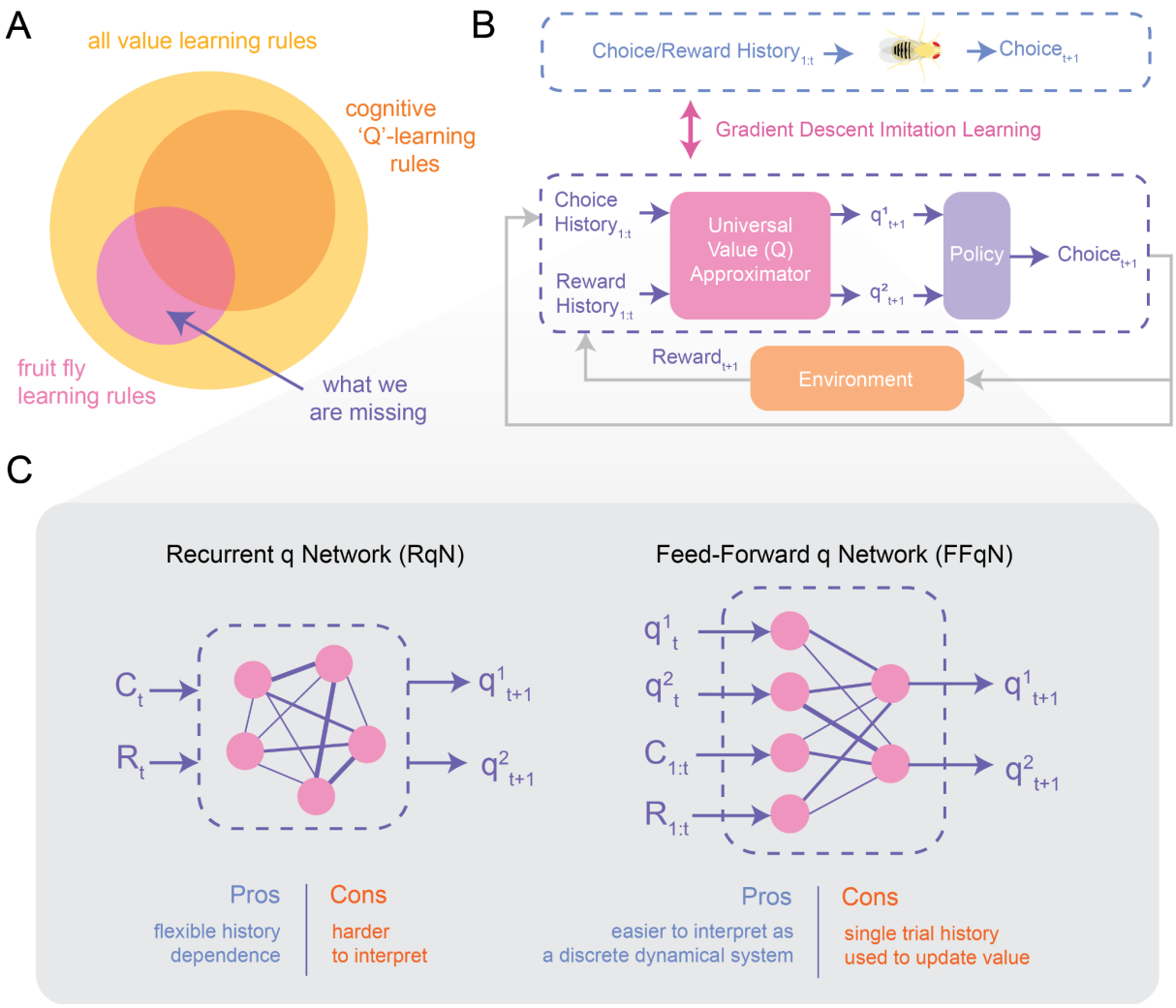


Figure 18. Neural networks can flexibly estimate the value learning rules via imitation learning.

(A) Venn Diagram of how we think the space of value learning rules are organized. The space of value learning is much bigger than the space we sample using our cognitive feature Q-learning models and is constrained by many assumptions. The actual space of learning rules that the fly uses may only partially overlap with our models; therefore, we need a way to sample the space with minimal assumptions.

(B) Our framework of value learning essentially needs a black box Universal Value (Q) Approximator (pink) that is capable of taking all of the histories and using it to predict the acceptance probabilities (q^1 and q^2; representative of odor value). These probabilities are then transformed by the Accept-Reject policy (see methods) and sampled to give the choices. The choices are then associated with rewards from the environment. In order to find what this black box does, we can use a Neural Network

to try and imitate the behavior observed constrained by the same value learning framework.

(C) We can use many different architectures for the neural network to approximate the behavior. However, the two leading types of artificial neural networks (ANNs) used by Machine-Learning (ML) researchers are Recurrent Neural Networks (RNNs) and Feedforward Neural Networks (FFNNs). We use these ANNs to create two different classes of value estimation networks. Recurrent q-Networks (RqNs) take in the entire sequence of past histories to predict the acceptance probabilities in the subsequent trial. Feedforward q-Networks take only the acceptance probabilities of the last trial and update them using the choice and reward from the current trial to update the acceptance probabilities for the subsequent trial.

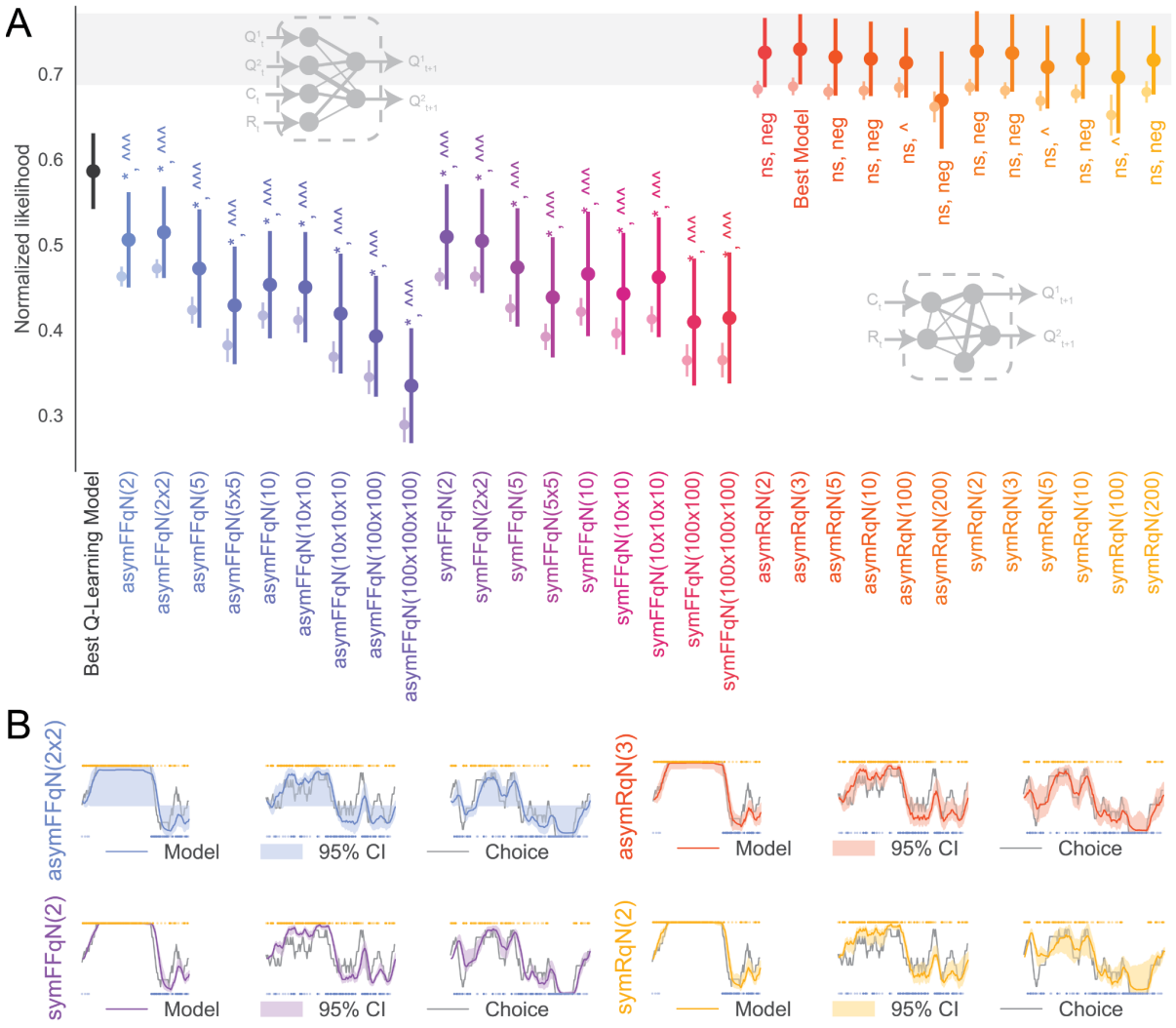


Figure 19. Small neural networks can explain fly behavior by estimating the dynamics of changing value of odors.

(A) Comparison of the goodness of fit and predictive power estimated using Normalized Likelihood on training data and testing data, respectively, for different architectures of neural networks trained to estimate value from data and predict the choices. Light and dark error bars represent the mean and standard error of training and test Normalized Likelihood, respectively. Test Normalized Likelihood of each of the models is compared to the best model (symRqN(3)) using a bootstrap-corrected two-sided paired samples t-test (stars for statistical significance) and bootstrap-corrected paired cohen's d effect size (carets for effect size) (m=3 flies, n=25 ensembles for bootstrap correction; see methods). See Table 17 for p-values

and effect sizes, including a comparison of training Normalized Likelihood using the same statistical measures.

(B) Smoothed predicted choice probabilities for 3 test flies with a 95% confidence interval estimated from 25 ensemble models for the best network architectures from each network class/variant overlaid on smoothed choice probabilities estimated from the data with a ten trial window (see methods).

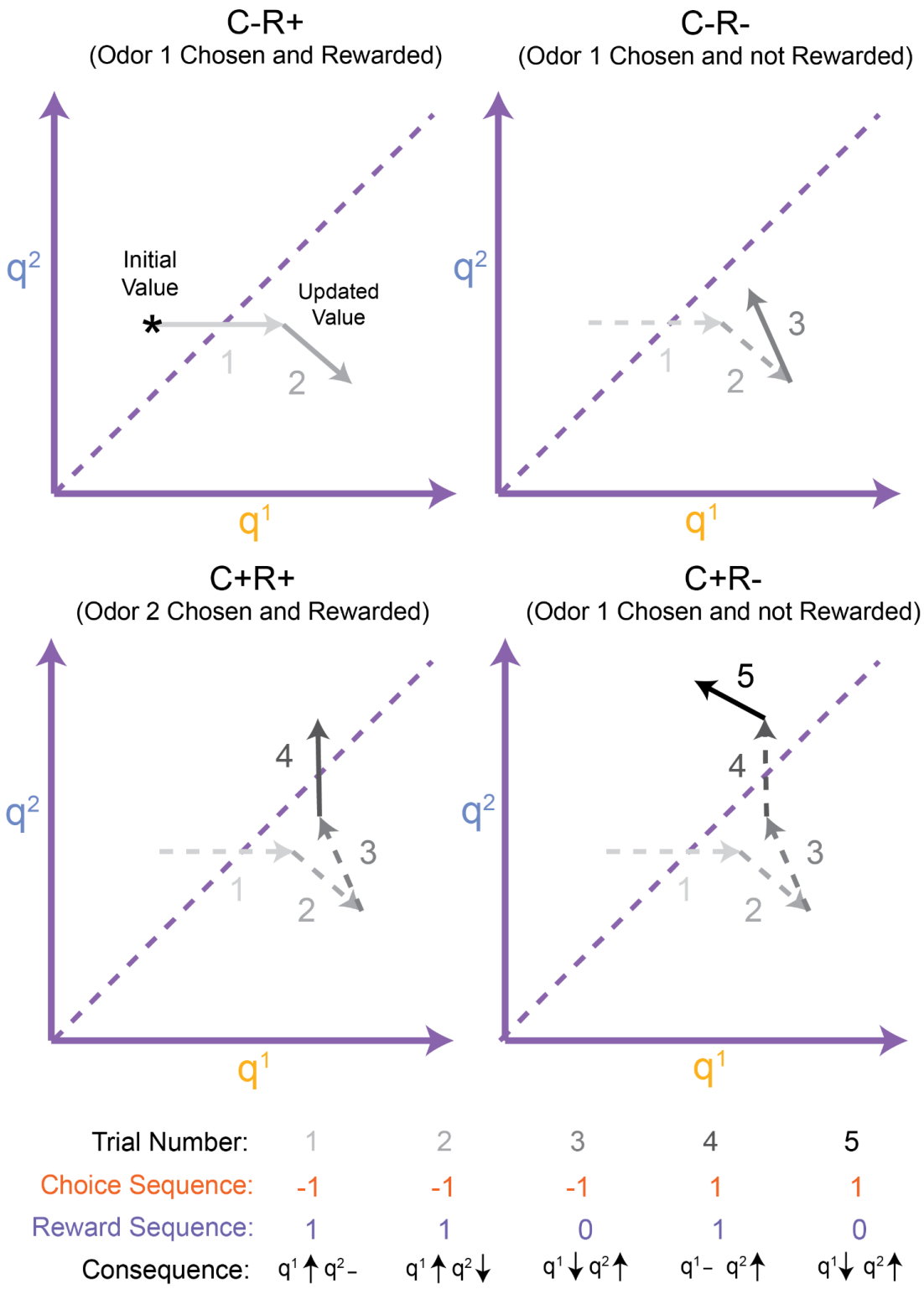


Figure 20. ***Understanding FFqNs as a conditional first-order discrete dynamical system.***

Consider an example sequence of choices and rewards starting at trial 1. The value of the two odors initially can be anywhere on the space of q^1 and q^2; say it is at the point marked by the asterisk. In the first trial, where odor 1 is chosen and rewarded,

the acceptance probability of odor 1 increases, and odor 2 remains the same (See arrow 1 in the C-R+ space). Since the change in probability only depends on the initial position and the (C, R) condition, the vector update is always uniquely defined for every point in the space. Similarly, in the subsequent trial, the update continues on the same condition C-R+, but this time the acceptance probability of odor 2 might reduce at this new point in the space. In the subsequent trial, the condition changes to C-R- where an independent vector field is defined, which leads to a decrease in the acceptance probability of odor 1 and an increase in odor 2. This vector continues over different (C, R) conditions over successive trials resulting in the overall behavior. However, any trajectory is fully defined by the four vector fields for the four conditions, the sequence of (C, R) conditions, and the initial conditions.

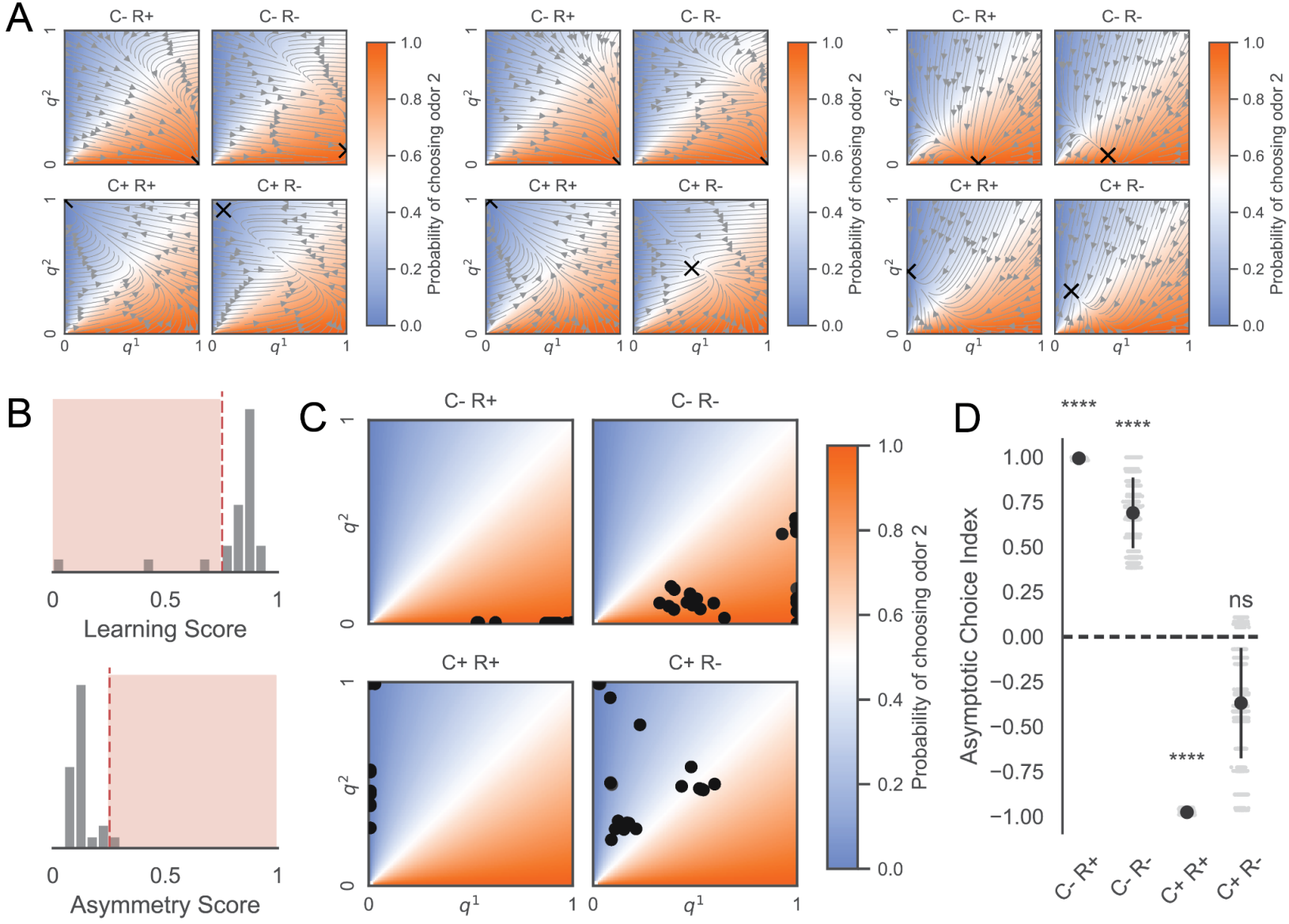

A
C- R+
C- R-
C+ R+
C+ R-
q^1
q^2
Probability of choosing odor 2
B
Learning Score
Asymmetry Score
C
D
Asymptotic Choice Index
****
ns
C- R+
C- R-
C+ R+
C+ R-

Figure 21. Dynamical systems analysis of an asymmetric FFqN reveals a system of unreliable attractors with weak perseverance.

(A) Vector fields for the acceptance probability update under the four Choice-Reward conditions represented as flows with the final choice probability represented as a heatmap with the simulated fixed points (from 100 independent initializations) marked with a cross. The estimated vector fields for three independent trained networks from the ensemble are shown.

(B) Histograms of learning and asymmetry scores for all the trained networks from the ensemble (for an explanation of scores, see methods).

(C) Position of all the fixed point attractors across the trained and filtered ensemble of asymmetric FFqNs marked on the space of acceptance probabilities with a black dot.

(D) Predicted preference of odors at the fixed point attractors of the different choice-reward conditions for all trained and filtered asymmetric FFqNs of the ensemble compared from zero using a two-sided bootstrap test (stars for significance; p=0.000 for all values other than C+R- where p=0.176).

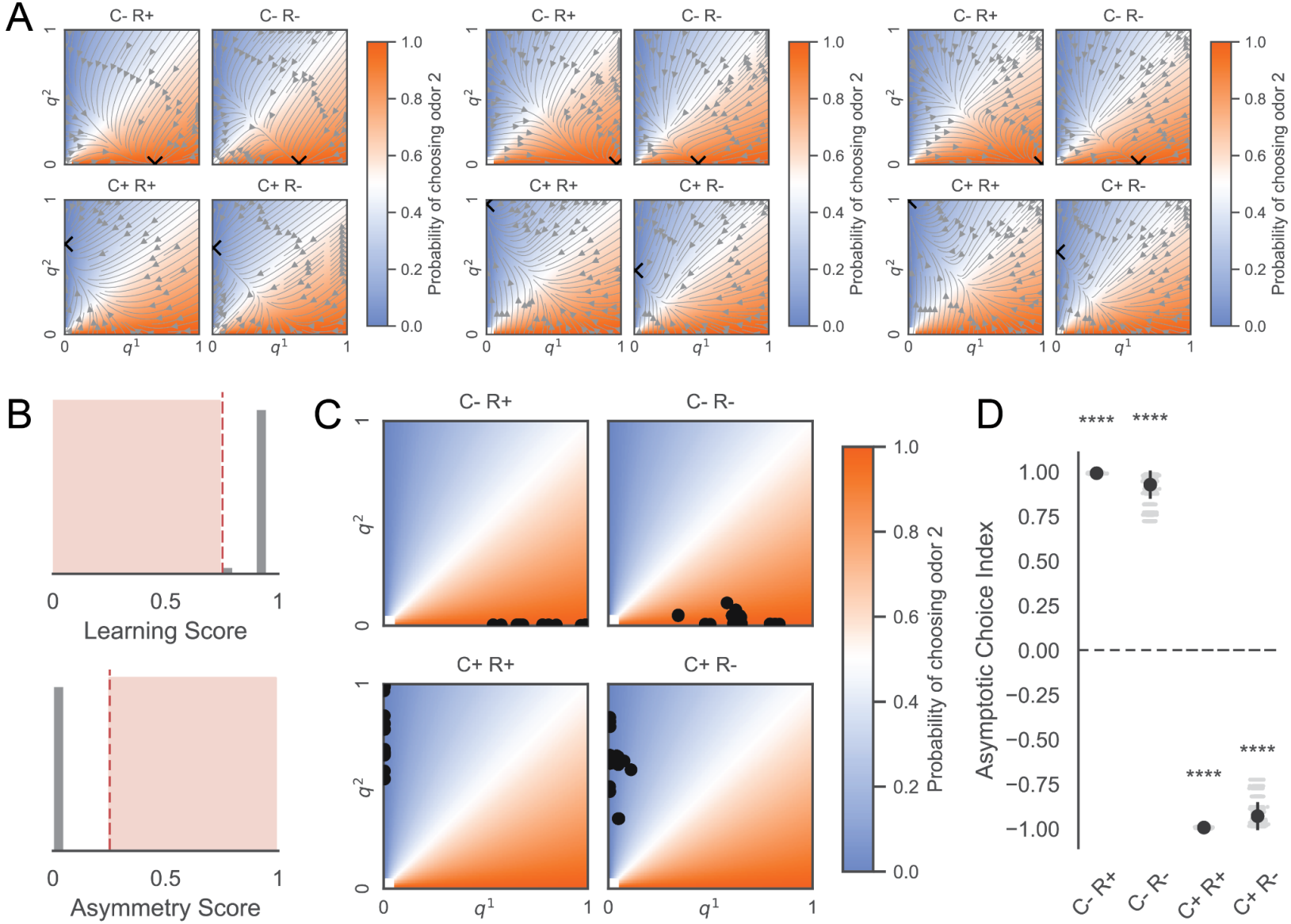

A
C- R+
C- R-
C+ R+
C+ R-
q^2
q^1
0
1
Probability of choosing odor 2
0.0
0.2
0.4
0.6
0.8
1.0
B
0
0.5
1
Learning Score
Asymmetry Score
C
D
****
Asymptotic Choice Index
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
C- R+
C- R-
C+ R+
C+ R-

Figure 22. Dynamical systems analysis of a symmetric FFqN reveals a system of reliable attractors with stronger perseverance.

(A) Vector fields for the acceptance probability update under the four Choice-Reward conditions represented as flows with the final choice probability represented as a heatmap with the simulated fixed points attractors (from 100 independent initializations) marked with a cross. The estimated vector fields for three independent trained networks from the ensembles are shown.

(B) Histograms of quantified learning and asymmetry scores for all the trained networks from the ensemble. Same as Figure 21..

(C) Position of all the fixed point attractors across the trained and filtered ensemble of symmetric FFqNs marked on the space of acceptance probabilities with a black dot.

(D) Predicted preference of odors at the fixed point attractors of the different choice-reward conditions for all trained and filtered symmetric FFqNs of the ensemble compared from zero using a two-sided bootstrap test (stars for significance; p=0.000 for all values).

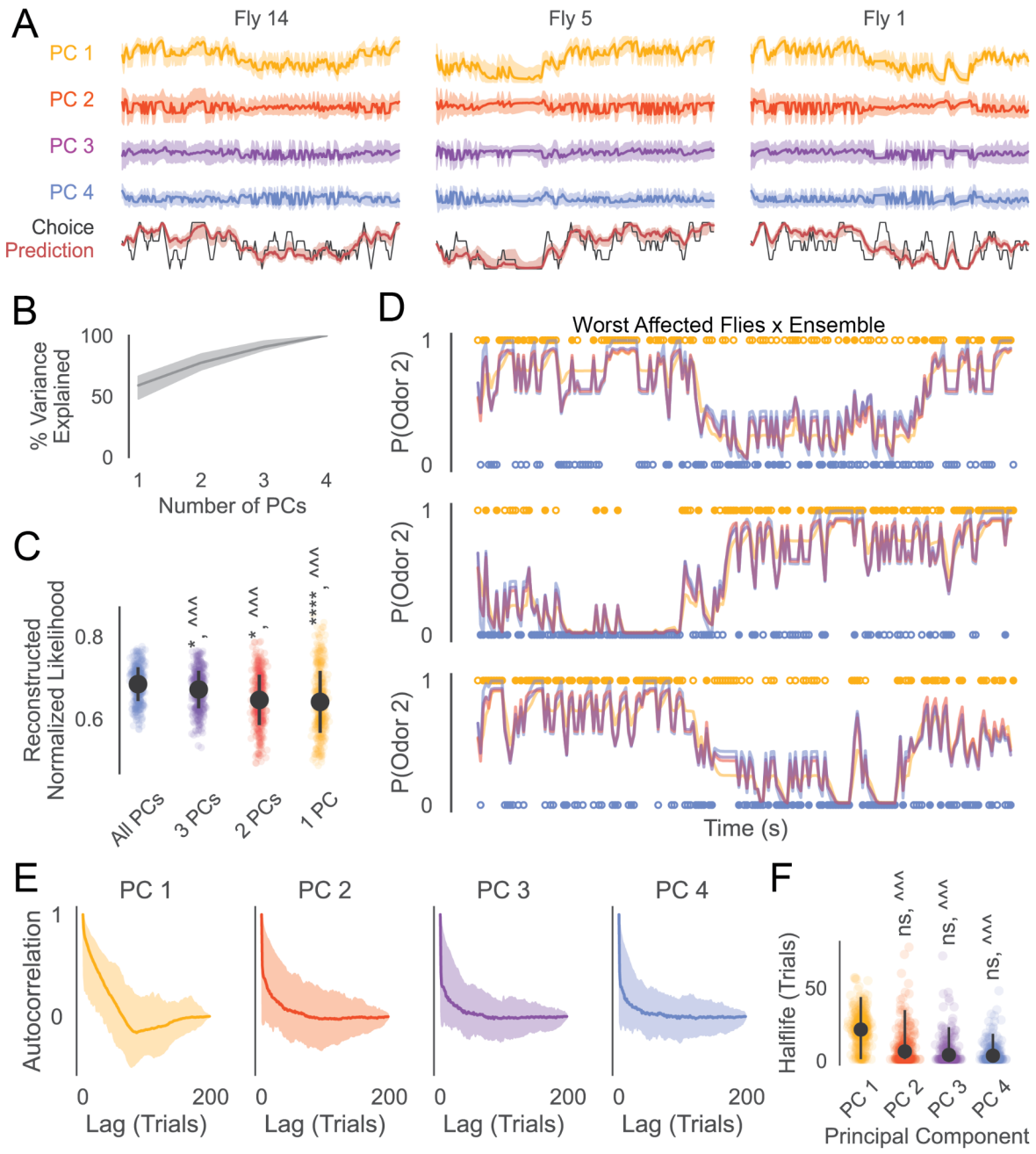


Figure 23. Dissecting the symmetric RqN reveals a possible separation of timescales that improves the performance of the RqN.

(A) Four Principal Components (PCs) of the hidden dynamics of the best symmetric RqN (symRqN(2) with four effective, hidden neurons) recovered using Principal Component Analysis (PCA) over the time axis. A 95% confidence interval (shaded area) was estimated from 25 trained networks from the ensemble shown alongside the smoothed choice probabilities from the data (black) and the prediction (red) (bottom; see methods). PCs were aligned by maximizing the correlation between

different trained networks to account for sign degeneracy in PCA methods. Data is shown for three flies from the training data (see subfigure D).

(B) Cumulative variance explained by the principal components with 95% confidence estimated over 25 trained networks from the ensemble.

(C) Contribution of each principle component was explored using a reconstruction Normalized Likelihood calculated by sequentially removing the PCs with the least contribution and then reconstructing the hidden dynamics fed to the decoder and policy to generate predictions for choice probabilities. Reconstructed normalized likelihood compared to log likelihood where all PCs are preserved using bootstrap-corrected two-sided Mann-Whitney-Wilcoxon test (stars for statistical significance; p=0.0397, 0.0131, 7.62e-6 respectively) and bootstrap-corrected matched-pairs rank biserial correlation effect size (carets for effect size; r = 0.837, 0.805, 0.665 respectively) (m=18 flies, n=25 ensembles for bootstrap correction; see methods)

(D) Effect of removing principal components on the smoothed predicted choice probabilities for the three most affected flies in the ensemble most affected by removing the last three principal components. Color of the lines represent the different removed components described in subfigure C

(E) Autocorrelation plot of the different PCs with a 95% confidence interval estimated with 18 flies across 25 ensembles.

(F) Halflife of the autocorrelation lag quantified for the different PCs. Black bars represent a 95% confidence interval calculated using 18 flies across 25 ensembles. Lag for the first PC is compared to the rest using two-sided Mann-Whitney-Wilcoxon test (stars for statistical significance; p=0.5011, 0.4044, 0.1682 respectively) and bootstrap-corrected matched-pairs rank biserial correlation effect size (carets for effect size; r = 0.861, 0.933, 0.932 respectively) (m=18 flies, n=25 ensembles for bootstrap correction; see methods)

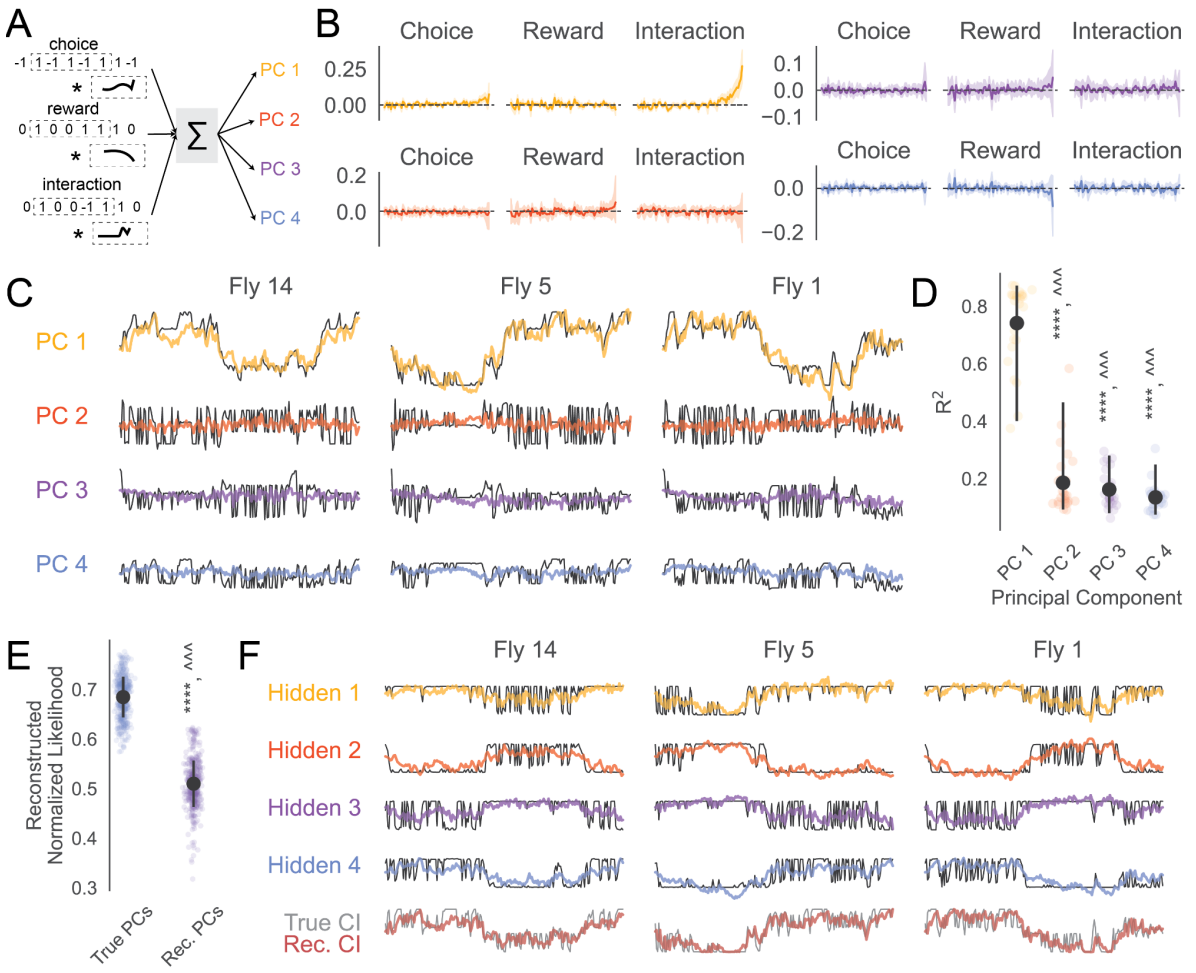


Figure 24. Kernel regression analysis to predict the principle components (PCs) of the hidden dynamics reveals the role of nonlinearity in the non-dominant PCs and suggests perseverance behavior.

(A) Kernel regression analysis applies convolutions with learned kernels on past time windows of choice, reward, and the interaction term (choice×reward). It sums them together to predict the future value of a PC of the hidden dynamics for the best symmetric RqN (symRqN(2)).

(B) Learnt kernels for the choice, reward, and interaction term to predict the values of different PCs with a 95% confidence interval estimated from 25 trained networks from the ensemble.

(C) Predicted (colored) and actual values (black) of the principle components predicted by a linear model for the same flies as Figure 23.

(D) Coefficient of determination (R2) for the linear fits for the four different PCs with the 95% confidence interval. The first PC is compared to the rest using two-sided Mann-Whitney-Wilcoxon test (stars for statistical significance; p=1.78e-7, 5.96e-8, 5.960e-8 respectively) and bootstrap-corrected matched-pairs rank biserial correlation effect size (carets for effect size; r = 0.987, 1.0, 1.0 respectively) (m=18 flies, n=25 ensembles for bootstrap correction; see methods)

(E) Predicted (colored) and actual (black) hidden dynamics for the four hidden neurons shown alongside the true (black) and predicted (red) trial-wise choice probabilities. Rec. stands for reconstruction.

(F) Reconstruction Normalized Likelihoods compared between a prediction with the actual PCs and the linear regression reconstructed PCs compared using a two-sided Mann-Whitney-Wilcoxon test (stars for statistical significance; p=7.629e-6 respectively) and bootstrap-corrected matched-pairs rank biserial correlation effect size (carets for effect size; r = 1.0 respectively) (m=18 flies, n=25 ensembles for bootstrap correction; see methods)

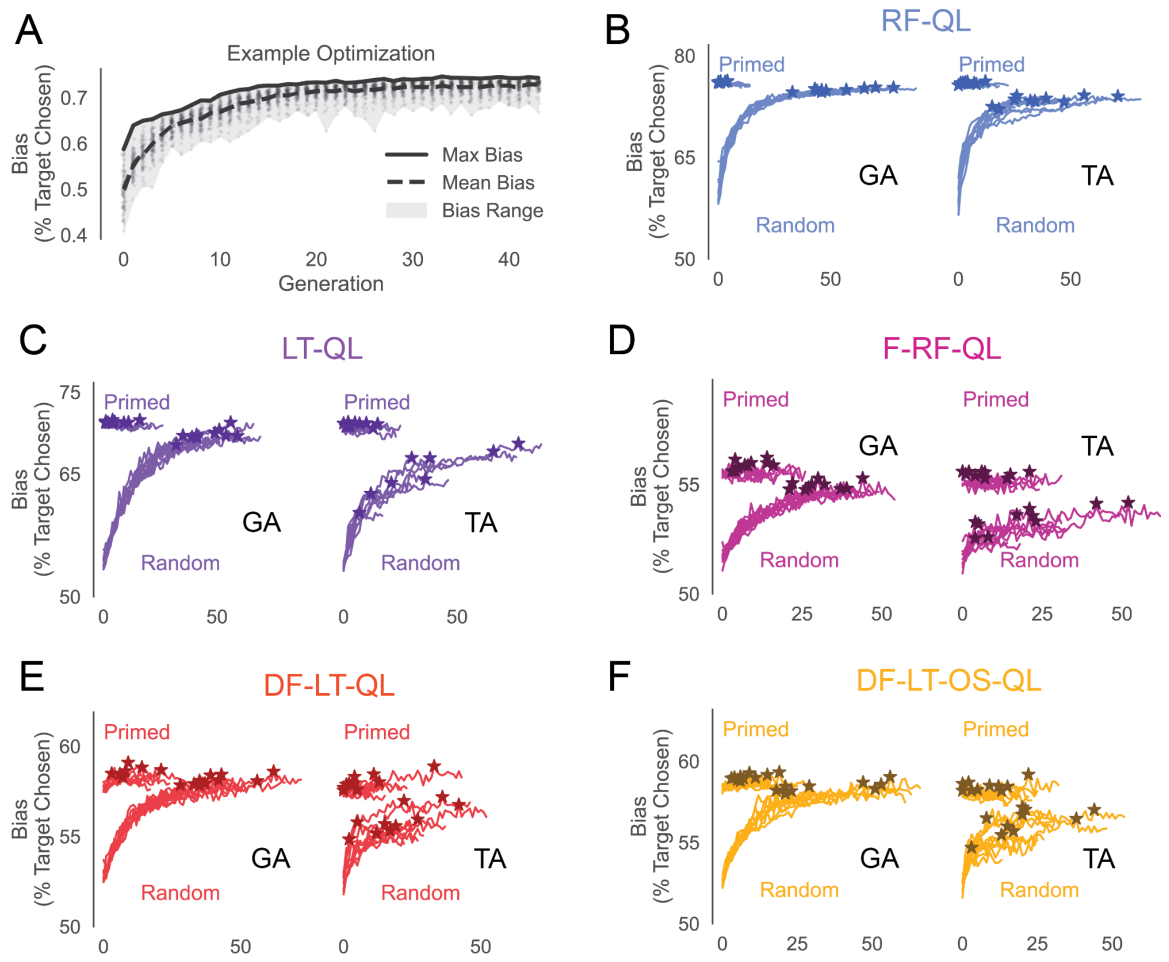


Figure 25. Optimization of choice engineering reward schedules.

(A) Example of a single stochastic optimization process for the DF-LT-OS-QL model using a genetic algorithm with the range of biases (shaded area), average bias (dotted line), and best bias (solid line) for the population of reward schedules.

(B–F) Traces of the best bias across multiple generations of stochastic optimization for five representative models. Replicates of different initializations (primed and random; see methods) and different optimization techniques are visualized. GA represents the Genetic Algorithm (left); TA represents Thermal Annealing (right)). Stars mark the final “best” discovered schedule for each initialization.

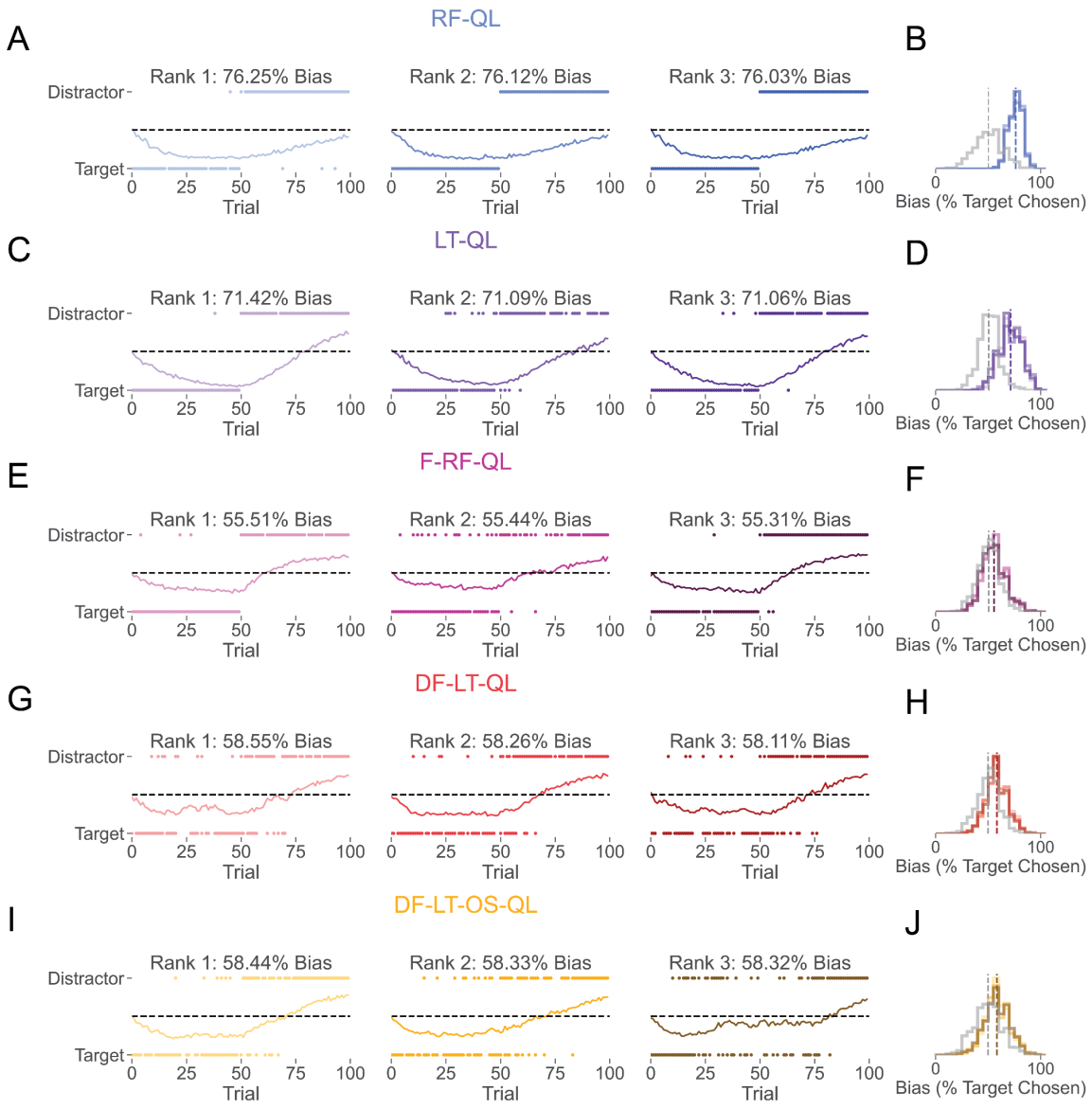


Figure 26. Choice Engineering provides candidate reward schedules for testing learning rules.

(A,C,E,G,I) Comparison of the top 3 maximally biasing reward schedules for five representative models. The top three schedules for each representative model are plotted. Dots represent the reward schedule i.e., rewards for the distractor and target odors for each trial. Absence of a dot represents the omission of reward on choice. Dotted lines represent no preference, and colored lines represent trial-wise bias for 1000 simulated agents.

(B,D,F,H,J) Distribution of overall biases over a 100 trial session for 1000 simulated agents for the top 3 maximally biasing schedules for five representative models compared to a schedule when equally spaced rewards are given identically on both odors (in gray).

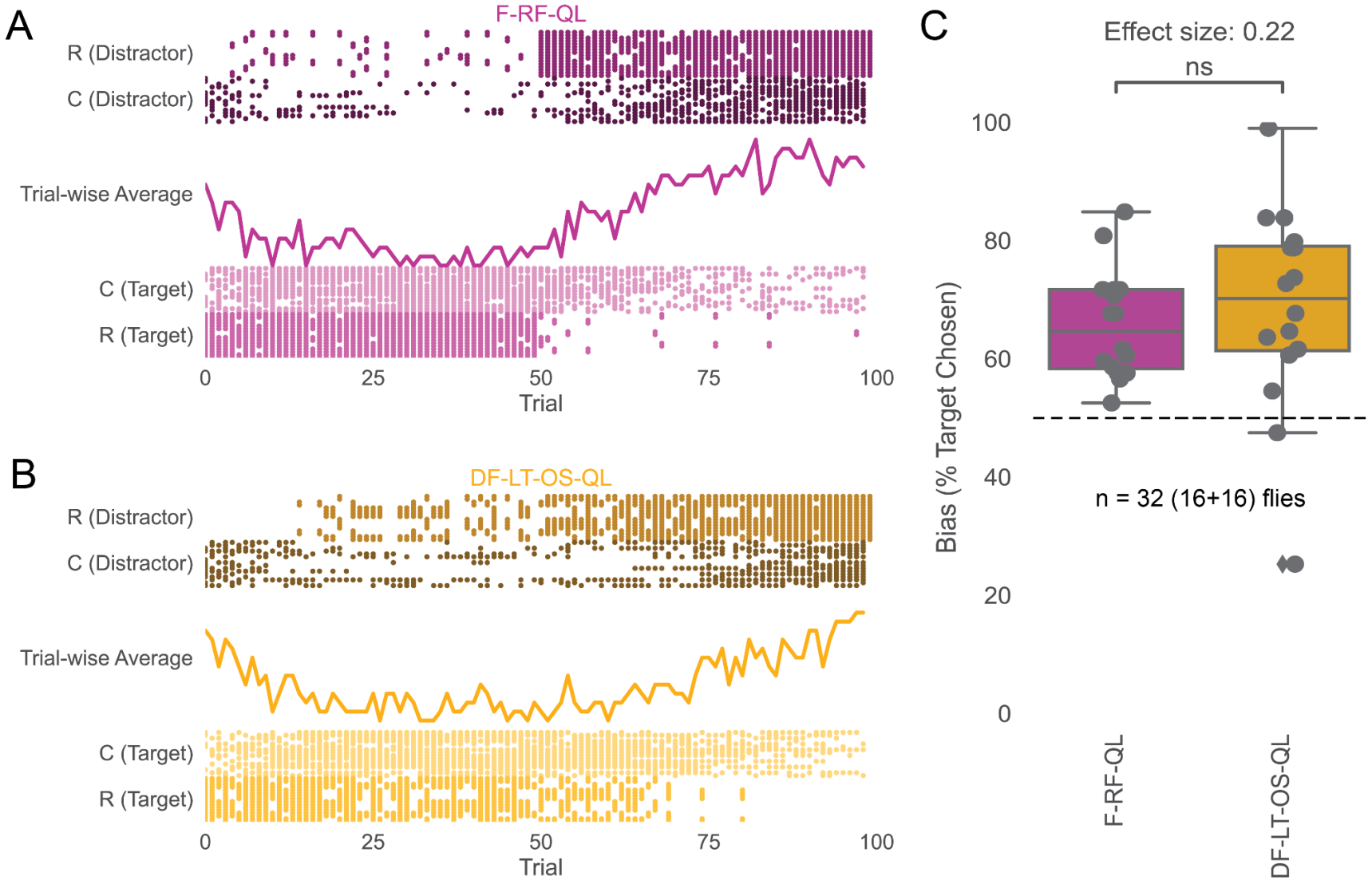

A
F-RF-QL
R (Distractor)
C (Distractor)
Trial-wise Average
C (Target)
R (Target)
0
25
50
75
100
Trial
B
DF-LT-OS-QL
R (Distractor)
C (Distractor)
Trial-wise Average
C (Target)
R (Target)
0
25
50
75
100
Trial
C
Effect size: 0.22
ns
100
80
60
40
20
0
Bias (% Target Chosen)
n = 32 (16+16) flies
F-RF-QL
DF-LT-OS-QL

Figure 27. Optimal schedules predicted by DF-LT-OS-QL models only show a weak increase in bias than those predicted by F-RL-QL models.

(A–B) Reward (dark) and choice (light) sequences for 16 flies tested in a single fly Y-maze for reward schedules predicted by both F-RL-QL and DF-LT-OS-QL models. For each set of schedules, eight flies were run with MCH at the target and eight with OCT as the target. The trial-wise average preference is visualized in the middle.

(C) Bias of each fly (% target chosen over a 100 trial session) for the two sets of schedules are found to be statistically non-significant but show a slight increase in bias ($p = 0.2994$; Mann Whitney U Test; $\delta = 0.2188$; Cliff's delta effect size) for 16 flies for each set of schedules.

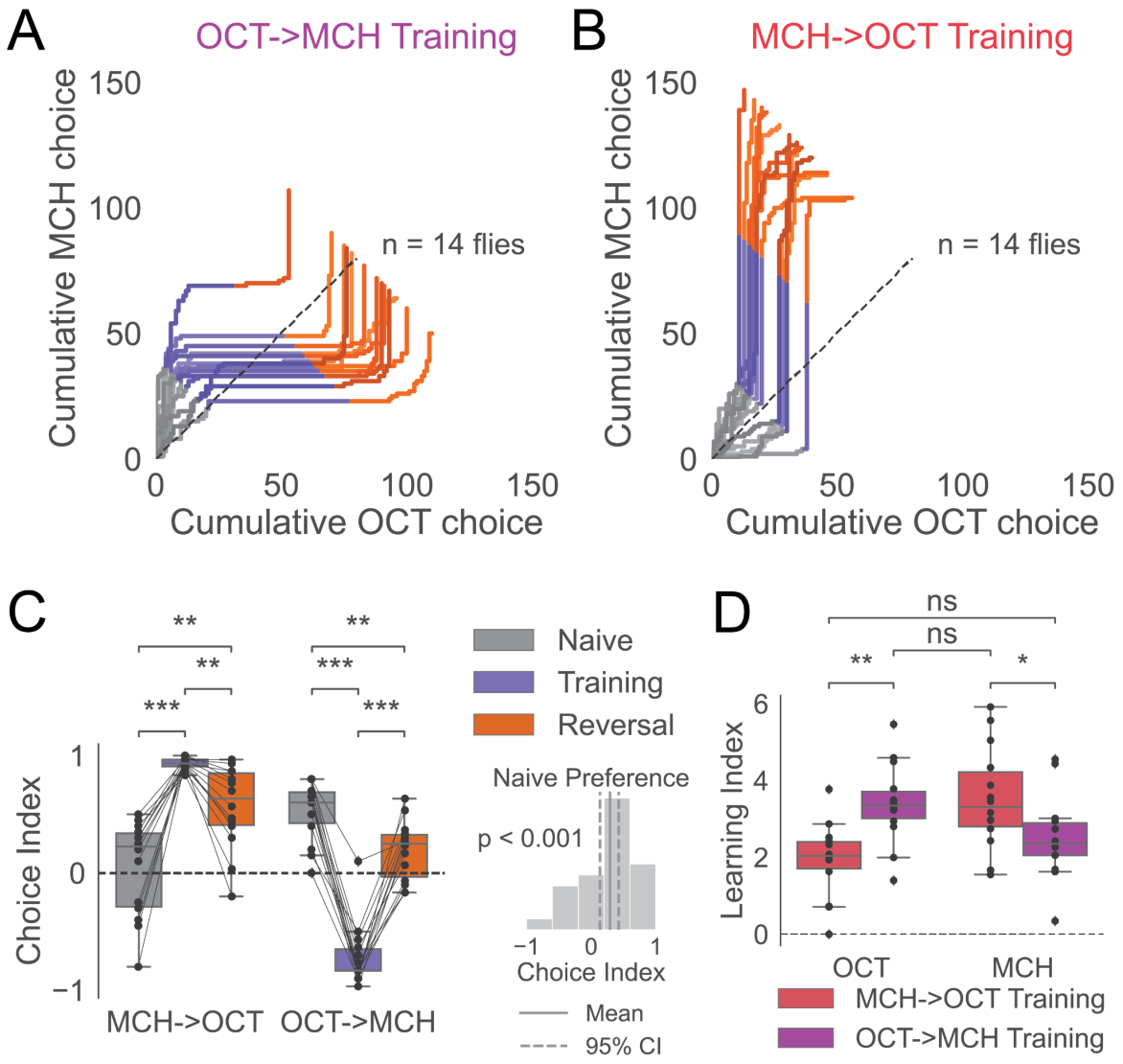


Figure 28. Strong Learning and asymmetric preference are observed for 24 hr starved flies in a high-throughput behavioral rig.

(A–B) Cumulative choices of OCT and MCH over time in experiments with 40 unrewarded trials (Naive) followed by 60 trials of pairing OCT(A)/MCH(B) with reward (Training), followed by 60 trials of pairing the opposite odor, i.e.,, MCH(A)/OCT(B) with certain reward (Reversal). The slope of the curve gives instantaneous preference.

(C) Choice index (+ve is MCH preference, -ve is OCT preference; see methods) quantified across the three phases (left). Values are compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see

Table 19). Overall naive preference with a 95% confidence interval (right) compared to zero with a two-sided one-sample t-test (stars for statistical significance; p = 9.99e-04).

(D) Learning index (+ve is reward association for the paired odor; see methods) quantified for the two odors under the training and reversal condition compared using two-sided Mann-Whitney U test (stars for statistical significance; see Table 20)

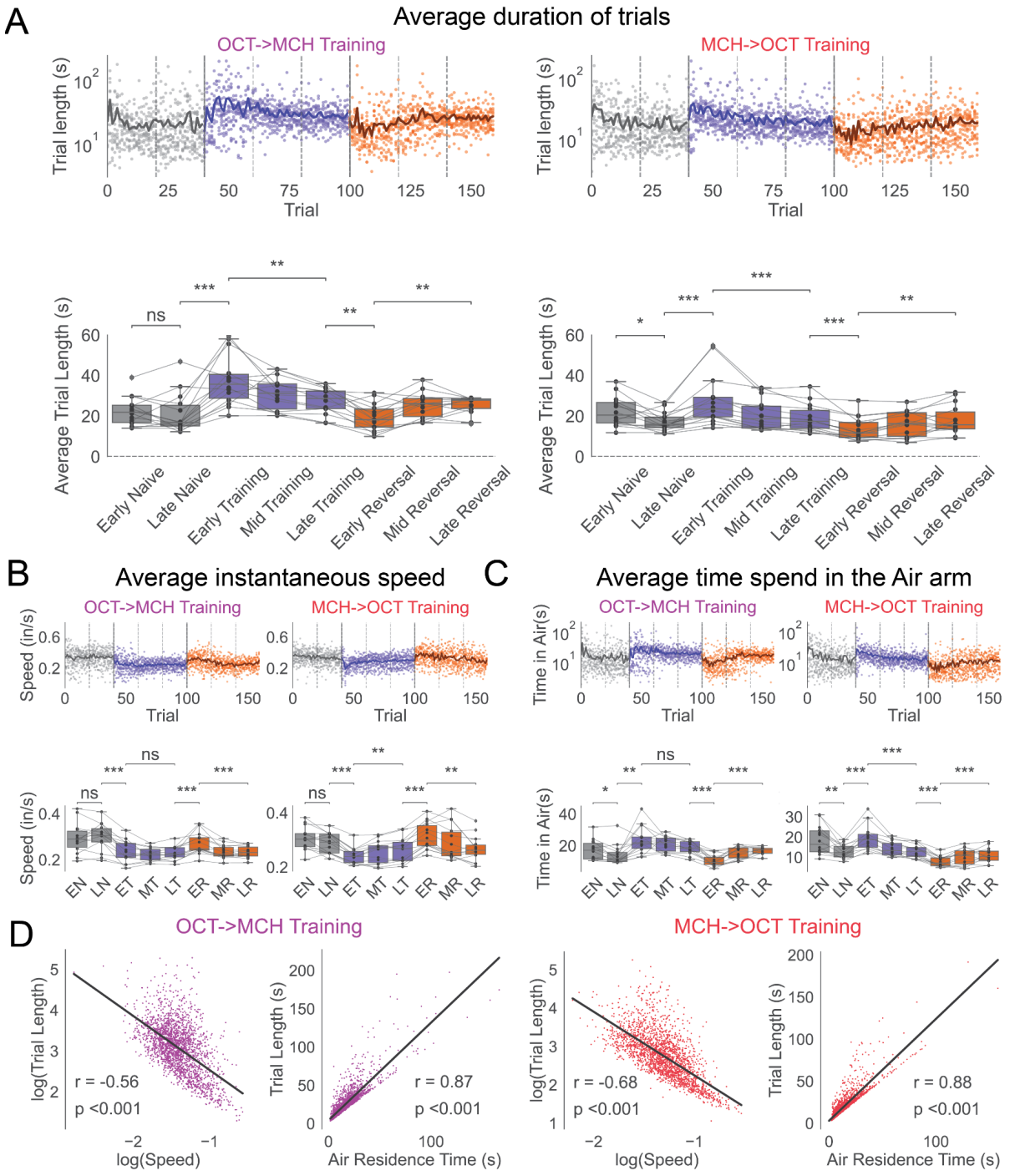


Figure 29. Slower choices on reward learning are explained by slower movement and residence in the last rewarded arm.

(A) Duration of each trial across all 14 experimental flies for the two learning experiments. Plotted along with the mean for each experiment (black) (top). Binned average trial times for flies across 20-trial subdivisions of the experimental phases (bottom) compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see Table 18)

(B) Average instantaneous speed in a trial across all 14 experimental flies for each of the two learning experiments. Plotted along with the mean for each experiment (black) (top). Binned average speeds for flies across 20-trial subdivisions of the experimental phases (bottom) compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see Table 18)

(C) Average time spent in the air arm for a trial across all 14 experimental flies for the two learning experiments. Plotted along with the mean for each experiment (black) (top). Binned average time spent in the air arm for flies across 20-trial subdivisions of the experimental phases (bottom) compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see Table 18)

(D) Log of trial duration compared to a log of average instantaneous speed (left) and trial duration compared to time spent in the air arm (right) for every trial across 14 flies for each experiment using Pearson's correlation (p=1.51e-182, 0.0, 2.37e-299, 0.0 from left to right).

EN: Early Naive Phase; LN: Late Naive Phase; ET: Early Training Phase; MT: Mid Training Phase; LT: Late Training Phase; ER: Early Reversal Phase; MR: Mid Reversal Phase; LR: Late Reversal Phase.

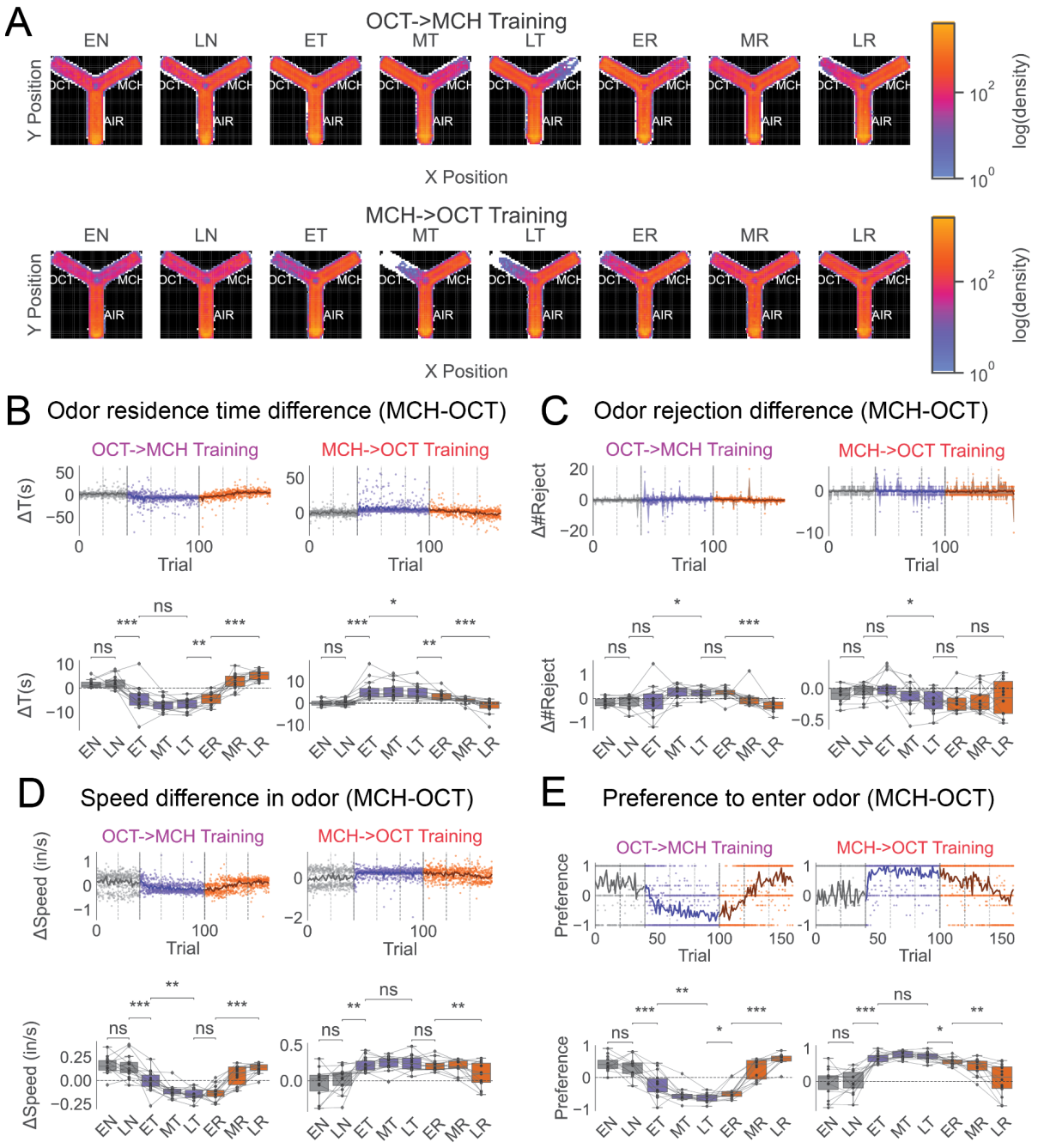


Figure 30. Change in odor preference as a function of reward history is a consequence of multiple kinematic factors.

(A) Residence of flies in the Y-arena (oriented to odor identity; left arm is OCT, the right arm is MCH; bottom is air) across each subdivision of the experimental phases in both experiments.

(B) Difference in the time spent in the MCH arm and the OCT arm in every trial across all 14 experimental flies for each of the two learning experiments. Plotted

along with the mean for each experiment (black) (top). The binned difference for flies across 20 trial subdivisions of the experimental phases (bottom) compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see Table 18)

(C) Difference in the number of times MCH is rejected and OCT is rejected in every trial across all 14 experimental flies for each of the two learning experiments. Plotted along with the mean for each experiment (black) (top). The binned difference for flies across 20 trial subdivisions of the experimental phases (bottom) compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see Table 18)

(D) Difference in the average instantaneous speed in the MCH arm and the OCT arm in every trial across all 14 experimental flies for each of the two learning experiments, along with the mean for each experiment (black) (top). The binned difference for flies across 20 trial subdivisions of the experimental phases (bottom) compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see Table 18)

(E) Difference in the fraction of times the MCH arm is entered, and the OCT arm is entered in every trial across all 14 experimental flies for each of the two learning experiments. Plotted along with the mean for each experiment (black) (top). The binned difference for flies across 20 trial subdivisions of the experimental phases (bottom) compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see Table 18)

EN: Early Naive Phase; LN: Late Naive Phase; ET: Early Training Phase; MT: Mid Training Phase; LT: Late Training Phase; ER: Early Reversal Phase; MR: Mid Reversal Phase; LR: Late Reversal Phase.

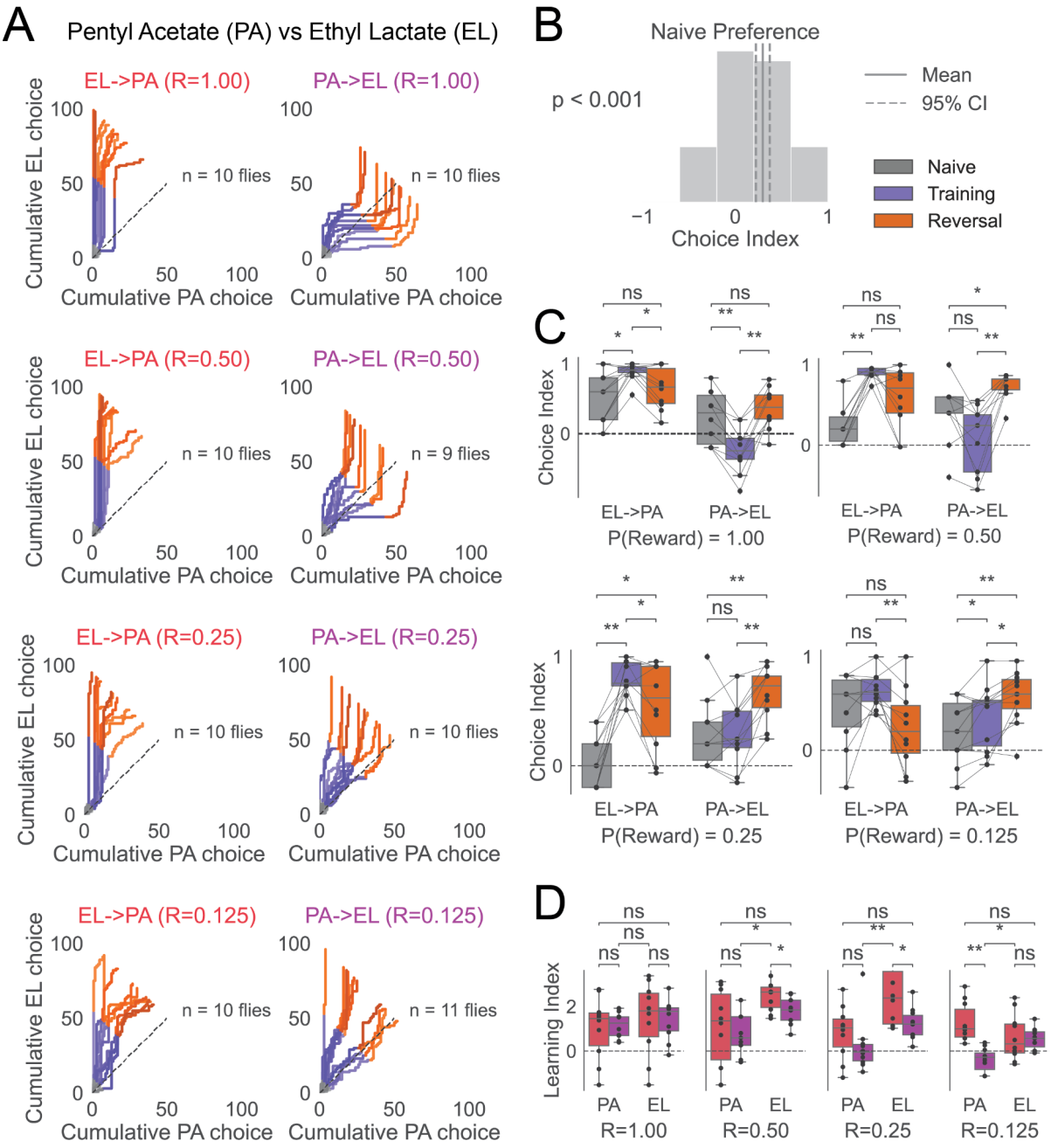


Figure 31. PA vs. EL choices show asymmetric, non-specific learning, especially at low reward probabilities, and a naive preference toward EL in 24 hr-starved flies.

(A) Cumulative choices of PA and EL over time in experiments with 10 unrewarded trials (Naive) followed by 45 trials of pairing EL/PA with reward (Training), followed by 45 trials of pairing the opposite odor, i.e., PA/EL with reward (Reversal). The reward pairing is varied to have different reward P(R) probabilities =0.125, 0.25, 0.5, and 1. The slope of the curve gives instantaneous preference.

(B) Overall naive preference with 95% confidence interval compared to zero with a two-sided one-sample t-test ($p = 0.00$) quantified using choice index (+ve is EL preference, -ve is PA preference; see methods).

(C) Choice index quantified across the three experimental phases and reward probabilities compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see Table 19).

(D) Learning index (+ve is reward association for the paired odor; see methods) quantified for the two odors under the training and reversal condition for different reward probabilities compared using two-sided Mann-Whitney U test (stars for statistical significance; see Table 20)

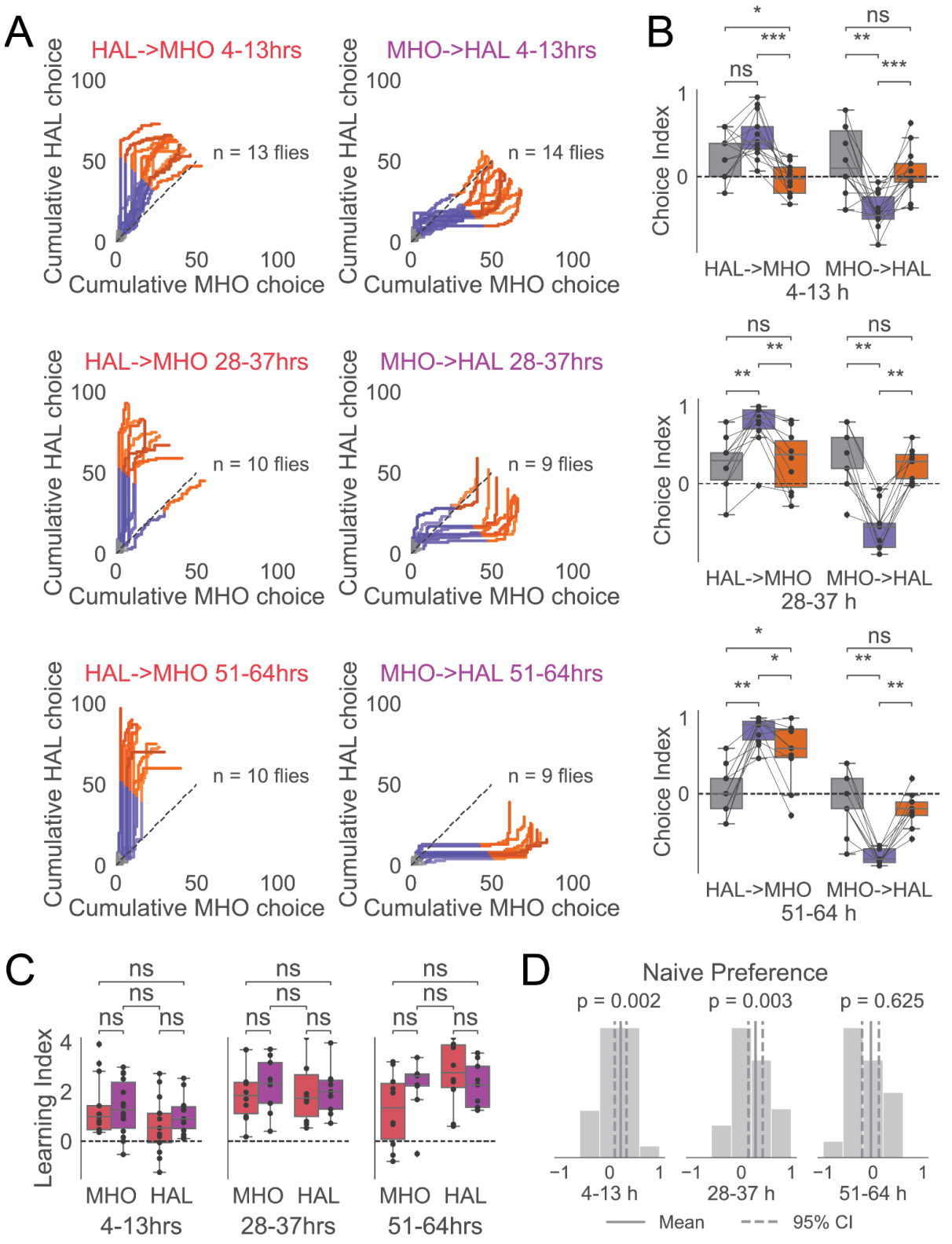


Figure 32. MHO vs. HAL choices show symmetric learning across starvation states with starvation-sensitive naive preference.

(A) Cumulative choices of HAL and MHO over time in experiments with 10 unrewarded trials (Naive) followed by 45 trials of pairing HAL/MHO with certain reward (Training), followed by 45 trials of pairing the opposite odor, i.e., MHO/HAL with reward (Reversal). The experiments were performed at different levels of starvation. The slope of the curve gives instantaneous preference.

(B) Choice index (+ve is HAL preference, -ve is MHO preference; see methods) quantified across the three experimental phases and levels of starvation. Values are compared using two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance; see Table 19).

(D) Learning index (+ve is reward association for the paired odor; see methods) quantified for the two odors under the training and reversal condition for different probabilities compared using two-sided Mann-Whitney U test (stars for statistical significance; see Table 20)

(B) Overall naive preference with 95% confidence interval across starvation levels compared to zero with a two-sided one-sample t-test using choice index (+ve is EL preference, -ve is PA preference; see methods).

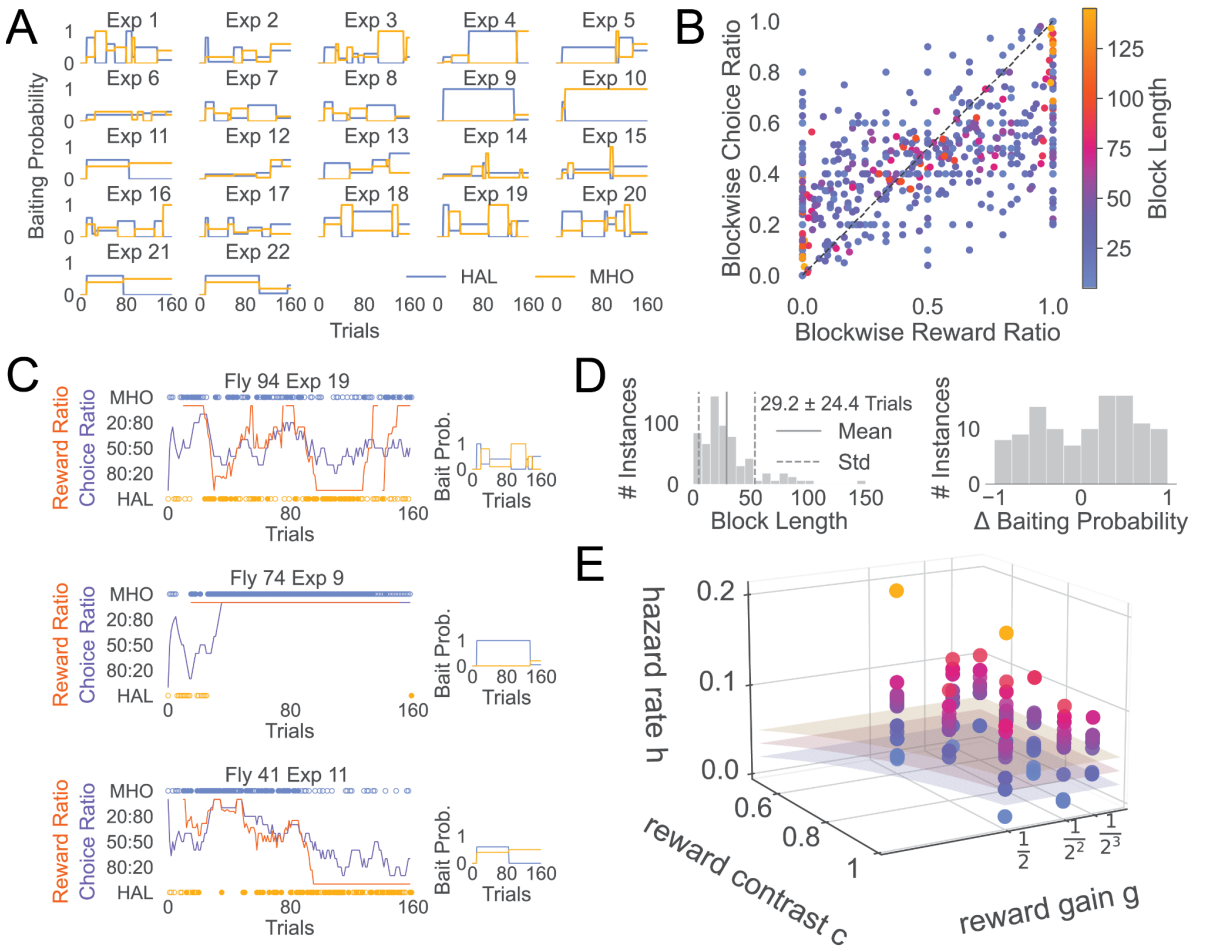


Figure 33. Mohanta (2022) "Variable Block" dataset shows probability matching across a broad sample of the space of dynamic baited-reward 2-alternative forced choice tasks.

(A) Set of baiting probabilities for 22 "forward" experiments that were run on three different flies each, along with three flies on "reciprocal" experiments where the odor identities were flipped.

(B) Blockwise reward ratios and Blockwise choice ratios are compared for the dataset colored by the length of the block in which the ratio is calculated.

(C) Three random example choice trajectories (left) from the data with the associated baiting probabilities (right). Orange and Blue dots in the reward schedule represent choosing Odor 1 and 2, respectively. Filled and empty dots represent the rewarded choice and unrewarded choices, respectively. The red and purple lines represent the reward and choice ratios calculated for 10 trials before the current trial (including the current trial).

(D) Histograms of the length of the blocks along with the 95% confidence interval (left) and the change in baiting probabilities of an odor between two successive blocks.

(E) Points of the task space (see methods) defined by reward gain, reward contrast, and estimated hazard rates sampled in the experiments with hazard rate estimated by looking at the reciprocal of the length of blocks observed in an experiment under each condition.

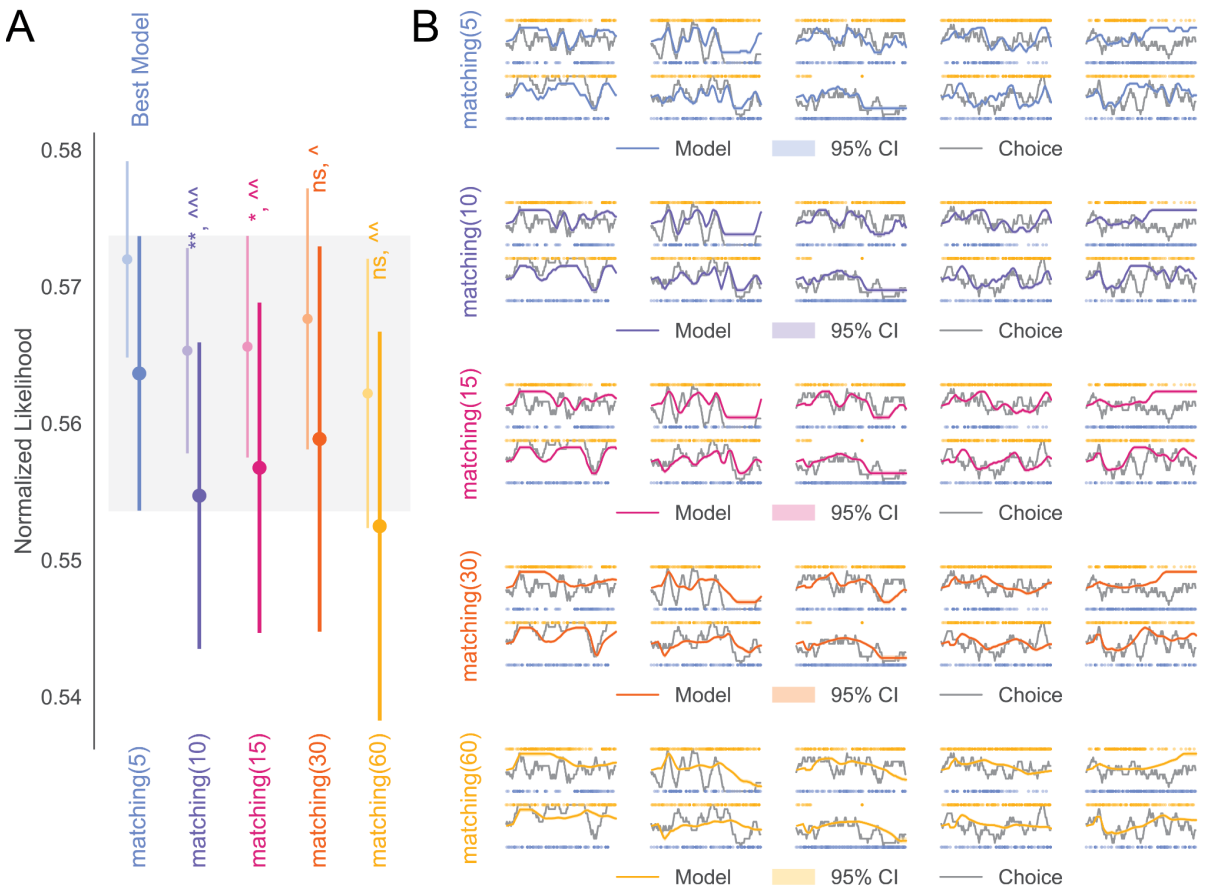


Figure 34. Constrained matching law models can predict future behavior with small integration windows.

(A) Comparison of the goodness of fit and predictive power estimated using Normalized Likelihood on training data and testing data, respectively, for different fits of the constrained matched law models (see methods) with different sizes of integration windows. Light and dark error bars represent the mean and standard error of training and test Normalized Likelihood fitted using 1000 bootstrapped samples on the training dataset. Test Normalized Likelihood of each of the models is compared to the best model (matching(5) - constrained matching law model with an integration time window of 5 trials) using a bootstrap-corrected two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance) and bootstrap-corrected matched-pairs rank biserial correlation effect size (carets for effect size) (m=44 flies, n=1000 bootstraps; see methods). See Table 24 for p-values and effect sizes, including a comparison of training Normalized Likelihood using the same statistical measures.

(B) Smoothed predicted choice probabilities for ten random test flies with 95% confidence interval estimated from 1000 bootstrap fits overlaid on smoothed choice probabilities estimated from the data with a 10 trial window (see methods) for the five matching models.

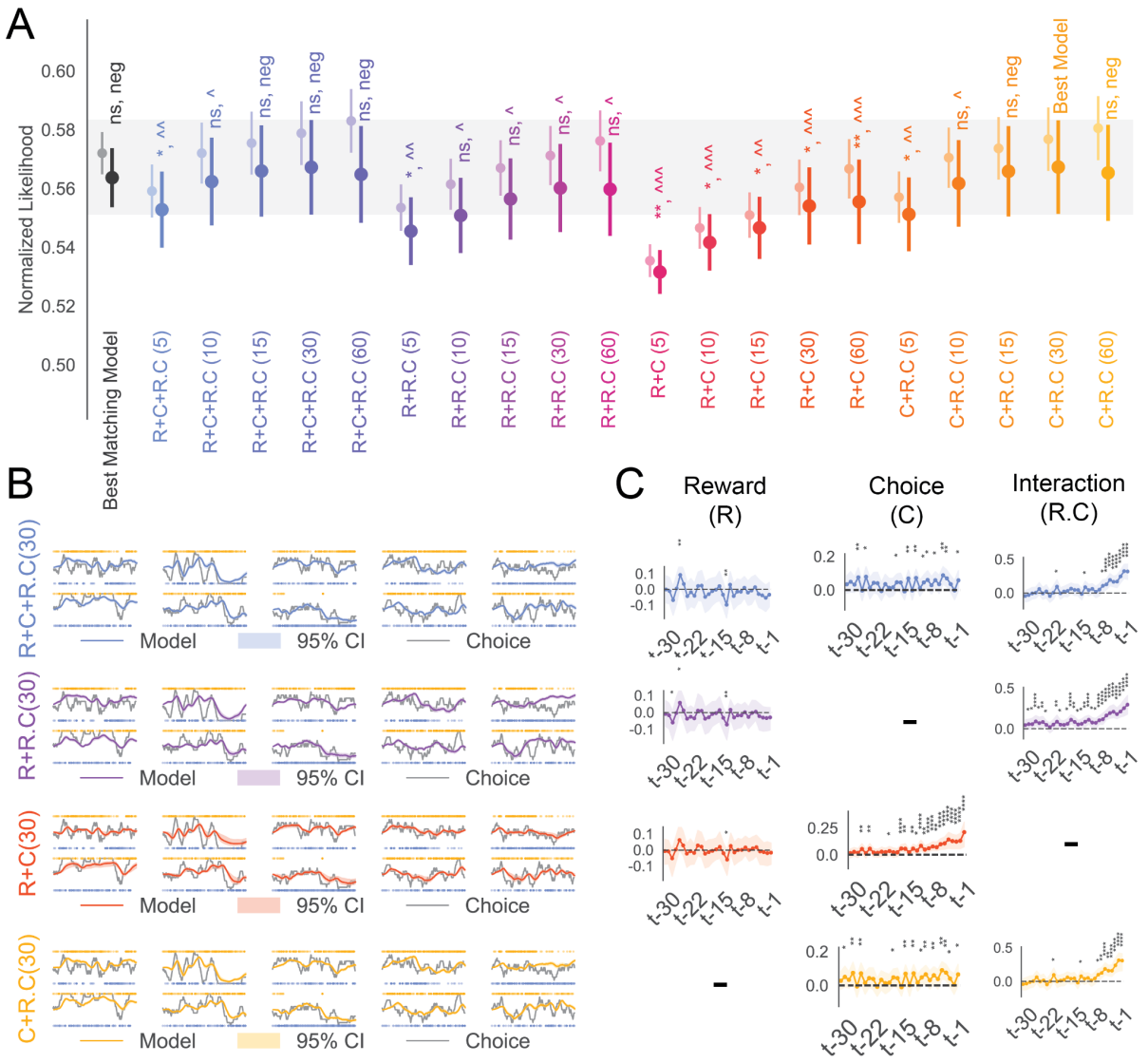


Figure 35. Logistic kernel regression models perform only as well as the best matching models

(A) Comparison of the goodness of fit and predictive power estimated using Normalized Likelihood on training data and testing data, respectively, for different logistic kernel regression models (see methods) with different sizes of integration windows. Light and dark error bars represent the mean and standard error of training and test Normalized Likelihood fitted using 1000 bootstrapped samples on the training dataset. Test Normalized Likelihood of each of the models is compared to the best model (C + R·C (30) Model with a 30 trial integration window) using a bootstrap-corrected two-sided paired samples Mann-Whitney-Wilcoxon test (stars for statistical significance) and bootstrap-corrected matched-pairs rank biserial correlation effect size (carets for effect size) (m=44 flies, n=1000 bootstraps; see

methods). See Table 25 for p-values and effect sizes, including a comparison of training Normalized Likelihood using the same statistical measures.

(B) Smoothed predicted choice probabilities for ten random test flies with a 95% confidence interval estimated from 1000 bootstrap fits overlaid on smoothed choice probabilities estimated from the data with a 10-trial window (see methods) for the four models with a 30-trial integration window.

(C) Kernel Regression Coefficients ($K_{X:t}$; see methods) for different terms estimated for the four models with a 30-trial integration window across 1000 bootstrap fits compared from zero using a two-sided bootstrap test (stars for significance). See Table 25, Table 26, Table 27, Table 28, Table 29 for the values of the coefficients and associated statistics.

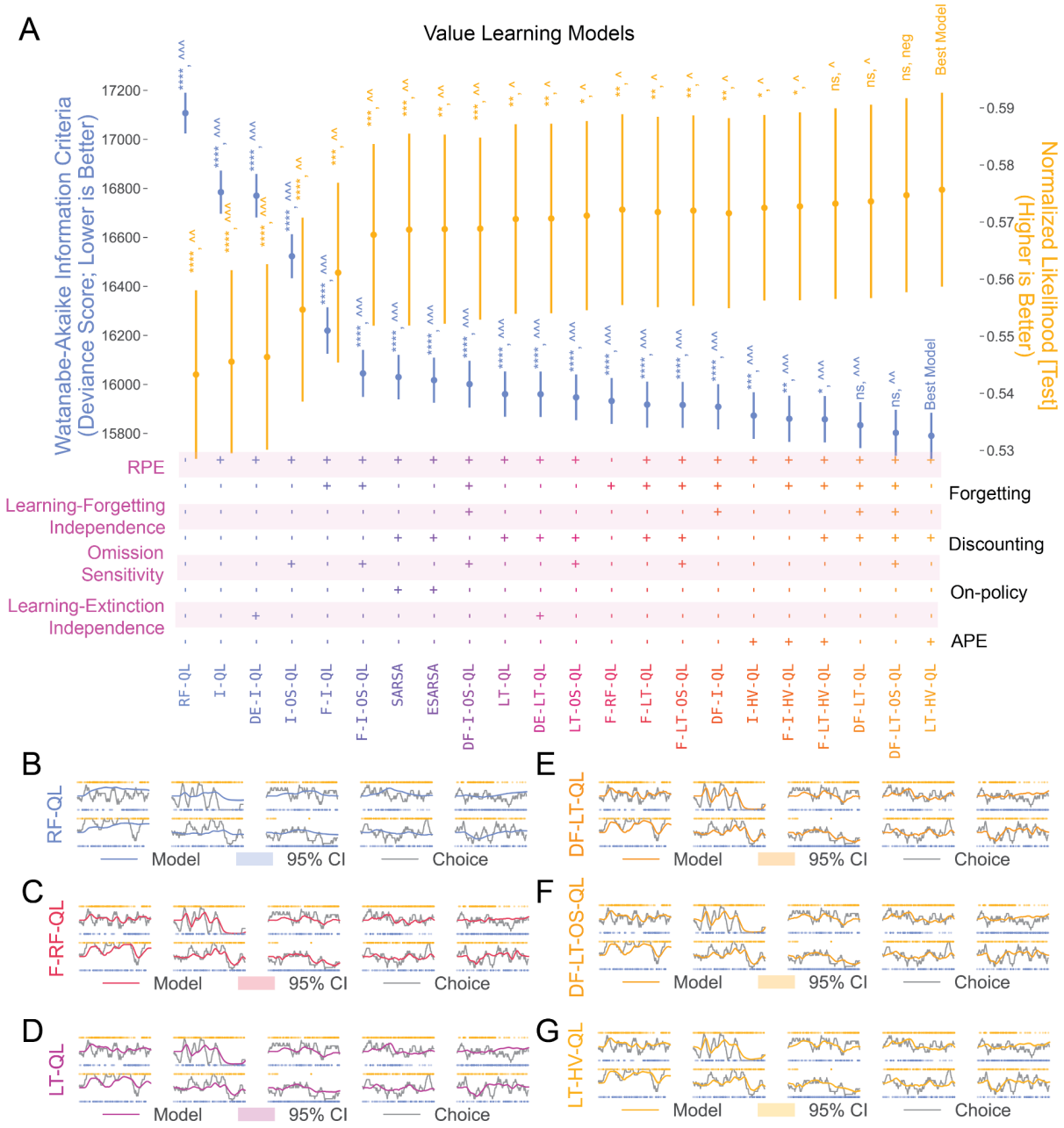


Figure 36. Results from fitting Q-learning models on Mohanta (2022) "Variable Block" dataset roughly reproduces the results from Rajagopalan "Fixed Block" (2022) dataset.

(A) Goodness of fit is estimated using the deviance-scaled Watanabe-Akaike Information Criterion (WAIC; blue). The difference of each model's WAIC relative to the best model is compared using a two-sided z-test (stars for statistical significance; see methods) and Cohen's d (carets for effect size). Predictive accuracy estimated using Normalized Likelihood [Test] (yellow) is compared relative to the best model using a bootstrap-corrected two-sided paired samples t-test (m=44 flies, n=100

bootstraps; see methods) (stars for statistical significance and paired Cohen's d (carets for effect size). The '+' and '-' symbols at the bottom signify which cognitive features (see Table 7) are included in the model. Error bars show Standard Error for WAIC and Normalized Likelihood [Test]. See Table 30 for statistics, p-values, and effect sizes.

(B–G) Smoothed predicted choice probabilities for ten random test flies with a 95% confidence interval estimated from 100 bootstrap fits overlaid on smoothed choice probabilities estimated from the data with a 10-trial window (see methods) for the six representative models from the dataset.

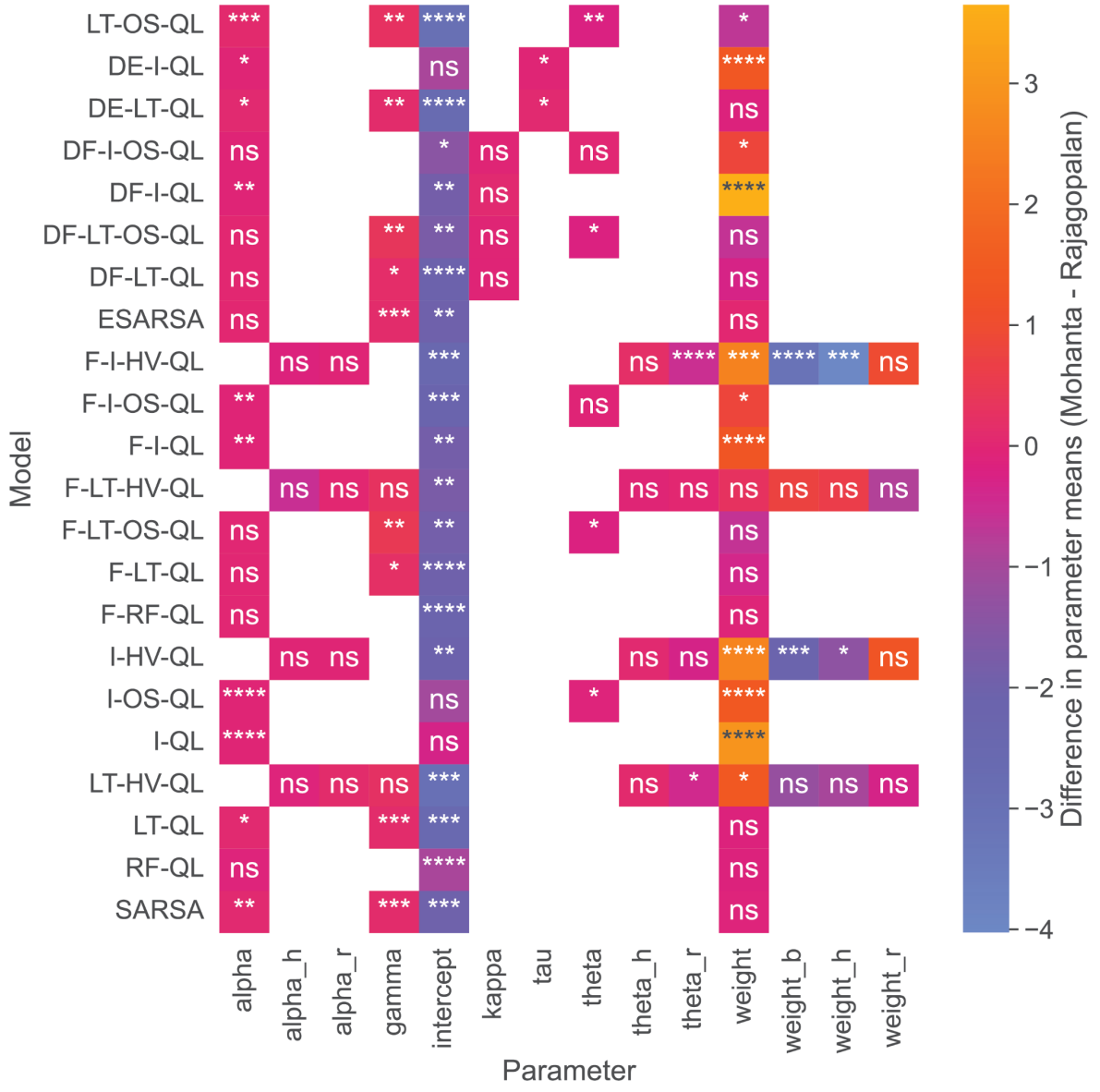


Figure 37. Difference between the parameter estimates from the Mohanta (2022) and Rajagopalan (2022) "Fixed Block" datasets.

Heatmap of the difference in the means of parameter estimates from the two datasets and the difference is tested using a simple z test (stars for statistical significance).

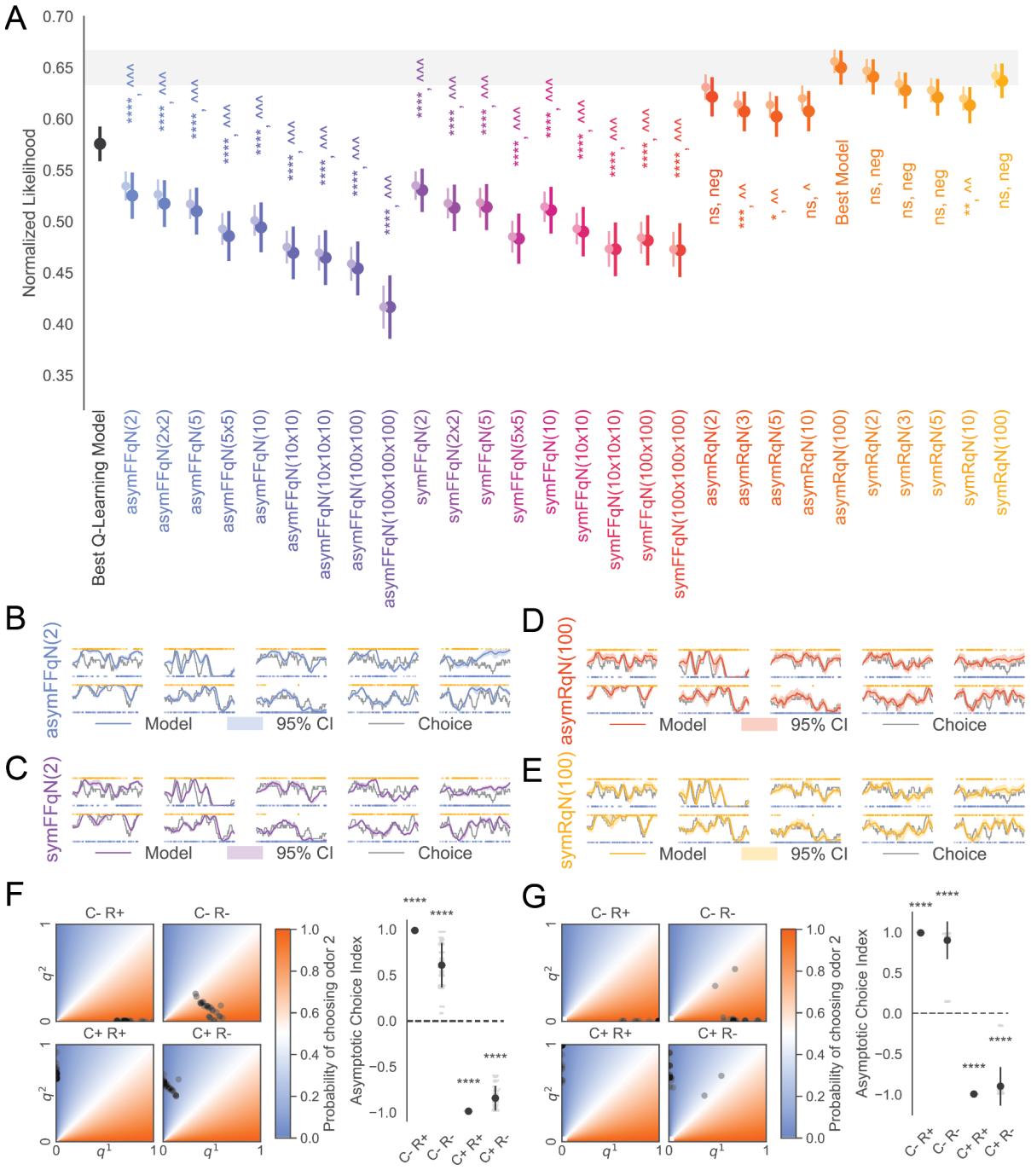


Figure 38. Results from fitting neural networks to the Mohanta (2022) "Variable Block" dataset also roughly reproduces the observations from Rajagopalan (2022) "Fixed Block" dataset.

(A) Comparison of the goodness of fit and predictive power estimated using Normalized Likelihood on training data and testing data, respectively, for different neural network architectures trained to estimate value from data and predict the choices. Light and dark error bars represent the mean and standard error of training

and test Normalized Likelihood, respectively. Test Normalized Likelihood of each of the models is compared to the best model (asymRqN(100)) using a bootstrap-corrected two-sided paired samples t-test (stars for statistical significance) and bootstrap-corrected paired cohen's d effect size (carets for effect size) (m=44 flies, n=25 ensembles for bootstrap correction; see methods). See Table 22 for p-values and effect sizes, including a comparison of training Normalized Likelihood using the same statistical measures.

(B–E) Smoothed predicted choice probabilities for ten random test flies with a 95% confidence interval estimated from 25 ensemble models for the best network architectures from each network class/variant overlaid on smoothed choice probabilities estimated from the data with a 10-trial window (see methods).

(F) Position of all the fixed point attractors across the trained and filtered ensemble of asymmetric FFqNs marked on the space of acceptance probabilities with black dots (left). Predicted preference of odors at the fixed point attractors of the different choice-reward conditions for all trained and filtered asymmetric FFqNs of the ensemble compared from zero using a two-sided bootstrap test (stars for significance; p=0.000 for all values).

(G) Position of all the fixed point attractors across the trained and filtered ensemble of symmetric FFqNs marked on the space of acceptance probabilities with black dots (left). Predicted preference of odors at the fixed point attractors of the different choice-reward conditions for all trained and filtered symmetric FFqNs of the ensemble compared from zero using a two-sided bootstrap test (stars for significance; p=0.000 for all values).