Title: Dopamine tracks adaptive learning of action representations

Authors: Maxime Come[1,2*], Arnaud Lespart[1*], Aylin Gulmez[1], Loussineh Keshishian[1,2], Joachim Jehl[1,2], Elise Bousseyrol[1,2], Steve Didienne[1,2], Eleonore Vicq[1,2], Tinaïg Le Borgne[1,2], Alexandre Mourot[1,2], Philippe Faure[1,2**]

Affiliations:

1. Brain Plasticity Laboratory, CNRS UMR 8249, ESPCI Paris, PSL Research University, Paris, France.
2. Sorbonne Université, Inserm, CNRS, Neuroscience Paris Seine - Institut de Biologie Paris Seine (NPS - IBPS), 75005 Paris, France.

*These authors contributed equally

** Corresponding author: phfaure@gmail.com

Abstract: (~150 word, here 146)

Flexible decision-making requires not only updating values, but redefining which features constitute an action in a given context. We recorded nucleus accumbens (NAc) dopamine release while mice navigated a three-target intracranial self-stimulation foraging task in which outcomes were evaluated under three distinct reward delivery rules. Despite a constant motor repertoire, dopamine transients reorganized across contingencies and generalized linear models revealed context-dependent dopamine signal reflecting action direction, recent outcome-history, or target identity. Reinforcement-learning model comparison showed that these signatures are best explained by distinct reward prediction errors (RPEs) defined over different state–action representations, rather than a single fixed model-free scheme. A single deep reinforcement-learning agent trained by temporal-difference learning, recapitulated both the rule-specific policies and the corresponding dopamine signature. These results identify NAc dopamine as a dynamic readout of representation learning, remapping prediction errors onto the task features that define successful action as contingencies change.

Main text

Introduction

Solving a puzzle often requires more than simply learning from outcomes—it also demands selecting which features matter depending on the context: A toddler might learn to sort objects by color in one game and by shape in another. This ability to flexibly reconfigure which environmental features are used to guide learning is fundamental to intelligent behavior. In computational terms, it requires not only the ability to learn action values from outcomes (reinforcement learning) but also to identify the relevant state space, i.e., the internal representation of the environment from which these values are derived. This process, known as representation learning, is increasingly recognized as a central challenge for neuroscience [1–4]. While reinforcement learning theories formalize how agents update value expectations through reward prediction errors (RPEs), such mechanisms only operate effectively when the agent has selected an appropriate set of task-relevant variables. Critically, selecting the appropriate state representation allows animals to overcome a fundamental computational bottleneck known as the *curse of dimensionality* [5,6] . In natural settings, choices often involve multiple features (e.g., spatial location, action history, outcome probability) that vary across contexts and time. Learning values for all feature combinations becomes intractable for standard reinforcement learning mechanisms, both because of the combinatorial explosion in possible states and because the mapping between features and reward can itself change [7]. Flexible behavior therefore requires a mechanism to extract and update low-dimensional, task-relevant representations—effectively, learning *what counts as an action* in a given context—so that values can be learned efficiently.

Understanding how animals form and update such internal representations is essential for describing flexible behavior, particularly in environments with changing or ambiguous task structures [3]. This issue has gained increasing importance as researchers shift their focus from experiments with a simple task structure to more elaborate tasks [8–12] that more closely resemble natural decision-making, with multiple (and possibly overlapping or competing) features that animals may use as state representations, as well as potentially abrupt changes over time in the state representations being used. Capturing the dynamics of representation change requires dedicated experimental and analytical approaches. While theoretical and behavioral studies have highlighted strategies to cope with high-dimensional environments[7], identifying neural signals that reflect the internal representations used to guide learning and decision-making remains a major challenge. This identification can provide critical insight into which features are effectively used by animals to build a relevant internal model of the world — to predict values, compute errors, and guide goal-directed actions.

Dopamine (DA) is a very well-established substrate to signal value and compute reward prediction error (RPE) [13–24], integrating outcome-related dimensions in a common currency [25–27], and driving reinforcement learning and decision making [19,25,26,28–31]. Because RPEs are necessarily computed within

65 a specific state and action space, variation of DA transients may reveal which representation the animal
66 is currently using to compute value. Recent work shows that DA can reflect inference over latent task
67 states and context dependent expectation [8,32] , as well as probability distributions over hidden causes [33] ,
68 suggesting sensitivity to abstract structure beyond raw sensory features. However, direct evidence that
69 DA dynamically reconfigures to track qualitatively different representations as animals shift between
70 fundamentally distinct task contingencies remains lacking.
71 Here, we address this gap using a spatial decision-making task in freely moving mice. We imposed three
72 reward rules [9,10,34,35] *i.e.,* deterministic, sequence-complexity and probabilistic, while keeping the same
73 physical actions and outcomes, isolating changes in the computational structure of the problem from
74 changes in motor demands. We combined this behavioral assay with fiber photometry recordings of DA
75 release in the nucleus accumbens (NAc) and a two-level computational modeling approach. First, we
76 used interpretable reinforcement learning models with hand-specified state and action features to formally
77 compare alternative candidate representations and identify which features best explain NAc DA dynamics
78 in each rule. Second, we trained a deep reinforcement learning (deep RL) agent with a simple feedforward
79 architecture that learns its own internal representations from minimal inputs, and asked whether a single
80 network can reproduce both the rule-specific behavioral policies and the rule-dependent modulation of
81 RPE-like signals. This approach tests whether representation learning—extracting task-relevant features
82 through error-driven plasticity—provides a parsimonious mechanism for the adaptive structure observed
83 in DA signals. Together**,** these analyses show that DA signals dynamically reflect RPEs computed from
84 context-specific features, revealing the internal model used by the animal in each condition, and that a
85 deep RL agent trained by temporal-difference learning naturally recapitulates these rule-specific patterns
86 under a fixed exploration policy. These findings position NAc DA as a real-time indicator of representation
87 learning, providing a window into how the brain adaptively constructs and updates action-relevant
88 structure in complex environments.
89
90 **Each reward context is associated with a specific reward-seeking strategy**
91 Using different versions of a spatial bandit task adapted for mice [9,10,34,35], we aimed to obtain rule-specific
92 and feature-dependent strategies. In this task, animals learned to navigate between three marked
93 locations in an open field, each associated with an intracranial self-stimulation (ICSS) of the medial
94 forebrain bundle (MFB). Mice could not receive two consecutive ICSS at the same location; and therefore,
95 had to alternate between rewarding locations, resulting in a sequence of movements and binary choices
96 (i.e., trials) **(Fig. 1A)**. Despite the apparent simplicity of this self-generated, goal-oriented behavior, mice
97 can use different features of the environment to guide their actions and obtain rewards (**Fig. 1B**). Mice
98 were initially trained in a deterministic context (*Det*) where all locations consistently delivered ICSS.

Although reward delivery was guaranteed at every location, this context nonetheless required animals to learn and internalize a simple structural rule: no two consecutive rewards could be obtained at the same location. As a result, mice had to alternate between the three available targets, producing sequences of choices constrained by spatial foraging dynamics. Initially, animals explored the environment with irregular trajectories and variable speeds. With training, they progressively optimized their behavior, developing stereotyped circular trajectories with minimal U-turns and ballistic movement profiles **(Fig. 2A)**. These behavioral adaptations reflect the acquisition of a minimal internal model of the task structure, allowing animals to predict that remaining at the same location would not yield further reward. This is further supported by the progressive increase in the number of trials **(Fig. S1A)**, reflecting improved efficiency in task execution throughout learning. Thus, although the Det context does not require complex credit assignment or value updating, it provides an essential baseline for evaluating the subsequent emergence of more sophisticated decision strategies when the reward structure becomes ambiguous or non-deterministic **(Fig 2B, Fig S1B-C)**.

In the complex context (Cplx) **(Fig 2B** middle**)**, reward delivery was determined by the variability in recent choice patterns **(**see methods and **Fig S2A)**[10], computed relative to the last nine decisions using a complexity algorithm. This rule requires animals to overcome a key limitation of model-free learning strategies. In contrast to the Det and Proba rules, where individual actions or spatial locations could be associated with consistent reward outcomes, the Cplx rule imposes no such direct contingency. Instead, animals must maximize variability in their behavior over time, independently of any fixed spatial preference or repeated motor strategy. This configuration renders previously learned policies a priori ineffective, and creates a setting in which trial identity loses its predictive power **(Fig S2B)**. Importantly, solving the task does not require tracking a large set of state-action pairs, which would be computationally costly, but rather adopting a *compressed representation* of the task, in which increasing sequence variability becomes the feature associated with reward [10]. As such, the Cplx rule constitutes a strong test of representation learning, assessing whether animals can adapt their internal model when the structure of reward contingencies is non-trivial and abstract.

In the probabilistic context (Proba), each location was associated with a fixed probability of reward delivery (p100, p50, or p25) **(Fig. 2B** right**)**. Unlike the Cplx rule, this setting involves stochastic outcomes and requires animals to integrate probabilistic feedback over time to optimize their choices. Successful performance relies on estimating expected values for each location and adapting spatial preferences accordingly. This context thus engages value-based decision-making based on feature-specific learning, with spatial location becoming the task-relevant dimension. In contrast to Cplx, where individual outcomes are largely uninformative without reference to prior behavior, the Proba condition allows a more classical

132 reinforcement learning strategy to emerge — but still requires flexible updating of value representations
133 as animals recover from the behavioral variability induced by the previous rule [35].
134 These three distinct reward structures resulted in distinct trajectory patterns **(Fig 2B)**, success rates **(Fig**
135 **2C)**, and decision-making strategies **(Fig 2D-G)**. In *Det*, animals tended to adopt circular trajectories with
136 minimal U-turns (~20%) **(Fig 2D)**. In contrast, the *Cplx* rule resulted in random trajectory patterns
137 characterized by high sequence complexity (**Fig 2E, Fig S2B)**. In *Proba*, mice exhibited a bias toward
138 locations with higher probability of reward delivery **(Fig 2G)**, resulting in a high percentage of U-turns (**Fig**
139 **2D**) and a preference for p100 and p50 (**Fig 2F**). We ensured that those differences in decision strategy
140 were not due to motivation or vigor to perform the three versions of the task **(Fig S3A-B)**. Overall, while
141 the basic design of the task remained constant, each rule instantiated a distinct causal relationship
142 between actions and rewards. The evolution of behavior across contexts demonstrates that mice are
143 capable of extracting and adapting to these contingencies, dynamically modifying their decision
144 strategies. This makes the task well-suited to track both behavioral flexibility and its neural correlates over
145 time, in particular through the lens of DA dynamics and internal representation learning.
146
147 **Nac Dopamine dynamics reveal expectations built upon rule-specific features**
148 We next examined DA release dynamics during the task, across the three rules, using the fluorescent
149 sensor $GRAB_{DA2m}$ [36] expressed in the lateral shell of the nucleus accumbens (NAc) in a new cohort of
150 wild-type male mice (**Fig 3A, Fig S4A**). Positive events in DA release occurred upon receiving expected
151 rewards, whereas negative events were observed when expected rewards were omitted **(Fig 3B-C, Fig**
152 **S4B)**, indicative of a negative RPE (for simplicity, these events, whether positive or negative, are referred
153 to as transients). Similar responses were observed while recording Ventral Tegmental Area (VTA) DA
154 neurons activity with GCaMP in DAT-iCre mice, ensuring consistency in the interpretation of DA dynamics
155 between release and somatic processes **(Fig S4C-D)**. Analysis of the amplitude distribution of DA
156 transients (positive and negative) across the different rules showed greater variability compared to
157 unexpected random stimulation in a rest cage, suggesting an active mechanism related to reward
158 expectation modulating the DA response, rather than being a mere response to the ICSS (**Fig 3D**). This
159 modulation is particularly important given that the ICSS directly activates MFB fibers, raising the possibility
160 of DA release being directly driven by stimulation. The observed variability across task contexts thus
161 strongly supports a dynamic, expectation-dependent component in DA signaling that add to the electrical
162 stimulation signal. Additional experiments with unexpected rewards delivered either during the task but
163 off-target (i.e. when the animal was in-between rewarded locations, **Fig 3E**) or in a rest cage (**Fig S5A-**
164 **D**) demonstrated a larger response compared to expected rewards during the task, yet only after
165 conditioning, further supporting the role of expectation in modulating DA release. We also controlled for a

potential impact of sensor fatigue and found no effect on DA signal when stimulations were given in the rest cage with varying durations in-between stimulation (matching those observed in the task, typically from 2s to 7s) **(Fig S5E-F)**. Altogether, these findings confirm that DA positive transients during the task are not merely driven by the delivery of MFB stimulation, but are modulated by reward expectation. The amplitude of DA signals varied depending on whether outcomes were expected or not, and this modulation differed across contexts and over time. These results are consistent with the notion that DA release reflects reward prediction error (RPE) signals, with positive transients following unexpected rewards and negative transients after unexpected omissions. Notably, similar ICSS pulses evoked different responses depending on behavioral context—highlighting that DA signals are influenced by ongoing task engagement and learning history, rather than being reflexive responses to stimulation alone. These observations support the use of DA transients as dynamic markers of outcome expectation. To better understand the basis of these expectations, we next asked: what task features are DA signals sensitive to, and how does this sensitivity vary across rules? To do this, we applied generalized linear models (GLMs) to analyze fluctuations in DA transients amplitudes across trials, running separate regression analyses for each individual mouse at the end of each rule (last two sessions) **(Fig 4A)**. The predictors included current and previous trial outcomes (P: reward or omission), the specific target where outcomes occurred (T: locations pA, pB, and pC; or p100, p50 and p25 in Proba), the direction taken (UT: Forward movement or U-turn) and side (RL: left or right choice). These variables were selected based on their relevance to the specific constraints of each task rule, and their potential to capture the different features that mice could use to build internal models of the environment. In the *Det* setting, where all trials were rewarded, we observed that the key predictor for differentiating trials was direction (UT) but not side nor target **(Fig 4B).** In the *Cplx* setting, trial outcome accounted for the biggest part of DA variation (positive for rewards, negative for omissions, **Fig 4C**), with an additional positive effect of previous outcome (having received an omission at trial n-1 increases DA signal at trial n), regardless of targets, side or directions. In *Proba*, this effect of previous outcome disappeared, and the target probability (T) significantly influenced DA variations **(Fig 4D)**.

Overall, the GLM analysis revealed that the primary drivers of DA fluctuations changed depending on the task setting, with direction, trial outcome, and target probability each playing distinct roles. Direct examinations of DA transients, categorized by direction, previous outcome or target, supported the strength of these effects. In *Det*, DA release varied with the direction **(Fig 4E, Fig S6A)** but not with the target (**Fig S6B**). In *Cplx*, omission on previous trial led to greater reward-induced and shallower omission-induced transients (**Fig 4F, Fig S6C**), while neither the target nor the direction showed significant effects (**Fig S6D-E**). At the end of the *Proba* setting, the DA signals were negatively influenced by target probability, with higher probabilities resulting in smaller positive transients for rewards and deeper

negative transients for omissions **(Fig 4G, Fig S6F)**. Finally, no effect of direction was observed on DA transients **(Fig S6G)**, and regarding outcome at previous trial, we observed a small effect only for rewarded trials **(Fig S6H).** Altogether, these results reveal specific patterns in the modulation of DA transients across task settings. Notably, DA fluctuations were not tied to the same task features across rules: outcome history dominated in Cplx, direction in Det, and target identity in Proba. This dissociation indicates that the features driving DA transients are not fixed, but flexibly selected depending on task structure—supporting the view that phasic DA signals adapt to the relevant contingencies of each context.

NAc dopamine cannot be explained by single model-free reinforcement learning but encodes state-specific RPEs

GLM analyses demonstrated that DA transients are modulated by distinct task features depending on the current rule—such as direction in Det, outcome history in Cplx, and target identity in Proba. However, they do not clarify *how* such features emerge nor whether the signal reflects a genuine reward prediction error (RPE). To directly test the RPE hypothesis, we used reinforcement learning (RL) models to generate trial-by-trial RPEs from the actual sequences of choices made by the animals (**Fig S7)**. A critical aspect of RPE computation is that it depends on the internal representation used to track expected value and on a well-defined set of competing actions with distinct consequences: from a given state, each choice must correspond to a different state transition (i.e., reaching a given location, such as pA, is uniquely determined by one action from that state, rather than being attainable via multiple equivalent actions). For instance, RPEs can be derived from predictions based on action direction (e.g., U-turn vs. forward), spatial target (e.g., pA vs. pB), or simply trial outcome regardless of action or location. These different assumptions correspond to different candidate representations of the task structure. By comparing theoretical RPEs from each model to recorded DA transients, we can infer which representation best matches the neural data in each context.

Based on our experimental results, we tested four models reflecting distinct representational schemes (**Fig. 5A**). Model 1 (M1) used a trial-based representation in which all choices updated a single global value, ignoring action or target identity. Model 2 (M2) assigned separate values to U-turn and forward actions, capturing a motor-based distinction. Model 3 (M3) tracked expected value for each target location independently. Finally, Model 6 (M6) implemented a non-abstracted, fully specified model-free scheme in which values were associated with each unique state-action pair (e.g., target × direction), reflecting learning without any structured generalization. All models were implemented using the Rescorla-Wagner rule to generate trial-by-trial RPEs, which were then used to assess their ability to reproduce key features of the DA signal in each task context.

In the Deterministic (Det) context, mice developed a directional bias in their choices, favoring forward movements over U-turns, which was mirrored by asymmetries in DA transients (**Fig. 4E**). Consistent with this, models M2 and M6 produced direction-dependent RPEs matching the fiber photometry data (**Fig. 5B–C**), whereas M1 and M3 did not (**Fig. 5C, Fig. S8A**), with M1 even predicting the opposite asymmetry. This supports the idea that, in Det, DA fluctuations reflect value representations incorporating action direction. In the complexity (Cplx) rule, where rewards depended on variability in recent choice sequences, DA transients were modulated by recent outcome history with no dependence on action or spatial features (**Fig. 4F**). Only the trial-based model M1 captured this pattern (**Fig. 5D–E**), whereas models with directional or spatial structure introduced effects not supported by the data (**Fig. 5E, Fig. S8B**), suggesting that animals relied on a simplified, history-tracking representation under Cplx. In the probabilistic (Proba) context, mice progressively developed spatial preferences aligned with fixed reward probabilities, and DA transients showed corresponding target-specific modulation (**Fig. 4G**). Models M3 and M6 reproduced these graded RPE differences across locations (**Fig. 5F–G),** although both tended to introduce additional modulations; among these, only the outcome-history effect (and not direction) was supported experimentally (**Fig. S6H** vs **Fig. S8C**).

Together, these results show that the DA signal does not reflect a unitary, fixed model-free mechanism (M6), but rather encodes context-specific RPEs based on dynamically selected internal representations. This flexible alignment between task structure and neural computation supports a key role for dopamine in representing the features that matter most for decision-making in each context.

A deep RL model captures behavior and Nac dopamine signals

To understand the computational principles that could jointly generate the behavioral trajectories and DA signals observed across rules, we next turned to a deep reinforcement learning (deep RL) approach. Rather than prescribing a specific representation a priori (trial-based, direction-based or spatial), we asked whether a simple neural network trained by RPEs could *learn* appropriate internal representations and reproduce both behavior and DA-like signals. We implemented a minimal feedforward network with one hidden layer trained by temporal-difference learning (Methods, **Fig. 6A**). On each trial, the network received the animal's current location (A, B, or C) and produced five action values (Q-values): two motor actions (Left, Right) and three target-specific actions (go to A, B, or C). Because motor and spatial descriptions can be partially redundant (e.g., "Left" and "go to A" may lead to the same choice), outputs are not mutually exclusive. The model sampled actions from a softmax over Q-values, received reward or omission according to the current rule, and computed a scalar RPE from obtained versus predicted outcome for the chosen action (Methods).

When exposed sequentially to the Deterministic, Complexity and Probabilistic rules, the deep RL agent reproduced (**Fig. 6B–F**) the main behavioral signatures observed in mice (**Fig. 2C–G**). Success rate decreased and U-turn rate increased at the Det to Cplx transition and at the Cplx to Proba transition (**Fig. 6B–C**). Sequence complexity increased specifically under Cplx (**Fig. 6D**), and the distribution of target visits and choice orientation reflected circular foraging in Det, random-like exploration in Cplx, and exploitation of high-probability targets in Proba (**Fig. 6E–F**). Thus, a single agent, with fixed architecture and learning rules, recapitulated the diversity of strategies expressed by animals across contexts.

To compare the model's internal "DA" signal with experimental data, we treated the trial-by-trial RPE as a proxy for NAc DA transients and applied the same GLM analysis to model-generated RPEs as to the recorded DA signals (**Fig. 6G**). The dominant predictors of model RPEs shifted across rules in close agreement with the data: direction in Det, current and previous outcomes in Cplx, and target identity (reward probability) in Proba (**Fig. 6G)**, with only a minor contribution of recent history in Proba. Thus, without explicitly encoding rule identity, the network learned input–action mappings that induced rule-specific RPE modulations aligned with the representational structure inferred from NAc DA.

Inspection of the network's output usage (**Fig. 6H**) clarified how these representations emerged. In Det, the agent relied predominantly on a single motor output (Left or Right), implementing a circular policy (clockwise or counter-clockwise) with few U-turns. In Cplx, Left and Right values converged toward similar levels and were used more evenly, generating high sequence variability and approximating a value-independent, randomness-seeking strategy. In Proba, the agent combined spatial and motor outputs to preferentially visit high-probability targets, yielding stable spatial preferences and frequent U-turns (**Fig. 6H**). Consistent with these learned value landscapes, Det produced a directional asymmetry in RPEs (smaller RPEs along the dominant direction than for U-turns), Cplx produced stronger positive RPEs after prior failures, and Proba produced target-dependent RPEs tracking exploitation of p100 and p50.

Together, these results show that a simple deep RL agent trained on the three rules can (i) reproduce the rule-specific behavioral policies observed in mice and (ii) generate RPE signals whose GLM fingerprints match rule-dependent NAc DA modulations, supporting the idea that flexible mappings from task features to RPEs can arise naturally from representation learning. Building on this, we next asked how these DA-coded representations relate to the *time-resolved* adaptation of choice policies across sessions in both animals and model.

DA representations track behavioral adaptation across rules

We next asked how the evolution of decision policies across sessions relates to changes in DA dynamics, and whether this coupling is consistent with the deep RL framework. We assessed the evolution of the betas from the GLM analysis of DA variations across sessions, along with the evolution of the betas from

the same analysis performed on the RPE generated by the deep RL model (see **Fig. S9A-C** for the evolution of behavior across session). By construction, these beta coefficients capture how strongly the internal value representation differentiates between options or outcomes, and can thus be directly related to evolution of choice bias and success rate.

In the Deterministic rule, the biased circularity profile could emerge from a value discrimination between two directional actions (right vs left), as suggest the evolution of βRL across sessions and its tendency to correlate with the circularity profile of mice (**Fig. 7A–B**). This behavior-RPE structure is strongly predicted by the deep RL model (**Fig. 7C–D**).

Upon transition to the Complexity rule, both mice and the deep RL agent initially experienced high omission rates because the previously optimal circular policy produced low sequence variability (**Fig. S9B**). Mice improved performance by increasing U-turns and redistributing them across trials to generate more variable sequences. Although U-turns could locally increase reward probability (**Fig. S10A**), animals did not adopt a simple "U-turn after omission" heuristic; instead, performance improved through a more global reorganization of sequences (**Fig. S10B–D**). The deep RL agent progressively equilibrate the value of the competing actions R and L, inducing a high entropy policy. As for the mice, this results in a redistribution of U-turns across sequences and a progressive increase of success rate. Notably, the emergence of a high-entropy policy did not require changes in the softmax exploration temperature β, which was held fixed. DA variations remained dominated by recent outcome history (larger responses after a prior reward than after a prior omission), with a tendency of this effect to decrease across sessions, and no correlation with improvements in sequence complexity or success rate (**Fig. 7E–F, Fig. S10E–F**). Regarding the deep RL agent: the history dependent RPE is already present since the start of complexity and decrease across sessions (**Fig. 7G–H**). The model provides a network-level mechanism for this effect with the presence of the hidden-layer which makes learning not confined to the selected output: updates propagate through common units and can partially influence other actions. Deeper simulation analysis demonstrate that this property could not be explained with a 1-layer model (**Fig S11**). Regarding the correlation between history effect and complexity of sequence of the deep RL agent (**Fig 7H**), it results from progressive evolution of both variable across sessions and not from a causal relationship. Thus, in this context, DA transients reflect a history-dependent RPE based on an abstract contingency (sequence variability), while policy adaptation proceeds through reconfiguration of representations and values rather than through DA-driven tuning of the softmax exploration temperature β.

The situation changed in the Probabilistic rule. Mice again exhibited an initial drop-in reward rate, now because targets differed in reward probability and the previously acquired variable policy was not yet aligned with these probabilities (**Fig. S9C**). Across sessions, animals increased success rate by biasing choices toward higher-probability targets, at the cost of more U-turns and more stereotyped trajectories.

334 In parallel, dependance of DA signal to target identity grew over time, and sessions with stronger
335 sensitivity of DA to target identity exhibited stronger preference for high-probability targets (**Fig. 7I–J**).
336 The deep RL agent mirrored these dynamics: under Proba, learned values diverged across targets,
337 producing increased RPE sensitivity to target identity and a progressive shift toward exploitation (**Fig. 7K–**
338 **L**). Across both Cplx and Proba, β remained constant, indicating that apparent changes in
339 "exploration/exploitation tradeoff" can arise from changes in learned value landscapes and internal
340 representations rather than from explicit modulation of a global exploration parameter (β).
341
342 **Discussion**
343 By recording dopamine (DA) release in the nucleus accumbens (NAc) while mice adapted to multiple
344 reward rules in a single spatial task, we asked how a canonical reinforcement-learning variable—reward
345 prediction error (RPE)—is implemented when the relevant state space itself must change. Across three
346 reward contexts, we found that DA transients did not simply encode a fixed scalar "value error" tied to
347 physical actions or locations. Instead, DA responses aligned with different internal descriptions of behavior
348 in each rule: direction-based in the Deterministic context, history-based in the Complexity context, and
349 probability-based in the Probabilistic context. By combining GLM-based model comparison, classical RL
350 models with hand-specified state features, and a deep RL agent that learns its own internal
351 representations, we show that NAc DA behaves as an RPE computed over flexible, learned action
352 representations, rather than over a rigid, experimenter-defined state space.
353 First, our results confirm and extend a consistent pattern observed across the dopamine literature,
354 wherein phasic DA carries information regarding both the obtained value and the RPE upon delivery or
355 omission of an expected reward [8,13–22]. It remains unclear whether this response stems from direct
356 stimulation of MFB DA fibers, resulting in DA release in the NAc, or whether it reflects a subjective value
357 mediated by circuits beyond the DA system alone [37,38]. While ICSS evoked positive transients across
358 contexts, their amplitude was strongly modulated by expectation. Unexpected rewards and omissions
359 triggered positive and negative transients respectively, consistent with classical RPE encoding [8,9,22,32].
360 Thus, our data reinforce the view that mesolimbic DA provides a relatively low-dimensional teaching signal
361 that can be read as a real-time RPE, even in freely moving animals performing spatially extended tasks
362 [8,22,27,39,40]. At the same time, the rule-dependent structure of DA modulation shows that this RPE is not
363 defined over a fixed "trial" or "arm" variable, but over an internal representation that can be reconfigured
364 as the task demands change.
365 Second, mice demonstrated substantial flexibility not only in their policy but also in the internal task-
366 state/action representation used for value learning. In the Deterministic rule, optimal performance was
367 achieved by adopting a stereotyped circular policy with few U-turns, and NAc DA was best explained by

368 an RPE defined over movement direction (U-turn vs forward), consistent with extensive evidence that
369 phasic dopamine in the nucleus accumbens reports reward prediction errors for well-learned action-
370 outcome contingencies[41] . In the Complexity rule, by contrast, the same circular policy became
371 detrimental because it produced low sequence variability. Mice recovered reward rate by increasing U-
372 turns and distributing them across trials, thereby raising global sequence complexity[10] . This aligns with
373 operant-learning literature showing that behavioral variability can itself be shaped by reinforcement
374 contingencies and brought under contextual control[42,43] . Strikingly, in this rule DA no longer differentiated
375 directions or targets, but instead was best accounted for by an RPE defined over recent outcome history,
376 consistent with an abstract contingency[33,44] that rewarded variability. Finally, in the Probabilistic rule,
377 animals exploited fixed reward probabilities associated with locations: they biased choices towards high-
378 probability targets, accepting more U-turns and more stereotypy. In this context, NAc DA reflected an
379 RPE over target identity and reward probability, with dependence of DA RPEs to locations growing in
380 parallel with the behavioral preference for high-probability sites, and consistence with classic work
381 showing that dopamine signals incorporate reward probability/uncertainty and subjective value[8,9,25,26].
382 Together, these observations indicate that mice not only adjust policy, but also switch which features of
383 actions and outcomes are treated as the relevant "state" variables for prediction and error computation, a
384 core idea in task-state/representation learning frameworks [1,6,45] , and that NAc dopamine tracks these
385 representation shifts, consistent with the view that RPE-like teaching signals can accompany (and
386 potentially drive) adaptive reconfiguration of learned representations[46] .
387 Beyond establishing a rule-dependent mapping between task features and DA-like RPEs, the deep RL
388 agent offers a minimal mechanistic account of *how* such mappings can emerge when actions admit
389 partially redundant descriptions. In our architecture, motor (Left/Right) and spatial ("go to A/B/C") outputs
390 can lead to the same choice, creating a credit-assignment problem in which the agent must attribute
391 outcomes to one of several partially overlapping action descriptions[47,48] . This simple "credit assignment
392 under redundancy" naturally produces the stereotyped circular policy in the Deterministic rule: value
393 concentrates on one directional output, making alternative choices effectively low-value deviations, and
394 generating the directional RPE asymmetry that matches the DA U-turn effect. Crucially, the same
395 architecture can also reproduce the Complexity signature—where DA is best captured by a history-
396 dependent, trial-level representation—because learning is not confined to the chosen output. Weight
397 updates propagate through shared hidden units, so that updating the value of a selected action also
398 partially updates the values of other, unchosen actions that rely on overlapping hidden-layer features. As
399 a result, the hidden layer rapidly comes to encode a coarse estimate of recent reward rate (or global
400 success), which is then broadcasted to multiple outputs. This provides a natural explanation for why the
401 "previous omission lead to larger subsequent reward transient" effect is present from the beginning of the

402 Complexity rule, both in the model and in DA recordings, without requiring a slow, rule-specific
403 accumulation process. More broadly, these results support a view in which dopaminergic teaching signals
404 are computed over *learned* internal representations rather than fixed experimenter-defined state spaces.
405 In the deep RL agent, plasticity in hidden layers both constructs such representations and shapes the
406 ensuing RPE signal, suggesting a plausible biological mechanism: the same dopaminergic error signal
407 could, via synaptic plasticity at multiple stages, progressively tune upstream networks to represent
408 precisely those features that best predict reward in a given context[47] . An important implication is that
409 individual "profiles" within a rule may emerge downstream of this shared rule-level representation, through
410 biased readout and attribution over the learned action manifold—an idea that can be directly tested by
411 quantifying whether hidden-layer activity is more strongly structured by rule identity than by within-rule
412 phenotypes.
413 The joint analysis of the evolution of dependence of DA RPEs of mice and RPE of the model to predictors
414 across sessions further constrains how we should think about DA's role in exploration and policy
415 adaptation. In the Complexity rule, mice increased sequence variability and improved performance without
416 developing any systematic DA dependence to specific actions, but displayed DA dependence to recent
417 reward vs omission. The deep RL agent exhibited the same qualitative pattern: it recovered reward rate
418 by making repetitive sequences predict low return and by increasing the entropy of its policy, yet did so
419 without any change in the softmax temperature β, which was held constant across rules. This implies that
420 the apparent increase in "exploration" in Cplx can emerge from changes in representations and value
421 estimates, rather than from a dedicated DA-driven modulation of a global exploration parameter.
422 Conversely, in the Probabilistic rule, both mice and model showed a tight coupling between the growth of
423 DA-RPE dependence to targets and the consolidation of an exploitative policy favoring high-probability
424 locations. Here, value-based asymmetries in DA (or RPE) and policy refinement evolved in parallel, again
425 under a fixed β in the model, because the learned value landscape became more uneven. These
426 comparisons argue that at least part of the DA–behavior relationship in our data can be explained by
427 standard RPE-based value learning over flexible state representations, without invoking an explicit DA
428 control signal for exploration temperature. This does not preclude additional, slower or circuit-specific
429 mechanisms for tuning exploration–exploitation balance[34,49,50] , but it suggests that they are not required
430 to account for the pattern we observe.
431 Several limitations of our approach point to directions for future work. First, fiber photometry in NAc
432 measures a population-averaged DA signal and cannot resolve heterogeneity across projection-defined
433 subpopulations or across core vs shell. Different subsets of DA inputs may carry distinct RPEs or
434 emphasize different state features[23,51,52] , which our bulk signal effectively averages. Second, our task
435 manipulates reward structure in a relatively simple, discrete set of rules; in more naturalistic

environments[8,53] , changes in contingencies may be more gradual or partially observable, and representation learning may involve hierarchical or belief-state-like structures that go beyond the models tested here. Third, our deep RL agent is intentionally minimal and does not include mechanisms such as meta-learning of learning rates, explicit uncertainty monitoring, or interactions with other neuromodulatory systems. Extending the framework to tasks with richer temporal structure or partial observability, and to models in which DA interacts with noradrenergic or cholinergic signals, will be essential to understand how different neuromodulators jointly shape internal models and exploration strategies.

In conclusion, our findings reveal that dopaminergic signals in the NAc carry a robust RPE-like teaching signal, but that this error is anchored to an internal, learned description of action and context that changes across reward rules. Behavioral adaptation in our task involves both policy changes and representation changes, and NAc DA tracks these representation changes in a way that can be captured by a single deep RL architecture. These results support a view of DA as a neural marker of representation learning in complex environments: not merely reporting errors in a fixed state space, but reporting errors in whatever task representation the brain has currently constructed to solve the problem at hand.

References

1. Niv, Y. Learning task-state representations. Nat Neurosci 22, 1544–1553 (2019).

2. Sutton, R. S. & Barto, A. G. Reinforcement Learning. (MIT Press, 1998).

3. Nakahara, H. & Hikosaka, O. Learning to represent reward structure: A key to adapting to complex environments. Neuroscience Research 1–7 (2012) doi:10.1016/j.neures.2012.09.007.

4. Gershman, S. J., Norman, K. A. & Niv, Y. Discovering latent causes in reinforcement learning. Curr. Opin. Behav. Sci. 5, 43–50 (2015).

5. Soltani, A. & Koechlin, E. Computational models of adaptive behavior and prefrontal cortex. Neuropsychopharmacology 47, 58–71 (2021).

6. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. Curr. Opin. Neurobiol. 20, 251–256 (2010).

7. Farashahi, S., Rowe, K., Aslami, Z., Lee, D. & Soltani, A. Feature-based learning improves adaptability without compromising precision. Nat. Commun. 8, 1768 (2017).

8. Krausz, T. A. et al. Dual credit assignment processes underlie dopamine signals in a complex spatial environment. Neuron (2023) doi:10.1016/j.neuron.2023.07.017.

9. Bousseyrol, E. et al. Dopaminergic and prefrontal dynamics co-determine mouse decisions in a spatial gambling task. Cell Rep. 42, 112523 (2023).

10. Belkaid, M. et al. Mice adaptively generate choice variability in a deterministic task. Communications Biology 3, 1–9 (2020).

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

12. Tervo, D. G. R. et al. Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. Cell 159, 21–32 (2014).

13. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. The Journal of neuroscience : the official journal of the Society for Neuroscience 16, 1936–1947 (1996).

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

15. Eshel, N., Tian, J., Bukwich, M. & Uchida, N. Dopamine neurons share common response function for reward prediction error. Nature Neuroscience 19, 479–486 (2016).

16. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47, 129–141 (2005).

17. Tobler, P. N., Fiorillo, C. D. & Schultz, W. Adaptive coding of reward value by dopamine neurons. Science (New York, N.Y.) 307, 1642–1645 (2005).

18. Schultz, W. Getting formal with dopamine and reward. Neuron 36, 241–263 (2002).

19. Steinberg, E. E. et al. A causal link between prediction errors, dopamine neurons and learning. Nature Neuroscience 1–10 (2013) doi:10.1038/nn.3413.

20. Roesch, M. R., Calu, D. J. & Schoenbaum, G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nature Neuroscience 10, 1615–1624 (2007).

21. Kutlu, M. G. et al. Dopamine signaling in the nucleus accumbens core mediates latent inhibition. Nat. Neurosci. 25, 1071–1081 (2021).

22. Mohebi, A. et al. Dissociable dopamine dynamics for learning and motivation. Nature 570, 65–70 (2019).

23. Jong, J. W. de, Liang, Y., Verharen, J. P. H., Fraser, K. M. & Lammel, S. State and rate-of-change encoding in parallel mesoaccumbal dopamine pathways. Nat. Neurosci. 27, 309–318 (2024).

24. Gershman, S. J. et al. Explaining dopamine through prediction errors and beyond. Nat. Neurosci. 1–11 (2024) doi:10.1038/s41593-024-01705-4.

25. Lak, A., Stauffer, W. R. & Schultz, W. Dopamine prediction error responses integrate subjective value from different reward dimensions. Proceedings of the National Academy of Sciences of the United States of America 111, 2343–2348 (2014).

26. Fiorillo, C. D., Tobler, P. N. & Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. Science (New York, N.Y.) 299, 1898–1902 (2003).

27. Hamid, A. A. et al. Mesolimbic dopamine signals the value of work. Nature Neuroscience (2015) doi:10.1038/nn.4173.

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

29. Bromberg-Martin, E. S. & Hikosaka, O. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron 63, 119–126 (2009).

30. Coddington, L. T. & Dudman, J. T. The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat Neurosci 21, 1563–1573 (2018).

31. Berke, J. D. What does dopamine mean? Nat Neurosci 21, 787–793 (2018).

32. Blanco-Pozo, M., Akam, T. & Walton, M. E. Dopamine-independent effect of rewards on choices through hidden-state inference. Nat. Neurosci. 1–12 (2024) doi:10.1038/s41593-023-01542-x.

33. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. Nat. Neurosci. 20, 581–589 (2017).

34. Dongelmans, M. et al. Chronic nicotine increases midbrain dopamine neuron activity and biases individual strategies towards reduced exploration in mice. Nat Commun 12, 6945 (2021).

35. Naudé, J. et al. Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking. Nature Neuroscience 19, 471–478 (2016).

36. Sun, F. et al. Next-generation GRAB sensors for monitoring dopaminergic activity in vivo. Nat. methods 17, 1156–1166 (2020).

37. Trujillo-Pisanty, I., Conover, K., Solis, P., Palacios, D. & Shizgal, P. Dopamine neurons do not constitute an obligatory stage in the final common path for the evaluation and pursuit of brain stimulation reward. PLoS ONE 15, e0226722 (2020).

38. Millard, S. J. et al. Cognitive representations of intracranial self-stimulation of midbrain dopamine neurons depend on stimulation frequency. Nat. Neurosci. 27, 1253–1259 (2024).

39. Coddington, L. T., Lindo, S. E. & Dudman, J. T. Mesolimbic dopamine adapts the rate of learning from action. Nature 1–9 (2023) doi:10.1038/s41586-022-05614-z.

40. Kutlu, M. G. et al. Dopamine release in the nucleus accumbens core signals perceived saliency. Curr. Biol. 31, 4748-4761.e8 (2021).

41. Hart, A. S., Rutledge, R. B., Glimcher, P. W. & Phillips, P. E. M. Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. Journal of Neuroscience 34, 698–704 (2014).

42. Grunow, A. & Neuringer, A. Learning to vary and varying to learn. Psychon B Rev 9, 250–258 (2002).

43. Neuringer, A. Operant variability: evidence, functions, and theory. Psychonomic bulletin & review 9, 672–705 (2003).

44. Babayan, B. M., Uchida, N. & Gershman, Samuel. J. Belief state representation in the dopamine system. Nat. Commun. 9, 1891 (2018).

45. Langdon, A. J., Song, M. & Niv, Y. Uncovering the 'state': Tracing the hidden state representations that structure learning and decision-making. Behav. Process. 167, 103891 (2019).

46. Alexander, W. H. & Gershman, S. J. Representation learning with reward prediction errors. arXiv (2021) doi:10.48550/arxiv.2108.12402.

47. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. Nat. Rev. Neurosci. 21, 335–346 (2020).

48. Richards, B. A. et al. A deep learning framework for neuroscience. Nat Neurosci 22, 1761–1770 (2019).

49. Cinotti, F. et al. Dopamine blockade impairs the exploration-exploitation trade-off in rats. Scientific Reports 9, 6770 (2019).

50. Beeler, J. A., Daw, N., Frazier, C. R. M. & Zhuang, X. Tonic dopamine modulates exploitation of reward learning. Frontiers in Behavioral Neuroscience 4, 170 (2010).

51. Lerner, T. N., Holloway, A. L. & Seiler, J. L. Dopamine, Updated: Reward Prediction Error and Beyond. Curr. Opin. Neurobiol. 67, 123–130 (2021).

52. Parker, N. F. et al. Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nat. Neurosci. 19, 845–854 (2016).

557 53. Solié, C. et al. Dopaminergic mechanisms of dynamical social specialization in mouse microsocie-
558 ties. bioRxiv 2025.02.10.637303 (2025) doi:10.1101/2025.02.10.637303.

559 54. Khabou, H. et al. Noninvasive gene delivery to foveal cones for vision restoration. JCI insight 3,
560 D358 (2018).

561 55. Sun, F. et al. A Genetically Encoded Fluorescent Sensor Enables Rapid and Specific Detection of
562 Dopamine in Flies, Fish, and Mice. Cell 174, 481-496.e19 (2018).

563 56. Choi, V. W., Asokan, A., Haberman, R. A. & Samulski, R. J. Production of recombinant adeno-asso-
564 ciated viral vectors. Current protocols in human genetics Chapter 12, Unit 12.9-12.9.21 (2007).

565 57. Aurnhammer, C. et al. Universal real-time PCR for the detection and quantification of adeno-associ-
566 ated virus serotype 2-derived inverted terminal repeat sequences. Human gene therapy methods 23,
567 18–28 (2012).

568 58. Lempel, A. & Ziv, J. On the Complexity of Finite Sequences. IEEE Trans. Information Theory 22,
569 75–81 (1976).

570

571

572

Acknowledgements:

This work was supported by the Centre National de la Recherche Scientifique CNRS UMR 8246 and 8249, INSERM U1130. We are grateful to the animal facilities (IBPS Sorbonne University) and Otilia de Oliveira and Emilie Tubeuf at ESPCI animal facilities. We also thank Jérémie Naudé and Clément Solié for their comments on the manuscript. We thank Deniz Dalkara and the viral core facility at the Vision Institute (Paris) for production of the AAVs. This work has received support under the Major Research Program of PSL Research University "PSL-Neuro" launched by PSL Research University and implemented by ANR (ANR-10-IDEX-0001). Our team is affiliated to DIM C-BRAINS, funded by the Conseil régional d'Ile-de-France.

Fundings:

The Foundation for Medical Research (FRM, Equipe FRM DEQ2013326488 to P.F, PhD fellowhsip ECO201806006688 to J.J., Fourth-year PhD fellowship FDT201904008060 to MC), the French National Cancer Institute Grant TABAC-19-020 and SPA-21-002 (to P.F.). French state funds managed by the ANR (ANR-19-CE16-0028 Bavar to PF, ANR-23-CE37-0017 VarSeek to PF). Memolife Labex, fourth-year PhD fellowship (to JJ and EV).

Authors contributions:

Conceptualization: MC, PF
Injection and implantation surgeries: MC
Behavioral experiments: MC, AG, LK
Fiber photometry recordings: MC, AG
Intracardiac perfusions and immunohistochemistry: MC, AG, EV, TLB
Data analysis: MC, AL, PF
Modelling: MC, AL, PF
Deep RL: AL, PF
Setups development: MC, JJ, AM, EB, SD, PF
Funding acquisition: PF
Writing - original draft: MC, AL, PF
Writing - review and editing: MC, TLB, AM, PF

Competing interests:

Authors declare that they have no competing interests

607 **Figure and Figure legends**

608

609

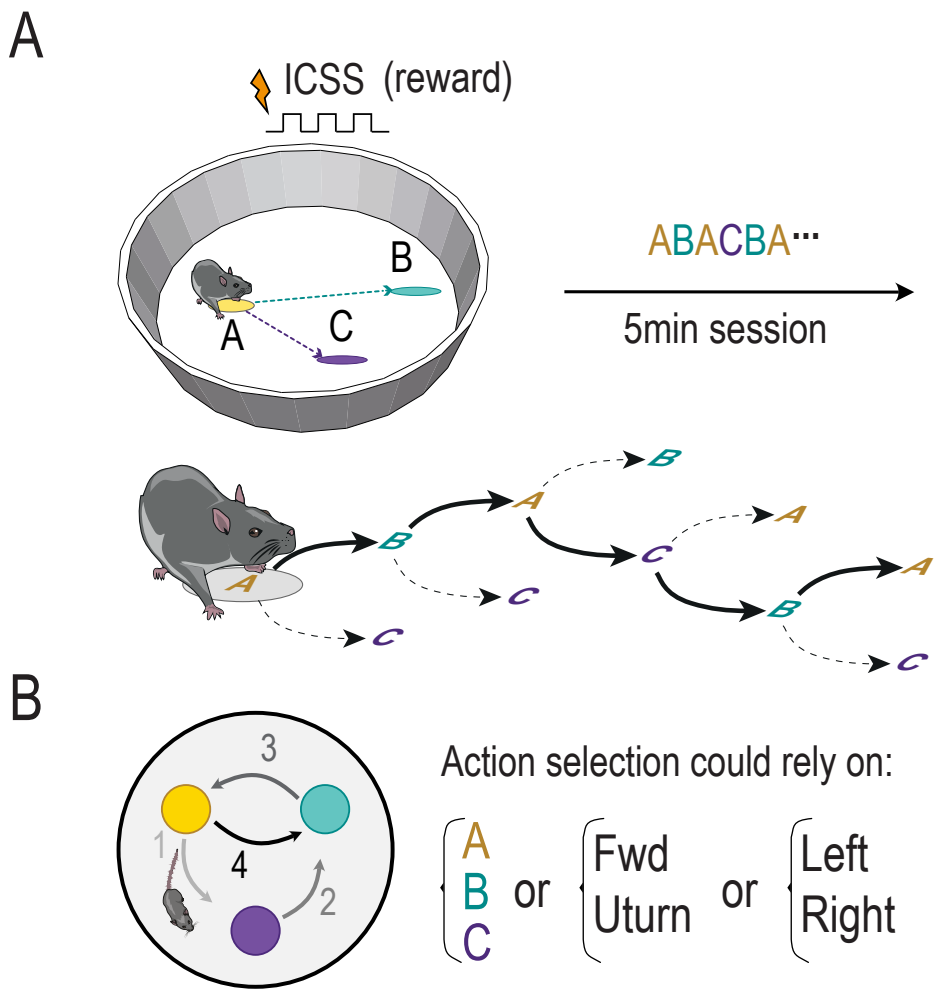

A
ICSS (reward)
B
A
C
ABACBA···
5min session
B
3
1
4
2
Action selection could rely on:
A
B
C
or
Fwd
Uturn
or
Left
Right

Figure 1

610 **Fig 1. Decision making context: (A)** Three explicit locations in an open field are associated with a
611 possibility of intracranial self-stimulation (ICSS)**.** Two stimulations could not be delivered consecutively in
612 the same location, therefore animals learned to alternate between target. Mice perform successive binary
613 choices to collect ICSS rewards. **(B)** Choice could rely on various overlapping sets of actions.
614
615

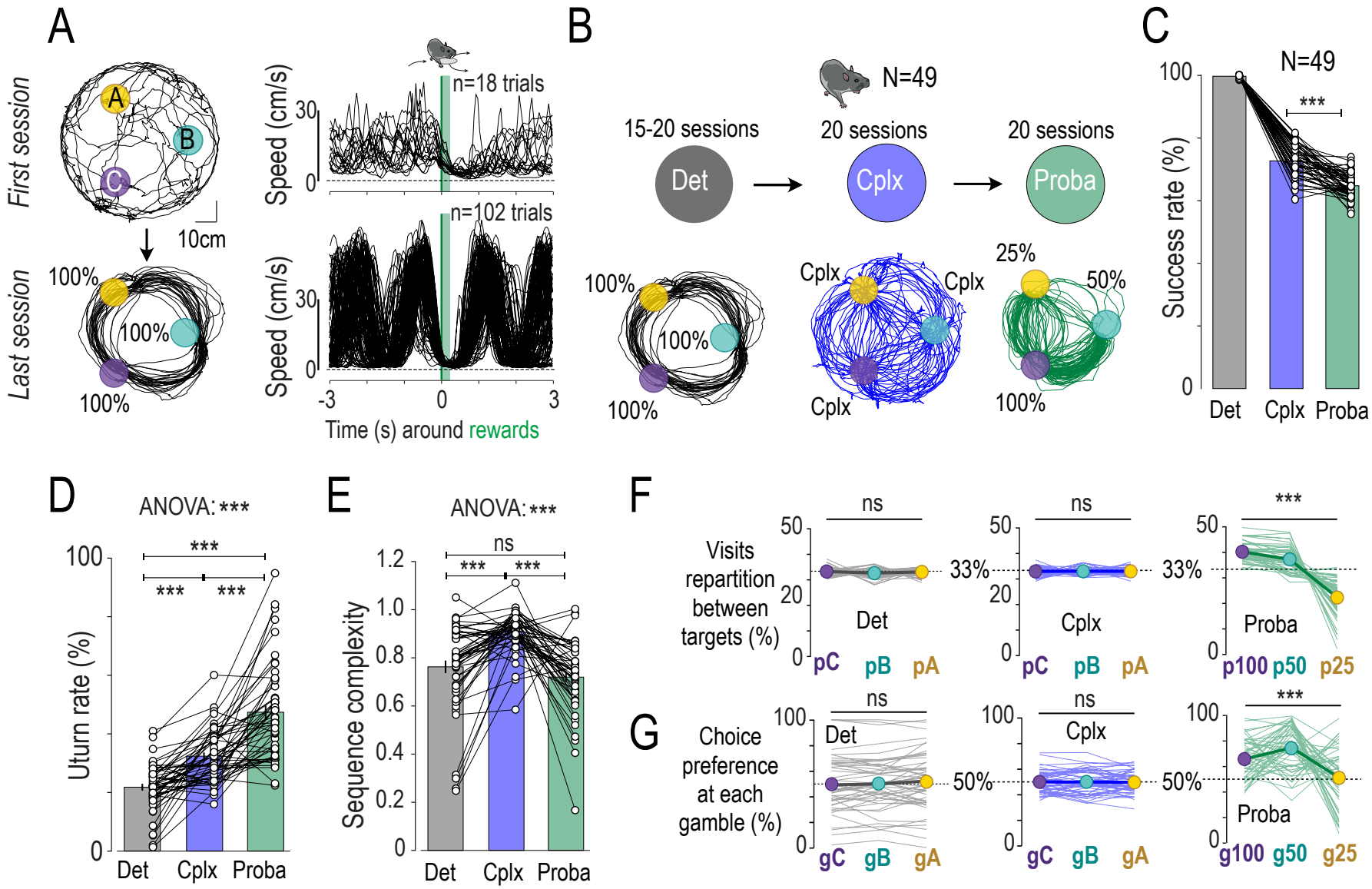

A
First session
Last session
A
B
C
10cm
100%
100%
100%
Speed (cm/s)
30
0
n=18 trials
n=102 trials
-3
0
3
Time (s) around rewards
B
N=49
15-20 sessions
20 sessions
20 sessions
Det
Cplx
Proba
100%
100%
100%
Cplx
Cplx
Cplx
25%
50%
100%
C
N=49
***
Success rate (%)
100
0
Det
Cplx
Proba
D
ANOVA: ***
***
***
***
Uturn rate (%)
100
0
Det
Cplx
Proba
E
ANOVA: ***
ns
***
***
Sequence complexity
1.2
1.0
0.8
0.6
0.4
0.2
0
Det
Cplx
Proba
F
Visits repartition between targets (%)
ns
ns
***
50
20
0
33%
Det
Cplx
Proba
pC
pB
pA
p100
p50
p25
G
Choice preference at each gamble (%)
ns
ns
***
100
0
50%
Det
Cplx
Proba
gC
gB
gA
g100
g50
g25

Figure 2

616 **Fig. 2. Mice display distinct reward seeking strategies adapted to each rule. A.** (Left) Trajectories
617 throughout conditioning and (Right) Speed profiles around ICSS in the deterministic context (First and
618 Last sessions). **B.** Three reward delivery rules were successively proposed: Deterministic (*Det*) where all
619 trials were rewarded (P=100%), Complexity (*Cplx*), where trials are rewarded based on sequence
620 variability, and Probabilistic (*Proba*), with each target associated to a given probability (P=25%, 50%, and
621 100%). **C-E** Comparison of success rate, sequence complexity and Uturn rate reveals distinct reward
622 seeking strategies across contexts. **F.** Locally, a mouse on one location (ex: p_A) has the choice between
623 the two other points (ex: p_B vs p_C), and therefore performs a gamble computed as $g_A = P(p_B|p_A)$. g>50%
624 corresponds to clockwise rotation for *Det* and *Cplx*, and to preference for highest probability of reward for
625 *Proba*. Proportion of target visits and choice preference at each gamble show a bias for circular foraging
626 in Det, exploitation in *Proba*, and randomness in *Cplx*. Data are shown as individual points, and mean
627 ±sem. N=49 mice (23 males and 26 females).
628
629

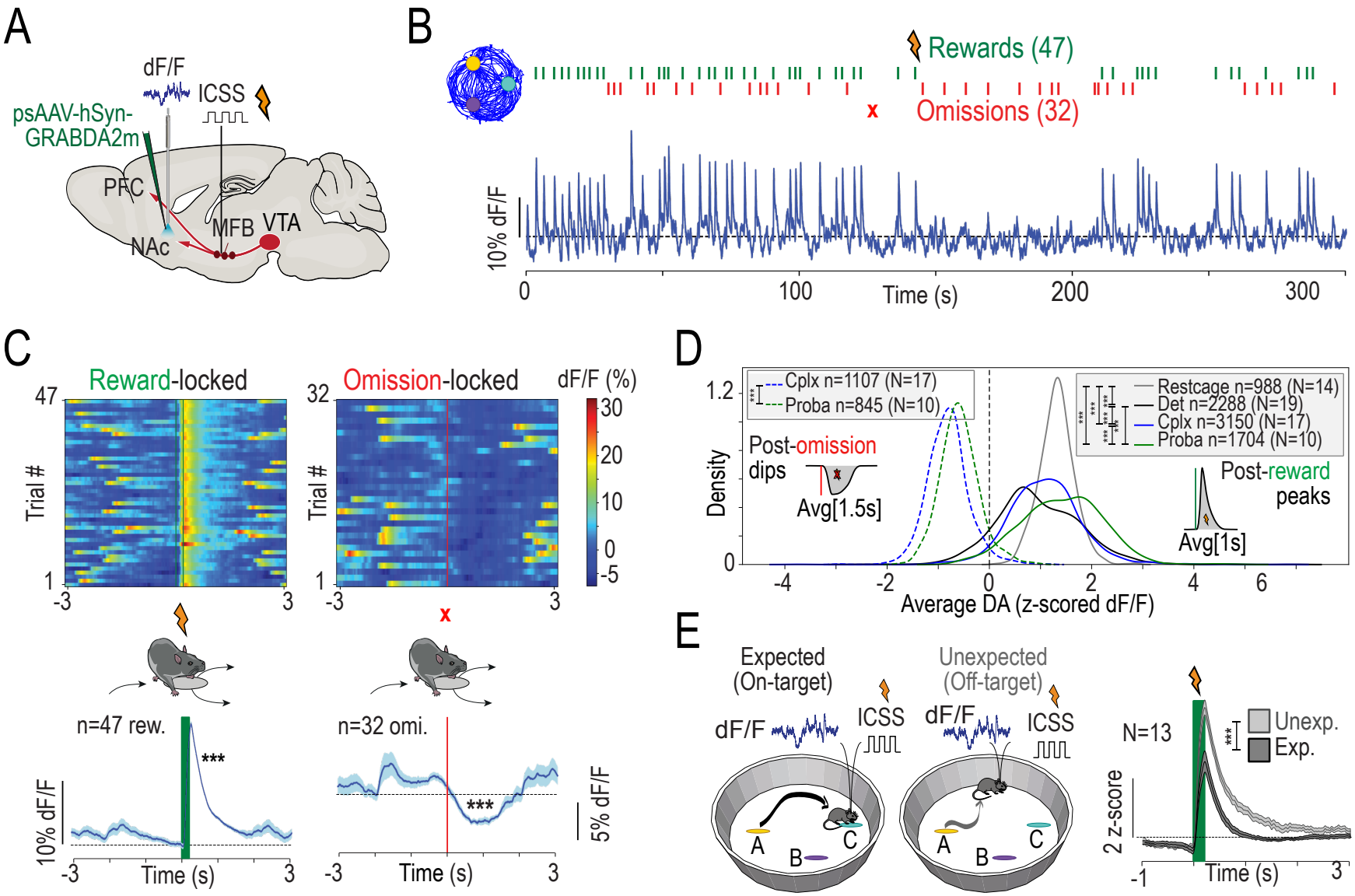

A
dF/F
ICSS
psAAV-hSyn-
GRABDA2m
PFC
MFB
VTA
NAc
B
Rewards (47)
Omissions (32)
10% dF/F
0
100
Time (s)
200
300
C
Reward-locked
Omission-locked
dF/F (%)
Trial #
n=47 rew.
n=32 omi.
10% dF/F
5% dF/F
Time (s)
D
Cplx n=1107 (N=17)
Proba n=845 (N=10)
Restcage n=988 (N=14)
Det n=2288 (N=19)
Cplx n=3150 (N=17)
Proba n=1704 (N=10)
Post-omission
dips
Avg[1.5s]
Post-reward
peaks
Avg[1s]
Density
Average DA (z-scored dF/F)
E
Expected
(On-target)
Unexpected
(Off-target)
dF/F
ICSS
N=13
Unexp.
Exp.
2 z-score
Time (s)

Figure 3

630 **Fig. 3: NAc DA release dynamics reveal expectations built upon rule-specific features. A.**
631 Schematic of the experimental design to record DA release during the task with chronic fiber photometry.
632 **B.** Representative signal from one 5-min session with rewards (green) and omissions (red). **C.** For the
633 same example session, signal is time-locked on location entry (t=0) and averaged. Rewards induce peaks
634 and omissions induce dips of DA release. **D.** Density distribution of averaged DA variations (Peak) for
635 rewards and omissions for the last two sessions of Det, Cplx or Proba, and for random stimulations in the
636 rest cage (performed on last day of Det). **E.** After conditioning, mice were randomly and unexpectedly
637 stimulated during the task outside of the rewarded zones (off-target), triggering DA peaks of greater
638 amplitude.
639
640

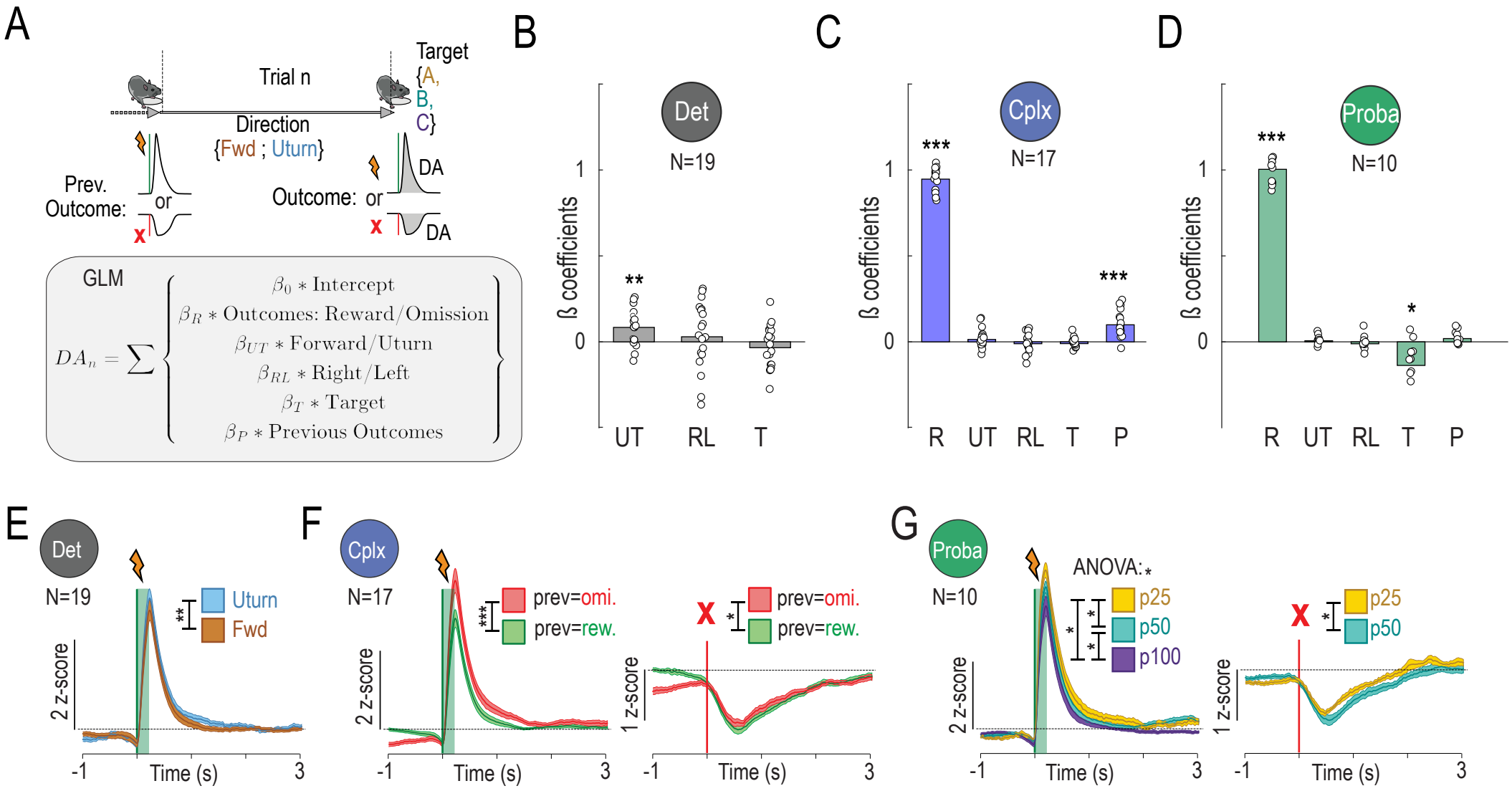

A
Trial n
Target
{A,
B,
C}
Direction
{Fwd ; Uturn}
Prev.
Outcome:
or
Outcome: or
DA
DA
GLM
$DA_n = \sum \left\{ \begin{array}{c} \beta_0 * \text{Intercept} \\ \beta_R * \text{Outcomes: Reward/Omission} \\ \beta_{UT} * \text{Forward/Uturn} \\ \beta_{RL} * \text{Right/Left} \\ \beta_T * \text{Target} \\ \beta_P * \text{Previous Outcomes} \end{array} \right\}$
B
Det
N=19
ß coefficients
UT
RL
T
C
Cplx
N=17
R
UT
RL
T
P
D
Proba
N=10
E
Det
N=19
Uturn
Fwd
2 z-score
Time (s)
F
Cplx
N=17
prev=omi.
prev=rew.
1 z-score
G
Proba
N=10
ANOVA:
p25
p50
p100

Figure 4

641 **Fig. 4: Defining sources of DA variations. A** Each trial is defined by predictors (outcome received,
642 previous outcome received, trajectory chosen to reach target, and target chosen) to fit DA amplitude using
643 GLMs. **B-D.** GLM results at the end of Det, Cplx and Proba. Features explaining DA variations vary across
644 contexts. **E-G.** Direct analysis of DA transients locked on those significant features (Uturn vs Fwd in Det
645 ; reward vs omission at previous trial in Cplx ; p25 vs p50 vs p100 in Proba). Data are shown as individual
646 points, and/or mean ±sem. n is the number of trials, N the number of mice in each condition.
647
648

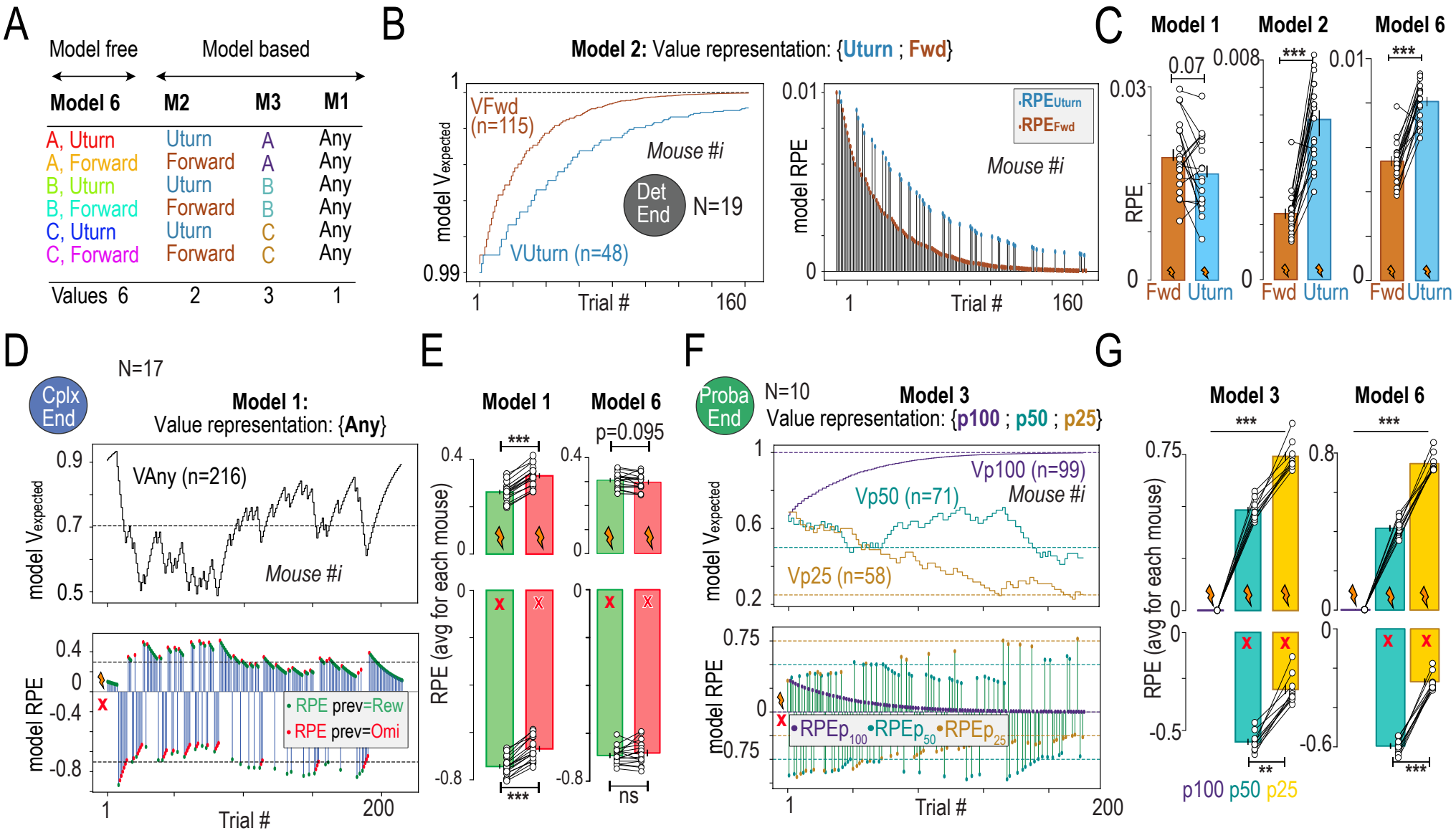

A
Model free
Model based
Model 6
M2
M3
M1
A, Uturn
A, Forward
B, Uturn
B, Forward
C, Uturn
C, Forward
Uturn
Forward
Uturn
Forward
Uturn
Forward
A
A
B
B
C
C
Any
Any
Any
Any
Any
Any
Values 6
2
3
1
B
Model 2: Value representation: {Uturn ; Fwd}
VFwd (n=115)
VUturn (n=48)
Mouse #i
Det End
N=19
model Vexpected
Trial #
RPEUturn
RPEFwd
model RPE
C
Model 1
Model 2
Model 6
0.07
0.008
0.01
RPE
Fwd
Uturn
D
N=17
Cplx End
Model 1:
Value representation: {Any}
VAny (n=216)
Mouse #i
model Vexpected
model RPE
RPE prev=Rew
RPE prev=Omi
Trial #
E
Model 1
Model 6
p=0.095
RPE (avg for each mouse)
ns
F
N=10
Proba End
Model 3
Value representation: {p100 ; p50 ; p25}
Vp100 (n=99)
Vp50 (n=71)
Vp25 (n=58)
Mouse #i
model Vexpected
model RPE
RPEp100
RPEp50
RPEp25
Trial #
G
Model 3
Model 6
RPE (avg for each mouse)
p100 p50 p25

Figure 5

649 **Fig. 5: DA signal embeds an RPE component, modelled from distinct value representations**
650 **specific to each rule. A.** Mice choice sequences were taken to train Reinforcement Learning (RL)
651 algorithms, testing three possible action representations to update values and compute corresponding
652 RPEs. Model 1 (M1) treats all trials equally with fluctuating { V_{Any} }. M2 updates a set of two distinct values
653 { V_{Fwd} ; V_{Uturn} }. A spatial model (M3) computes three independent values for each target { V_{pA} ; V_{pB} ; V_{pC}
654 }. M6 computes 6 independent value considering forward and Uturn for each position A, B or C. We then
655 trained another GLM assuming DA = $V_{obtained}$ + RPE, with trial RPEs generated from M1, M2, M3 and M6.
656 **B-G.** Models reproducing RPEs that explained DA variations vary across contexts. **B.** In M2-Det,
657 convergence toward 1 is slower for Uturns, leading to higher RPE_{Uturn} and reproducing DA data. **C.** M2
658 and M6 reproduce RPE data but not M1**. D.** In M1-Cplx, V_{Any} is always updated and fluctuates around
659 mean success rate. Plotting corresponding RPEs regarding current and previous outcomes mimic DA
660 data in M1 but not M6. **F-G.** In M3-Proba, value of each target converges and then fluctuates around its
661 probability, and corresponding RPEs reproduce DA data in M3 and M6. Data are shown as individual
662 points, and/or mean ±sem. n is the number of trials, N the number of mice in each condition.
663
664

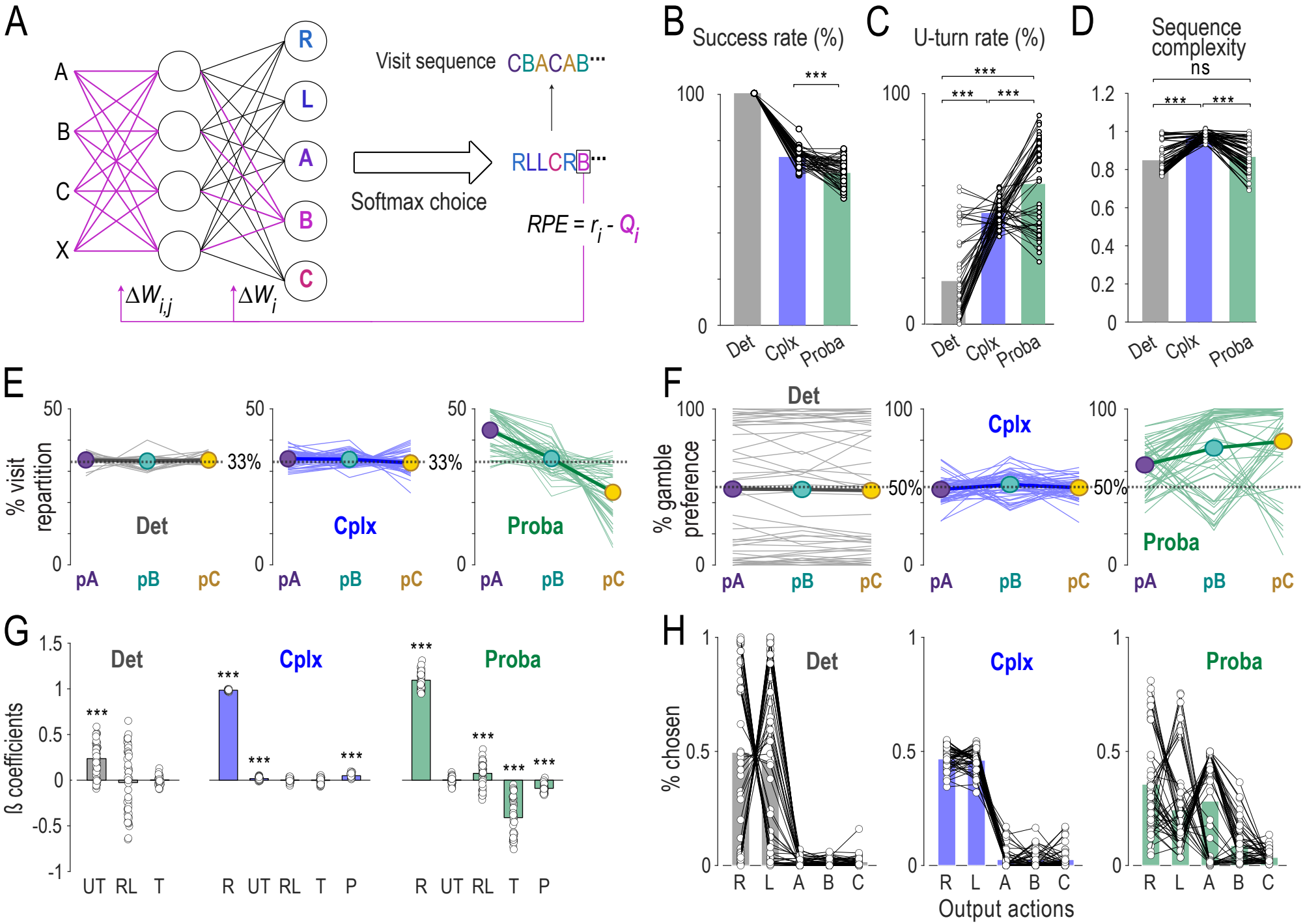

A
A
B
C
X
R
L
A
B
C
Visit sequence CBACAB···
Softmax choice
RLLCRB···
RPE = r_i - Q_i
$\Delta W_{i,j}$
ΔW_i
B
Success rate (%)
C
U-turn rate (%)
D
Sequence complexity
ns
Det
Cplx
Proba
E
% visit repartition
33%
Det
Cplx
Proba
pA
pB
pC
F
% gamble preference
50%
G
ß coefficients
UT RL T
R UT RL T P
H
% chosen
Output actions

Figure 6

Fig. 6: A deep RL network captures behavior and DA signals A. Schematic of the deep RL network, learning rule and sequence generation. The network outputs the values of a set of five actions: R: right; L: left; A, B and C: going to point A, B and C respectively. **B-D.** Comparison across rules of **B.** success rate, **C.** sequence complexity and **D.** Uturn rate **E.** Proportion of target visits **F.** Choice preference at each gamble. The latter shows a bias for circular foraging in Det, randomness in *Cplx* and exploitation in *Proba*. For Det and Cplx, the figure displays the preference for going right at each location, and for proba it displays the preference for the highest-value option at each location. **G.** Regression coefficients of a GLM at the end of Det, Cplx and Proba. Abreviations stands for: R: reward/omission; UT: forward/uturn; RL: right/left; T:target; P: previous trial's reward/omission. **H**: % of chosen action among the five output of the network. Data are shown as individual points for each simulation, and the bars represent the mean. N=50 simulations.

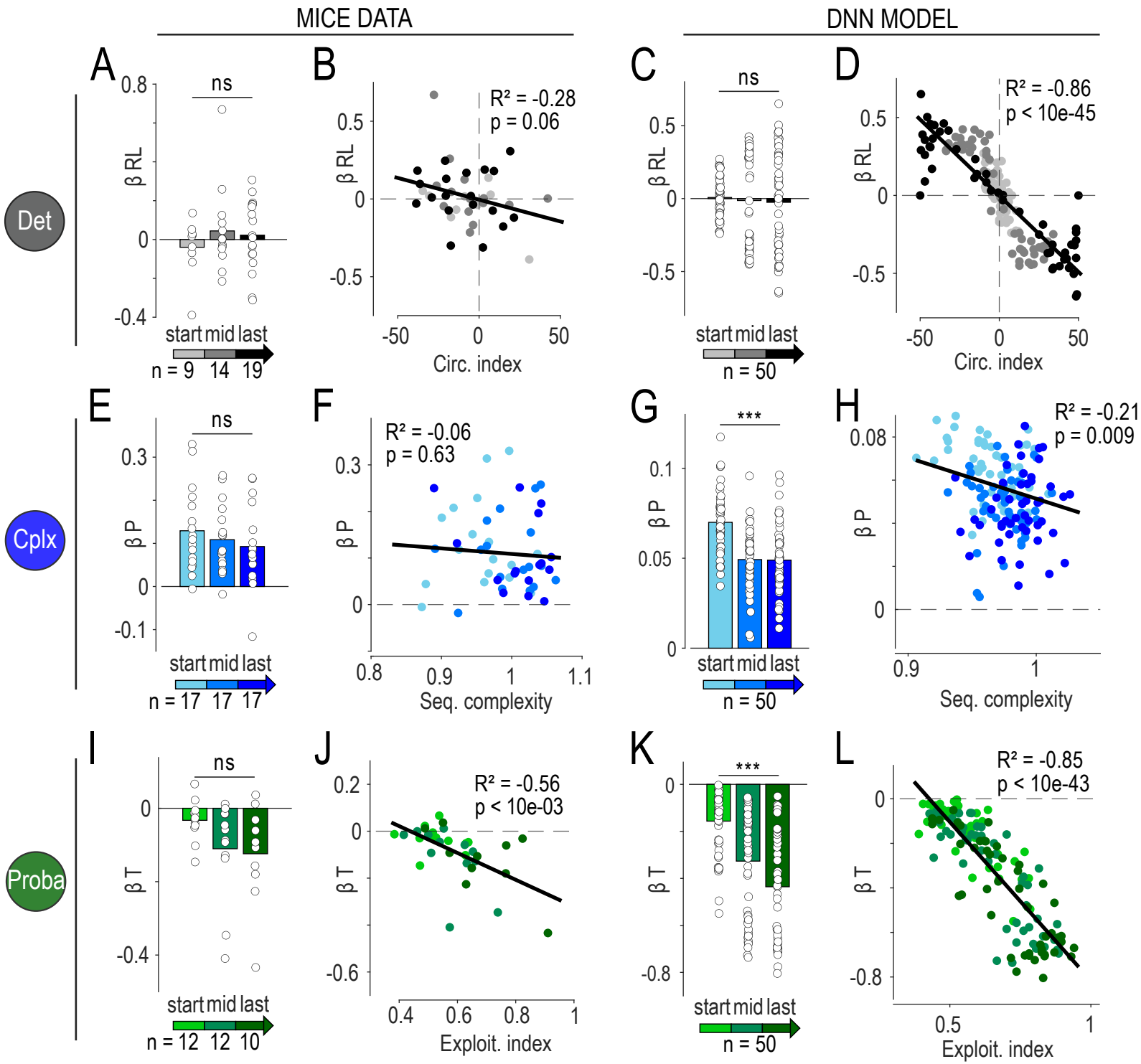

MICE DATA
DNN MODEL
Det
Cplx
Proba
A
B
C
D
E
F
G
H
I
J
K
L
ns
R² = -0.28
p = 0.06
R² = -0.86
p < 10e-45
R² = -0.06
p = 0.63
***
R² = -0.21
p = 0.009
R² = -0.56
p < 10e-03
R² = -0.85
p < 10e-43
β RL
β P
β T
Circ. index
Seq. complexity
Exploit. index
start mid last
n = 9 14 19
n = 50
n = 17 17 17
n = 12 12 10

Figure 7

Figure 7: Deep neural network shows similar performance and learning trajectories than mice behavior. A, E, I. Evolution of dependence of DA RPEs to principal predictor variable within each rule: **A.** 'right/left' (RL) in Det, **E.** previous outcome (P) in Cplx, **I.** 'target' (T) in proba. The evolution is shown across sessions spent in each rule with division of sessions into three categories 'start', 'mid' and 'last' which correspond to different stages in the training schedule. **C, G and K** show the same analysis performed on the deep RL simulations data. In figures **B,D,F,H,J,L** each data point represent a mouse (**B,F,J**) or a deep RL simulation (**D,H,L**) at one of the three stages of training schedule. **B, D.** In Det, linear regression between circularity index and dependence of RPEs to 'right/left' (RL) for mice (**B**) and deep RL simulation (**D**). **F, G.** In Cplx, linear regression between sequence complexity and dependence of RPEs to 'previous outcome' (P), for mice (**F**) and deep RL simulation (**G**). **J, L.** In Proba, linear regression between exploitation index and dependence of RPEs to 'target' (T) for mice (**J**) and deep RL simulation (**L**).

692 **Materiel and Methods**

693 **EXPERIMENTAL MODEL AND SUBJECT DETAILS**

694 Experiments were performed on adult C57Bl/6Rj wild-type mice (Janvier Labs, France). Both male and
695 female mice, weighing 20-30 g and 8 weeks old at the time of surgery, were used for behavioral
696 experiments. Only male mice were used in the $GRAB_{DA}$ fiber photometry cohorts. For cre-dependent
697 GCaMP experiments, DATiCre male mice were used. All mice were kept in an animal facility where
698 temperature (20 ± 2°C) and humidity were automatically monitored and a circadian 12/12h light–dark
699 cycle was maintained. All experiments were performed in accordance with the recommendations for
700 animal experiments issued by the European Commission directives 219/1990, 220/1990, and 2010/63,
701 and approved by Sorbonne University and ESPCI.
702

703 **METHOD DETAILS**

704 **AAV production**: AAVs for $GRAB_{DA2m}$ (pXR1-AAV-hSyn-GRAB-DA4.4) were produced as previously
705 described [54] (using the cotransfection method from plasmids generously provided by Dr. Yulong Li [36,55]
706 and purified by iodixanol gradient ultracentrifugation[56]). AAV vector stocks were tittered by quantitative
707 PCR (qPCR) [57] using SYBR Green (Thermo Fischer Scientific). AAV vectors for GCaMP6f (AAV1-EF1a-
708 DIO-GCaMP6f-P2A-nls-dTomato) and GCaMP7c (pGP-AAV1-syn-FLEX-jGCaMP7c variant 1513-
709 WPRE) were directly ordered from Addgene.
710 **Intracranial self-stimulation (ICSS) electrode implantation:** Male and female WT mice were
711 anaesthetized with a gas mixture of oxygen (1 L/min) and 1-3% of isoflurane (Piramal Healthcare, UK)
712 and then placed into a stereotaxic frame (Kopf Instruments, CA, USA). After the administration of a local
713 anesthetic (Lurocain, 0.1 mL at 0.67 mg/kg), a median incision revealed the skull, which was drilled at the
714 level of the median forebrain bundle (MFB). For ICSS, a bipolar stimulating electrode (PlasticOne 2
715 channels, stainless steel, 10 mm) was then implanted unilaterally (left or right, randomized) in the brain
716 using the following stereotaxic coordinates (from bregma according to Paxinos atlas): AP −1.4 mm, ML
717 ±1.2 mm, DV −4.8 mm from the brain). Dental cement (SuperBond, Sun Medical) was used to fix the
718 implant to the skull. An analgesic solution of buprenorphine at 0.015 mg/L (0.1 mL/10 g) was delivered
719 prior to awakening from the surgery and, if necessary, the following recovering days. After stitching, mice
720 were placed back in their home-cage and had a minimum of 5 days to recover from surgery. The efficacy
721 of electrical stimulation was verified through the rate of conditioning during the deterministic setting (see
722 Intracranial Self Stimulation (ICSS) bandit task). Out of the 54 mice implanted (27 for each sex), 49 were
723 included in the results (23 males and 26 females).

724

725 **Virus injections and fiber photometry recordings:** 3 cohorts of WT male mice (total of 24) were
726 anaesthetized (Oxygen 1 L/min, Isoflurane 1–3%) and implanted with an ICSS electrode as described
727 above. They were then injected unilaterally (randomized left/right side and ipsi/contralateral side regarding
728 the ICSS electrode) in the NAc lateral shell (1 µL, coordinates from bregma: AP +1.45mm; ML ±1.55mm;
729 DV −4.05mm from the skull) with an adeno-associated virus [36,55] to express $GRAB_{DA2m}$. An optical fiber
730 (200 µm core, NA = 0.39, Thor Labs) coupled to a metallic ferule (1.25 mm) was implanted 100 µm above
731 the injection site in target region and cemented to the skull with blackened cement. 5 DATiCre male mice
732 followed the same procedures for GCaMP experiments in the VTA (1 µL, coordinates from bregma: AP -
733 3.10mm; ML ±0.50mm; DV −4.20mm from the brain), 3 of them with GCaMP7c and 2 with GCaMP6f.
734 Viral expression typically took 10-15 days to achieve a satisfying signal and lasted for up to 3 months.
735 However, some mice exhibited a shorter duration of expression and were therefore excluded for the
736 analysis of later sessions. Although the mice performed the task on a daily basis, fluorescent recordings
737 were made only every 2 or 3 days to prevent sensor bleaching. Low power (100-200 mA) LEDs (465 nm
738 and 405 nm, Doric Lenses) coupled to a patch cord (500 µm core, NA = 0.5, Prizmatix) were used for
739 optical stimulation of the sensors in lock-in mode (572.205 Hz for the 465 nm LED, 208.616 Hz for the
740 405 nm LED) and collection of 520 nm fluorescence. 405 nm was used as the isobestic wavelength. The
741 optical stimulation patch cord was plugged onto the ferrule during all experimental sessions, even those
742 without recordings, to habituate animals and control for latent experimental effects. After the daily session,
743 a short recording of the autofluorescence signal $F(auto)$, coming from the patchcord only, was
744 performed with same LED intensities, no animal plugged and room in the dark. Raw 520 nm fluorescence
745 was demodulated by the software (Doric Lenses) to extract 465 nm and 405 nm signals. The 405 nm
746 signal was visually checked to account for instability artefacts coming from head movements or patch
747 cord unplugging during the session, and if needed correct the associated 465 nm signal accordingly,
748 otherwise it was not used for signal treatment. 465 nm signal F_i follows several treatment steps according
749 to this formula:

750
$$\frac{dF_i}{F_0} = \frac{F_i - F(auto) - F_i(fit)}{F_i(fit)} - 1$$

751 First F_i is subtracted with the constant value of autofluorescence *F(auto)* measured with patch cord only,
752 improving drastically the signal-to-noise ratio. Then, largest transients induced by ICSS were excluded to
753 perform a smoothing on the subsequent truncated signal. We then computed a mono-exponential fit
754 $F_i(fit)$ on this smoothed signal, which was also subtracted to F_i at each time point i to account for
755 exponential decay. The result is then divided by the same $F_i(fit)$ at each time point i to normalize the

signal around 1, and subtracted by the constant 1 to normalize to 0 and obtain positive or negative transients as dF_i/F_0 over an entire session (5 or 10min). In order to aggregate signals coming from different sessions for each mouse, and then pool mice for the analysis, we also applied a z-scoring on dF_i/F_0 over each entire session.

Intracranial self-stimulation (ICSS) bandit task. The ICSS bandit task [9,10,34,35], took place in a circular open-field with a diameter of 68 cm. Three explicit square-shaped marks (2 × 2 cm) were taped in the open field, forming an equilateral triangle (side = 35 cm). Entry in the circular zones (diameter = 6 cm) around each mark was associated with the delivery of a rewarding ICSS stimulation. A LabVIEW (National Instruments) application precisely tracked and recorded the animal's position with a camera (20 frames/s). When a mouse was detected in one of the circular rewarding zones, a TTL signal was sent to the electrical stimulator, which generated a 200 ms train of 5 ms biphasic square waves pulsed at 100 Hz (20 pulses per train). Two consecutive rewards could not be delivered on the same target, which motivated mice to alternate between targets and therefore generate sequences of binary choices. ICSS intensity was adjusted, within a range of 15-200 µA, during early conditioning sessions, so that mice would achieve between 50 and 120 visits per session (5 min duration) for two successive sessions. ICSS intensity was then kept constant for all the experiments, even when reward delivery rules changed. Mice with insufficient scores were excluded. Different reward delivery rules were used, and all animals went through all three protocols successively. The first is a deterministic (Det) setting, with 10 to 17 daily sessions of 5 min. All zones were associated with an ICSS delivery (P = 100%). The second, described previously in [10], is a complex (Cplx) setting where a grammatical complexity algorithm [58] analyses online the choice sequence that the mouse is producing, calculates the complexity of two potential sequences of length 10 (9 past targets + next target among the 2 available) and gives a reward only if the complexity of the sequence increases. Repeating patterns of low complexity will therefore lead to series of omissions, while increasing variability will increase success rate. Mice did daily sessions during 15-20 days. The third setting is probabilistic (Proba): each target is associated with a probability to obtain an ICSS stimulation among three (P = 25%, P = 50%, P = 100%), as described previously [9,34,35]. The probabilities at each location were pseudo-randomly assigned per mouse, and 15-20 sessions were performed. 2 cohorts of both male and female mice followed deterministic, complexity and probability settings successively, with no fluorescent sensor expression. Three cohorts of male mice expressing $GRAB_{DA}$ and implanted with an optical fiber implantation followed different settings: *i)* the first cohort performed only Det and Cplx, and recordings started only at the end of Det, *ii)* the second and third cohorts performed Det, Cplx and Proba, with recordings starting at the beginning of Det, and performed also some control experiments (especially, unexpected rest cage and off-target ICSS). Consequently, there is variation in animal numbers among

conditions in the figures. Finally, one cohort of DATiCre male mice was tested in Det and Cplx only. We defined three analysis windows within each condition (see Fig 7) - Start, Late, and End - corresponding to the following experimental session ranges: Deterministic (Det) sessions 6–7, 11–12, and 14–17; Complexity (Cplx) sessions 1–4, 5–8, and 10–13; and Probabilistic (Proba) sessions 1–2, 3–6, and 11–16.

QUANTIFICATION AND STATISTICAL ANALYSIS

Behavioral measures: For all those groups, the following measures were analyzed with custom codes in Python (using mostly Numpy and Pandas libraries, on PyCharm CE) and compared throughout the different rules: *i)* number of visits, *ii)* success rate, *iii)* time-to-goal, *iv)* choice repartition (proportion of visits at each location), *v)* percentage of U-turn (target n = target n+2) and *vi)* sequence complexity (applying the same complexity algorithm calculation but offline and for all choices during a session). Furthermore, the ICSS bandit task can be seen as a Markovian decision process: every transition can be considered as a binary choice between two options, since a zone cannot be reinforced twice in a row. The sequence of choices per session results from the succession of three specific binary choices, or gambles. For deterministic and complexity, $G_C = P(A|C)$ would be the total number of visits in target A divided by the total number of visits in targets A and B, when the animal is in target C. Similarly, $G_A = P(B|A)$ and $G_B = P(C|B)$. A gamble above 50% indicates that the animal has a preference for moving clockwise (or below 50% for moving counter-clockwise). In probabilistic, direction of conditional probabilities does not follow spatial repartition of locations, but rather preference for the high value option: G_{25} = 100% vs 50%, G_{100} = 50% vs 25% and G_{50} = 100% vs 25%. Applying this principle at each choice, those 3 gambles can be aggregated into single values to give circularity index (going in circle, no matter clockwise or counter-clockwise), exploitation index (always preferring the highest value option) or repetition index (always making the same choice at given gamble, no matter the direction or exploitation).

Fiber photometry analysis: All treatments and analyses were performed in Python using custom codes (mostly Numpy and Pandas libraries). After cleaning and processing each session signal to obtain dF/F values and z-scored dF/F values, events of interest were extracted to align the signal in [-3s:3s] time window in dataframes, t_0 being the exact time of location entry (triggering reward delivery or omission), with 1kHz sampling. Session-wise averages of given conditions for each mouse were then extracted, and averaged again over multiple mice for statistical analyses. In some conditions, especially when events of interest were rare (some scenarios of rewards or omissions chains in complexity, or some scenarios of locations transition in probabilistic), two or more sessions from one animal were pooled as if they were one (for instance, the last two sessions in a given context) to have enough trials for each animal in this

823 condition. For the same reason, the third cohort of $GRAB_{DA}$ mice followed 10 min long sessions (instead
824 of 5 min) in Cplx and Proba settings, with no particular effect on the overall quality of the signal, nor the
825 duration of $GRAB_{DA}$ expression (up to 3 months). For $GRAB_{DA}$, rewards-elicited positive transients
826 typically peaked around 250 ms after location entry (duration of ICSS being 200 ms) and decayed during
827 a bit less than 1s: we therefore extracted maximum and mean of the signal in a 1 s window post location
828 entry. Omissions-elicited negative transients were longer, reaching their minimum around 800 ms after
829 location entry and taking roughly 700-800 ms to go back to baseline: we therefore extracted minimum and
830 mean of the signal in 1.5 s window post location entry. For GCaMP, kinetics depended on the sensor
831 used: peaks reached maximum value around 250-300 ms post location entry for GCaMP6f and 350-400
832 ms for GCaMP7c, while dips reached minimum value around the same time (900-1050 ms post location
833 entry) for both sensors. However, return to baseline after reward-induced peaks was much shorter for
834 GCaMP6f (500-600 ms post location entry) than for GCaMP7c (2-3 s). For some correlation analyses
835 (using SciKit Learn Python library), especially the ones regarding z-scored peaks or dips amplitude
836 regarding outcome chain history, all trials of all mice were pooled together in a given condition.
837
838 **Generalised Linear Model (GLM) approach:** GLM was performed in Python using custom codes
839 (StatsModels or SciKit Learn library). To disentangle multiple factors that could explain DA signal, due to
840 high degree of behavioral and task-related variables correlated to each other from one trial to the next,
841 we designed a generalized linear model where a variable Y is explained by a linear combination of
842 multiples variables X_i, each of them weighted by a parameter w_i, plus a residual (or intercept) w_0.
843 $$Y = w_0 + w_1 . X_1 + w_2 . X_2 + \ldots$$
844 The model aims at fitting variations of Y by determining the weights w_i and their significance. Dependent
845 variable Y was post location entry 1s average for reward-induced peaks or 1.5s average for omission-
846 induced dips. Multiple X_i variables have been used, namely: *i)* reward or omission at previous and current
847 location, *ii)* Forward or U-turn at previous trial, *iii)* current target visited (spatially A, B or C, or in Proba
848 p_{100}, p_{50} or p_{25}), and *iv)* time since last stimulation (in Restcage stimulation condition). A single GLM was
849 applied for each mouse in a given condition, then w_i parameters resulting from all those GLMs were
850 averaged among mice, and the average was statistically compared to 0. Significance, either with positive
851 or negative weight, indicates that this variable explains part of DA variations.
852
853 **Reinforcement Learning (RL) models:** We used Reinforcement Learning (RL) to compute Reward
854 Prediction Errors (RPEs) from actual mice choice sequences and see how they match DA data. Before
855 each trial, the agent contains a set of expected values for each possible action. As one of these actions
856 is selected, it leads to either a reward ($V_{obtained}$ = 1) or an omission ($V_{obtained}$ = 0), then RPE is calculated

as $V_{obtained}$ - $V_{expected}$, and a new expected value of this action is fed back into the agent's set for next trials. From both behavioural and photometry results, we hypothesised and tested three possible value representations in the bandit task. First, we proposed a simple, one-order representation "going to any target" or "performing any trial" to get a reward. In this case, all trials are similar, regardless of target or trajectory choices, and we simply compute and update $V_{expected}$ = { V_{Any} } at each trial. Second, a representation of internal directionality with a set of two actions and $V_{expected}$ = { V_{Fwd} ; V_{Uturn} }. In this case, RPEs are specific and computed separately for each of the two actions. Third, a spatial representation "going to target X" with a set of three actions and $V_{expected}$ = { V_{pA} ; V_{pB} ; V_{pC} }. Again, RPEs are computed for each target independently. Modelling the RPE values resulting from each of those three representations allowed us to compare them and determine which simulation better replicates DA data in each context. Initial $V_{expected}$ were set consistently with behavior in the task. For Det End, they were all set to 0.99. For both Cplx End and Proba End, they were set as mean success rate computed from the two previous sessions. For example, for a given mouse, initial V_{Uturn} to initiate the RL model with choice sequence from sessions 9-10 is the proportion of rewarded Uturn trials from sessions 7-8. Exception is for V_{p100} in Proba End where it was also set to 0.99. We arbitrarily tested several learning rates α = {0.001; 0.01; 0.05; 0.2; 0.4}. Results were consistent with experimental data for α = {0.01; 0.05; 0.2; 0.4}. Smaller α (0.001) led to convergence that was too slow considering mice number of trials provided to models, while larger α made convergence in Det too quick. In Fig 3 and Fig S6, α is set to 0.05. We next assumed that in our recordings, DA = $V_{obtained}$ + RPE, and tested which representation accounted most in the error component using GLMs on top of our RL-computed RPEs (taking as input variables $V_{obtained}$ = {1; 0} for rewards or omissions, and theoretical RPEs computed from Model 1, 2 and 3). Similarly, models were applied for each mouse in a given context, then w_i parameters were averaged among mice for each context, and the average was statistically compared to 0. Significant weight indicates that this variable explains part of DA variations. Finally, we extended this compilation of RL-computed RPE values and GLM to fit RPE weights to DA data across sessions and contexts (Fig 4 and Fig S7). In this case, we started RL models with mice choice sequences in Det Start with all $V_{expected}$ equal to zero (naive agents), computed corresponding RPEs and updated corresponding $V_{expected}$. Consistent with mice progressively learning and updating values across sessions and contexts, the final $V_{expected}$ of a given time-point became the initial $V_{expected}$ of the next time point. For instance, from Det Start to Det Late (all $V_{expected}$ becoming closer to 1, but not at the same speed). Or from Cplx End to Proba Start ($V_{expected}$ of each target therefore starting to diverge). To allow for longitudinal comparisons, we next scaled (z-score) our data (both experimental DA and RL models-computed RPEs) on each time point, applied GLMs on each time point, and then compared the weights *i)* across sessions in a given context and at each transition between contexts, and *ii)* each of them regarding its difference with 0.

DEEP REINFORCEMENT LEARNING MODEL:

Three-point task simulation: We trained a two-layer deep neural network to predict the values of different actions inside a simulation of the three-point task. The simulation was designed trial-wise. In each trial, the model was located at one of the three points and had to choose an action to go to any of the two available points. We provided the model a set of 5 minimal actions, each corresponding respectively to go right, go left, and go to point A, B or C. After the network predicting the value of these 5 actions, a softmax function was used to convert the output value vector *y* into probabilities of choice using the following formula:

$$p(choice) = \frac{e^{\beta y}}{\sum e^{\beta y}} \quad (1)$$

Here β was a choice temperature constant, always set to 9 in our simulations to closely fit mouse data, although other reasonable β values produced qualitatively similar results. The chosen action was then sampled using the probabilities. Note that at each trial, one of the go to A, B or C actions, was made unavailable due to the model being already present at this location. It was thus excluded from the softmax computation. As we gave the model abilities to use location information in the available actions (with A, B, C) we also provided information about its current location in the input. We also gave the model an input that would always be equal to one regardless of the position, to facilitate the non-use of location information if they were not necessary to predict the values. Thus, the three possible input vectors were the following:

$$A : [1\ 0\ 0\ 1], \quad B : [0\ 1\ 0\ 1], \quad C : [0\ 0\ 1\ 1] \quad (2)$$

where each corresponds to the location of the model on point A, B or C.

Architecture and initialization: The vector of actions values *y* is given by the equation:

$$y = XW^1W^2 \quad (3)$$

with *X* being the vector of inputs as defined in (2), W^1 the weights to compute the hidden layer vector from the input vector, and W^2 the weights to compute the output vector y from the hidden layer vector. We opted for a network with 4 hidden units (same as the number of inputs), even though a different number of hidden units produced similar results (tested with 1 and 10 hidden units). For each simulation, the initialization of $W^1$and W^2 were sampled from a uniform distribution in the interval [-0.01; 0.01] such that initial predicted values from y output vector were close to zero.

$$W^1_{init} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \quad \mathcal{U}(-0.01,\, 0.01) \tag{4}$$

$$W^2_{init} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} + \quad \mathcal{U}(-0.01,\, 0.01) \tag{5}$$

Training and Reinforcement: We trained each simulation for 400 trials in each rule set of the three-point task in the same order as the mice: first determinist, then complexity, and finally probabilistic. We updated the weights of the network using stochastic gradient descent at the end of each trial. At each rule switch, the weights of the network were preserved. Each update minimized the following loss function:

$$L = (y_{ch} - Stim)^2 \tag{6}$$

where y_{ch} is the value of the chosen action and Stim. is the result of the choice, taking the value 1 if the model receives a "stimulation" and 0 if not. From this loss, the gradients to update $W^1$and W^2 were computed using chain rule as follow:

$$\Delta W^1 = \alpha * 2 * X^T (y_{ch} - Stim) W^{2T}_{:,ch} \tag{7}$$

$$\Delta W^2_{:,ch} = \alpha * 2 * (XW^1)^T (y_{ch} - Stim) \tag{8}$$

where α is the learning rate and was set to 0.02. Note that the gradient for W^1were computed by only back-propagating information relative to the computation of the chosen action value, which correspond to the weights contained in only one column of the W^2 matrix. Similarly, the gradient for W^2 was a one column vector that would update the weights of the only column of the W^2 matrix that is necessary to compute the value of the chosen action. Because of this design, the actions were in competition for learning, which could favor the use of different actions depending on the actual rule set.

Figures and Statistics: Raw figures were plotted using Python custom codes (mostly MatPlotLib library). Graphics, typography and layout were formatted with Adobe Illustrator. All statistical analyses were computed using Python with Scipy library and custom programs. Results were most frequently plotted as individual data points and mean ± sem. The total number of observations in each group and the statistics used are indicated in figure legends and detailed statistics tables: unless specified, data points indicate the number of mice (N) on which the statistics were performed, and in some cases, they represent number

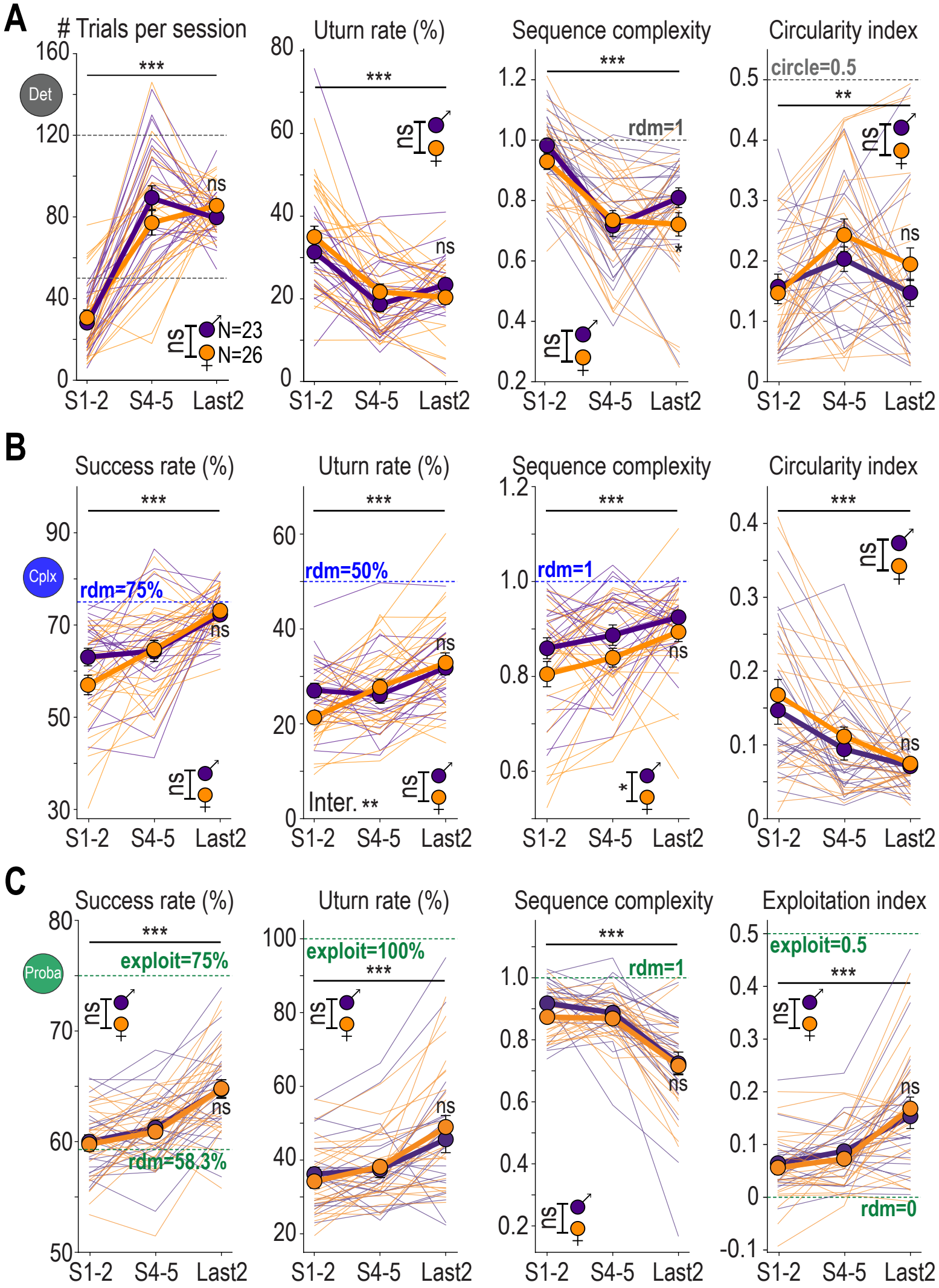
948 of trials (n) either for one example session from one animal, or from all sessions of all animals in a given
949 condition. Classical comparisons between means were performed using parametric tests (Student's t-
950 test, or ANOVA for comparing more than two groups, when parameters followed a normal distribution
951 (Shapiro test P > 0.05)), and non-parametric tests when the distribution was skewed (here, Wilcoxon or
952 Mann-Whitney U for one/two samples and whether comparison is paired or not, or Kruskall-Wallis for
953 more than two groups). More complex comparisons with several factors were performed using two-way
954 or mixed ANOVA regardless of normal distribution for simplicity, with no major impact on results
955 interpretation (see Fig S1, sex X session effects). Multiple comparisons were corrected using a
956 sequentially rejective multiple test procedure (Holm). Linear regressions were assessed either with
957 Pearson (parametric) or Spearman (non-parametric) tests. Probability distributions were compared using
958 the Kolmogorov–Smirnov (KS) test. All statistical tests were two-sided. p > 0.05 was considered not to be
959 statistically significant. In some cases, p > but close to 0.05 were indicated in the figure (see Tables of
960 detailed statistics for more information).
961
962 **Fluorescence immunohistochemistry:** After completing the successive rules of the task, mice from the
963 $GRAB_{DA}$ cohorts were euthanatized by IP injection of euthasol (0.1mL per 30g at 150mg/kg), immediately
964 followed by paraformaldehyde (PFA) intra-cardiac perfusion, and brains were rapidly removed and post-
965 fixed in 4% PFA for 2 to 4 days. Serial 60µm sections were cut with a vibratome (Leica).
966 Immunohistochemistry was performed as follows: free-floating VTA and NAc brain sections were
967 incubated for 1h at 4°C in a blocking solution of phosphate-buffered saline (PBS) containing 3% bovine
968 serum albumin (BSA, Sigma A4503) and 0.2% Triton X-100, and then incubated overnight at 4 °C with *i)*
969 a mouse anti-tyrosine hydroxylase primary antibody (TH, Sigma, T1299) at 1:500 dilution and *ii)* a chicken
970 anti-eYFP primary antibody (Life technologies Molecular Probes, A- 6455) at 1:1000 dilution, both in PBS
971 containing 1.5% BSA and 0.2% Triton X-100. The following day, sections were rinsed with PBS and then
972 incubated for 3 h at 22–25 °C with *i)* Cy3-conjugated anti-mouse secondary antibody (Jackson
973 ImmunoResearch, 715-165-150) at 1:500 dilution and *ii)* a goat anti-chicken AlexaFluor 488 secondary
974 antibody (711-225-152, Jackson ImmunoResearch) at 1:1000 dilution, both in a solution of 1.5% BSA and
975 0.2% Triton X-100 in PBS. After three rinses in PBS, slices were wet-mounted using Prolong Gold Antifade
976 Reagent with DAPI (Invitrogen, P36930). Microscopy was carried out with a fluorescent microscope Leica
977 DMR, and images captured in gray level using MetaView software (Universal Imaging Corporation) and
978 colored post-acquisition with ImageJ. Labeling for YFP in the NAc (along with satisfying signal during the
979 task) allowed to confirm $GRAB_{DA}$ expression, and fiber implantation in the NAc lateral shell was also
980 visually checked. Similar procedures were used to check for GCaMP7c and GCaMP6f expression in VTA
981 DA neurons. For GCaMP7c we used the same anti-TH and anti-eYFP antibodies as previously described.

982 For GCaMP6f we used a sheep anti-TH primary antibody (AB-1542, Milipore) at 1:500 dilution coupled
983 with a donkey anti-sheep secondary antibody (713-165-147, Jackson ImmunoResearch) at 1:500 dilution
984 to highlight DA neurons, and simply used the virus-associated tdTomato to validate expression in the VTA
985 and optic fiber implantation site. For MFB slices, 100 µm sections were performed and slices were directly
986 visualized with visible light to check for ICSS electrode implantations.
987
988 **Statistics and Reproducibility:** All experiments were replicated with success (several successive
989 cohorts of mice).
990

991 **Supplementary figures**

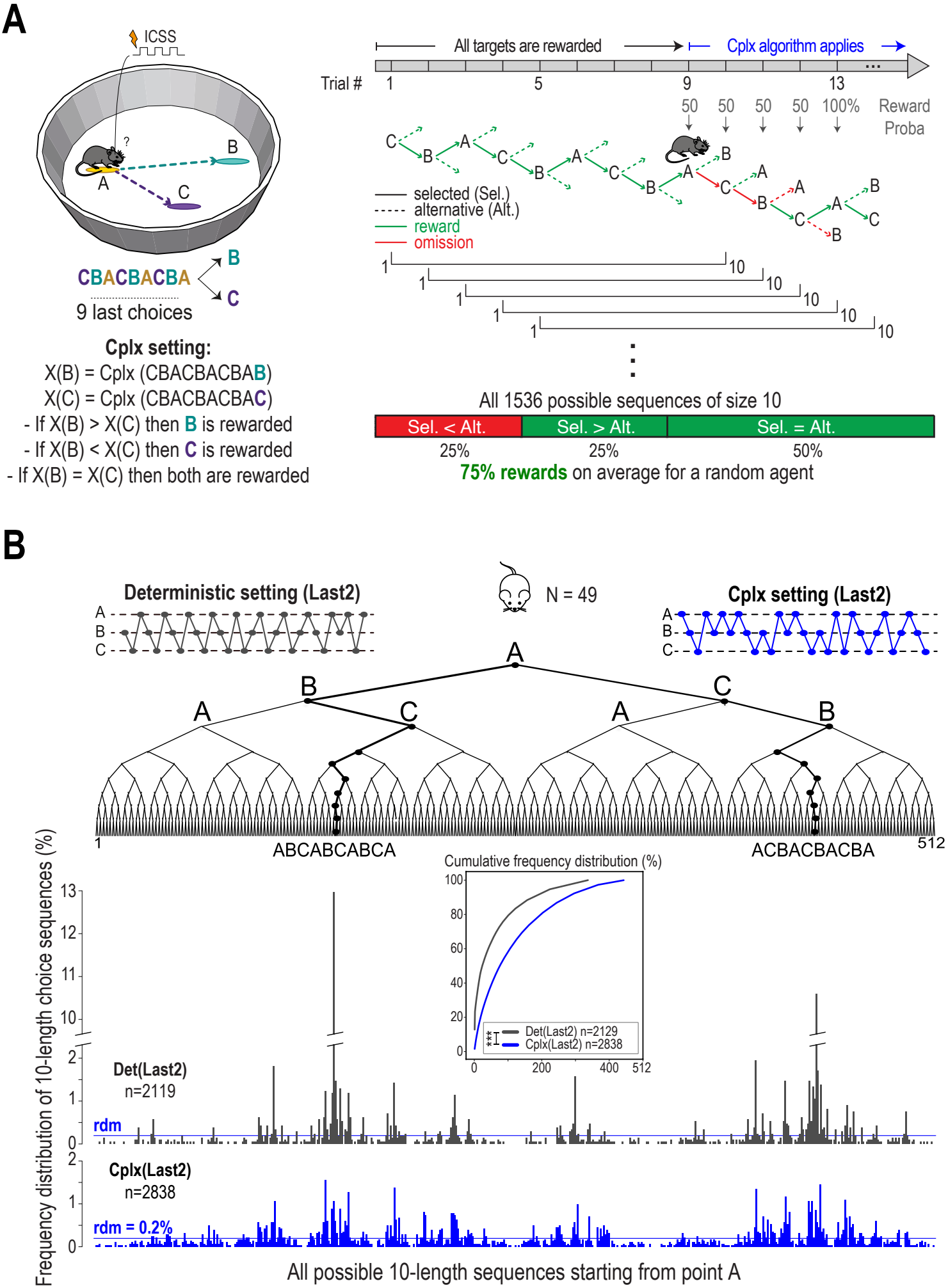
992

993



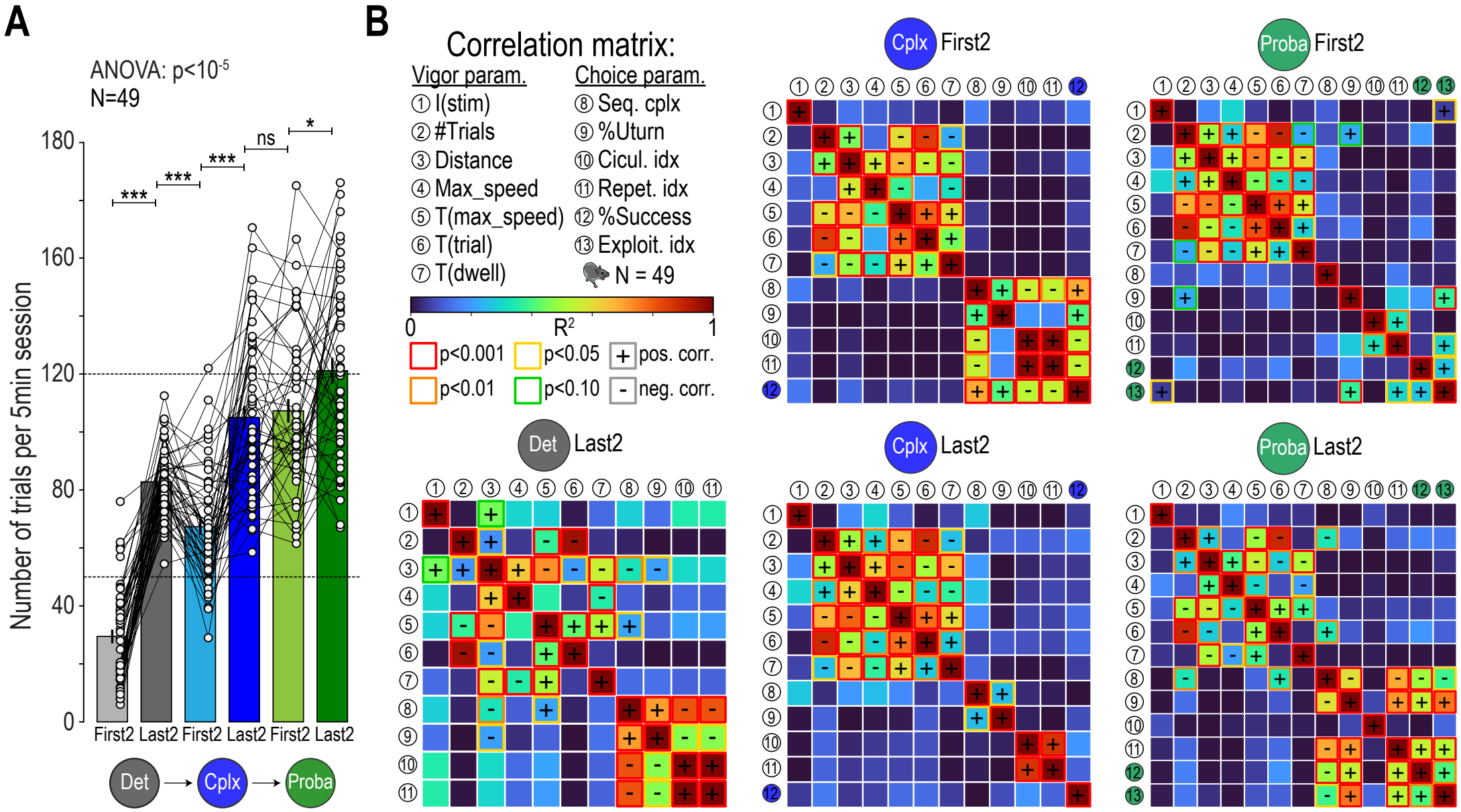
Supplementary figure 1

994 **Fig. S1: Evolution of decision behaviour across sessions, with no major sex effects.**
995 **A. Decision parameters throughout Det sessions for males and females.** Comparison of **(left)** the
996 number of trials per session, **(middle-left)** the Uturn rate, **(middle-right)** the sequence complexity, and
997 **(right)** the circularity index between sessions 1&2, sessions 4&5 and the last 2 sessions in male and
998 female mice. In addition, we also compared the final states (Last2) between males and females. A fully
999 circular mouse would have 0% Uturn, low seq. cplx and 0.5 circul. idx. **B. Same as in A. for Cplx**
1000 **sessions.** A mouse keeping circular strategy would have low success, 0% Uturn, low seq. cplx and 0.5
1001 circul. idx. A random mouse would have 75% success, 50% Uturn, seq cplx = 1 and circul. idx = 0. **C.**
1002 **Same as in A. for Proba sessions.** An exploitative mouse would have 75% success, 100% Uturn, low
1003 seq. cplx and 0.5 exploit. idx. A random mouse would have 58.3% success, 50% Uturn, seq cplx = 1 and
1004 exploit. idx = 0. (Data are shown as individual points, and mean ±sem. N = 23 male and 26 female mice.)
1005
1006



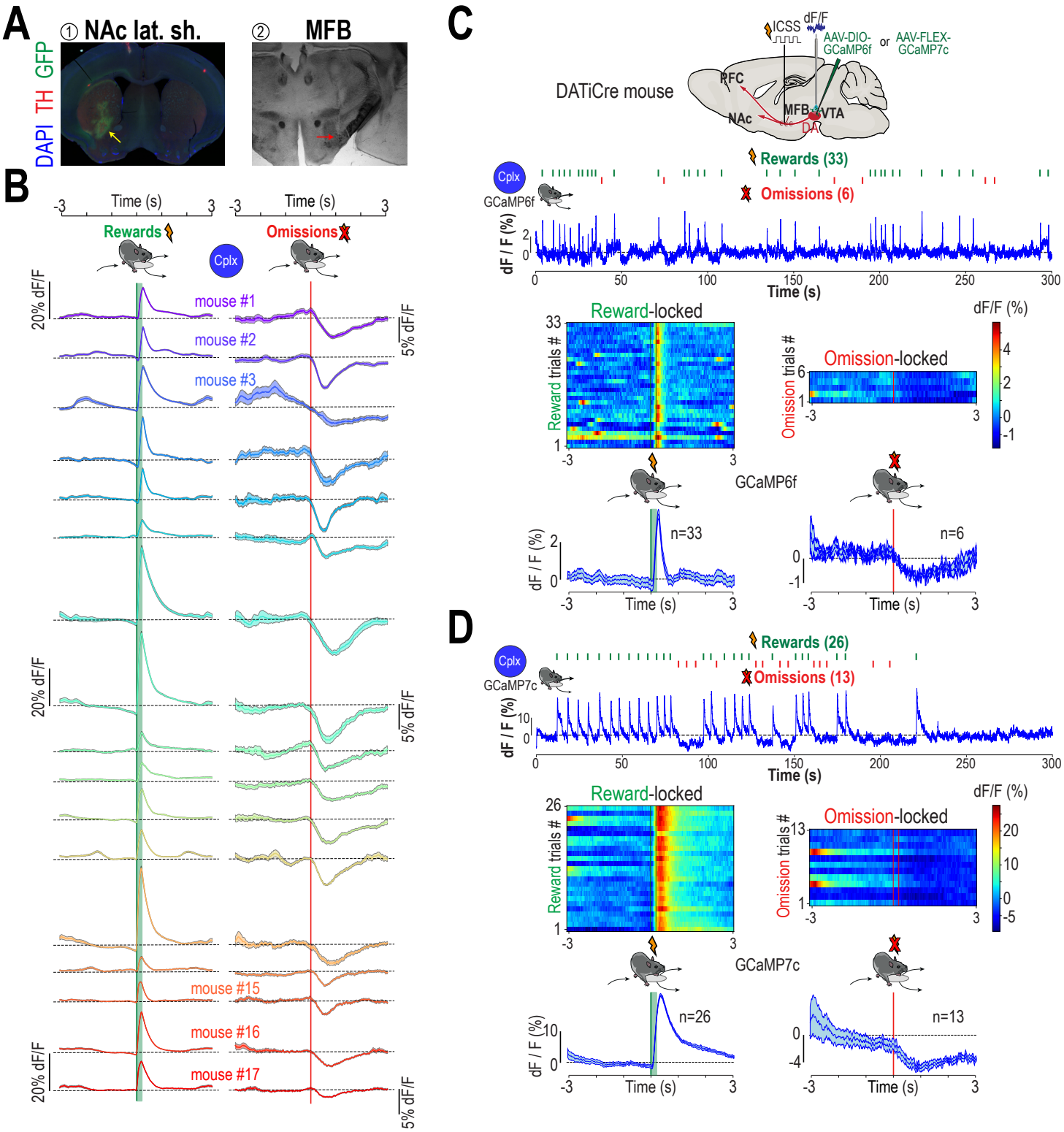
Supplementary figure 2

1007 **Fig. S2: Additional information on the Cplx rule and mice sequence patterns.**
1008 **A. Detailed schematic representation of the Cplx rule.** The first 9 trials of each session provide
1009 deterministic rewards (P=100%) to launch the Cplx algorithm, which then determines at each trial, in a
1010 sliding window, which target will lead to a reward by comparing the Lempel-Ziv grammatical complexity
1011 of the two potential sequences: 9 past choices + first remaining target VS. 9 past choices + second
1012 remaining target. The mouse will be rewarded only if it chooses the target that increases complexity. If
1013 both sequences have the same complexity, both targets will be rewarded **(see Methods)**. Taking all
1014 possible sequences of size 10 starting from one location, 75% of them are rewarded on the 10th trial.
1015 Therefore, a random agent exploring homogeneously this sequences tree will converge to 75% success
1016 rate. **B. Distribution of mice choice sequences of length 10 at the end of Det and Cplx.** Two
1017 distribution peaks (paths in the decision tree) appear in Det, corresponding to circling behavior (clockwise
1018 and counterclockwise), representing together roughly 25% of all produced sequences (among 512
1019 possibilities). In Cplx, these peaks strongly reduce in size, in favor of more distributed visits of all possible
1020 sequences. **(Insert)** Cumulative distribution comparison between Det and Cplx (Last2 sessions for each
1021 rule). (In B, n is the total number of sequences of length 10, computed from sessions-wise mice
1022 successive choices, from N=49 mice both males and females).
1023
1024



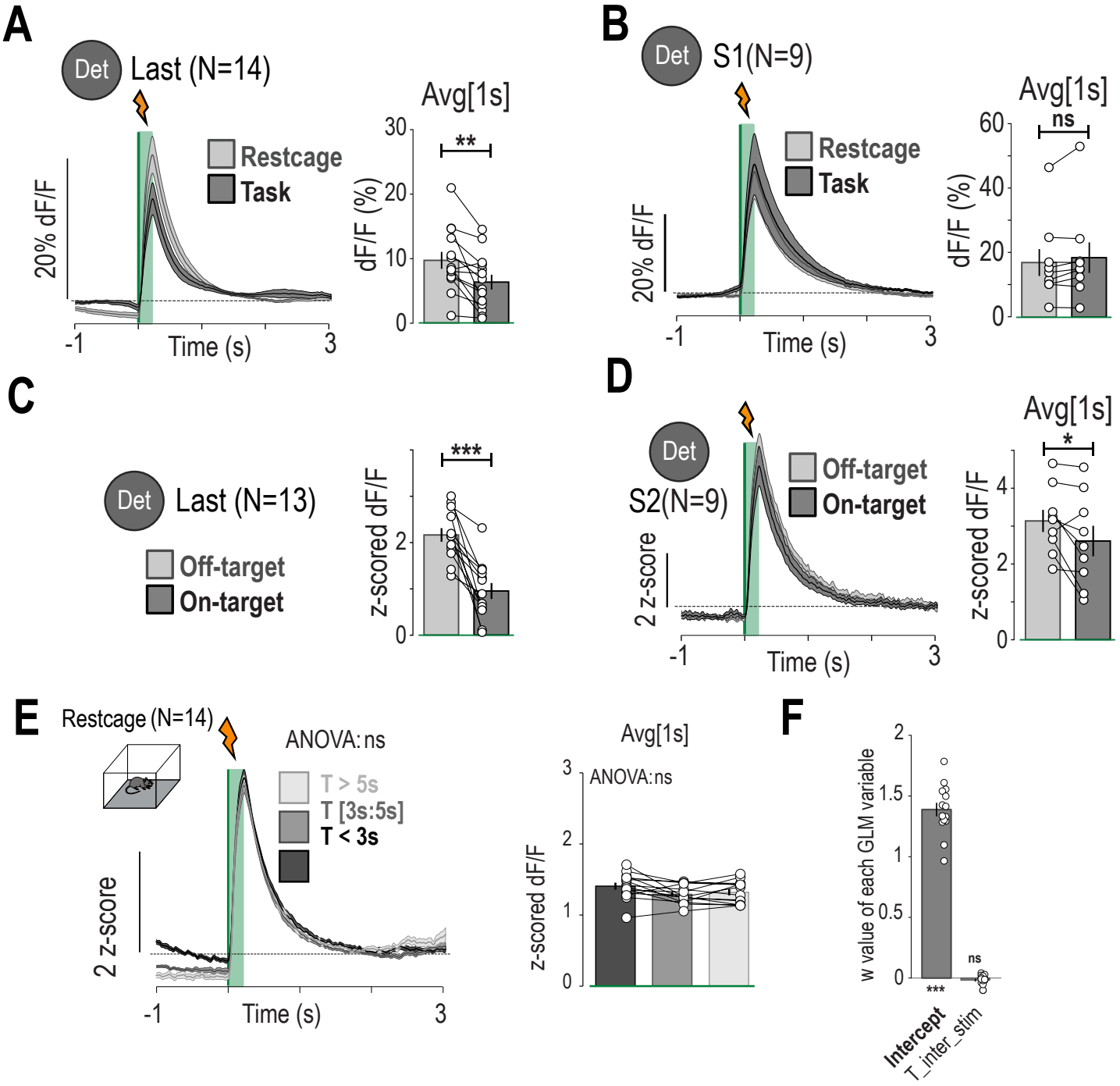
Supplementary figure 3

Fig. S3: Motivation throughout the task and decoupling between vigor and choice parameters. **A. Evolution of motivation across contexts and sessions.** Comparison of number of trials across contexts and sessions (mean of 2 sessions each time). **B. Correlation matrices between various vigor and choice parameters across mice in different contexts.** Parameters are computed for each mouse as the mean of 2 sessions (either First2 or Last2, for a given context). Each box represents the linear correlation between two parameters (Pearson for parametric, Spearman for non-parametric, each dot being a mouse). The filling color of each box represents the R_s value. The frame color of each box represents the p-value (after Bonferroni correction). The warmer the color, the more those two parameters are significantly correlated. **(Left)** Last2 sessions of Det (11 parameters, x66 Bonferonni correction). **(Middle)** First2 and Last2 sessions of Cplx (12 parameters, x78 Bonferonni correction). **(Right)** First2 and Last2 sessions of Proba (13 parameters, x91 Bonferonni correction). (In A, data are shown as individual points, and mean ±sem. In B, only R_s and corrected p-values are shown with color code. Individual data are available upon request. N is always the number of mice.)



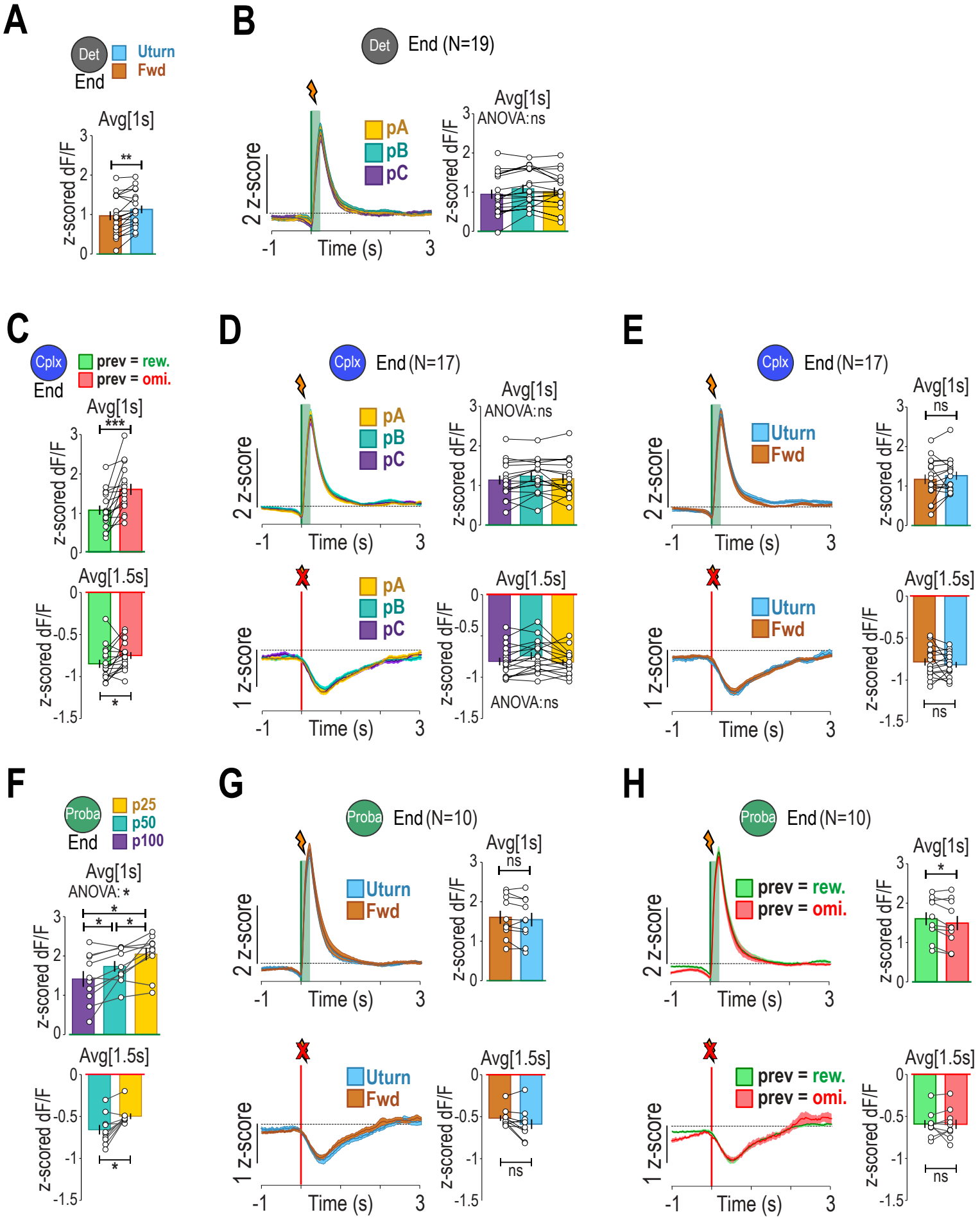
Supplementary figure 4

1040 **Fig. S4: DA fiber photometry signals with $Grab_{DA}$ and GCaMP6f A. NAc and MFB slices**
1041 **immunohistochemistry.** Post-hoc verification of optic fiber implant and $Grab_{DA}$ virus expression in the
1042 NAc lateral shell (left), and stimulation electrode implant in the MFB (right). **B. Individual mice NAc DA**
1043 **release for rewards and omissions in Cplx.** Each line and colour is an individual mouse, averaged for
1044 all trials during last Cplx session, in [-3s:3s] time window locked on location entry. Every single mouse
1045 included in the results displayed reward-induced peaks and omission-induced dips of DA release
1046 significantly different from zero (dashed black lines). **C-D. DA cell activity using GCaMP fiber**
1047 **photometry.** DATiCre mice were injected with an AAV to express either GCaMP6f or GCaMP7c in VTA
1048 DA neurons, implanted with an optic fiber in the VTA, and stimulation electrode in the MFB, to assess DA
1049 neuron activity in the task. **C. GCaMP6f.** Using similar experimental procedures and signal analyses in
1050 the Cplx context, calcium dynamics of VTA DA neurons show similar reward and omission-induced
1051 transients than NAc lateral shell DA release, in this case with smaller transients amplitudes (worse signal-
1052 to-noise-ratio) whether positive or negative, and faster kinetics for positive transients. **D. GCaMP7c.** Same
1053 as B for GCaMP7c, with greater transient amplitudes (better signal-to-noise-ratio) and slower kinetic,
1054 whether positive or negative. (In B, C, D, curves are shown as mean ±sem for a single session, n is the
1055 number of reward or omission trials in this session.)
1056
1057



Supplementary figure 5

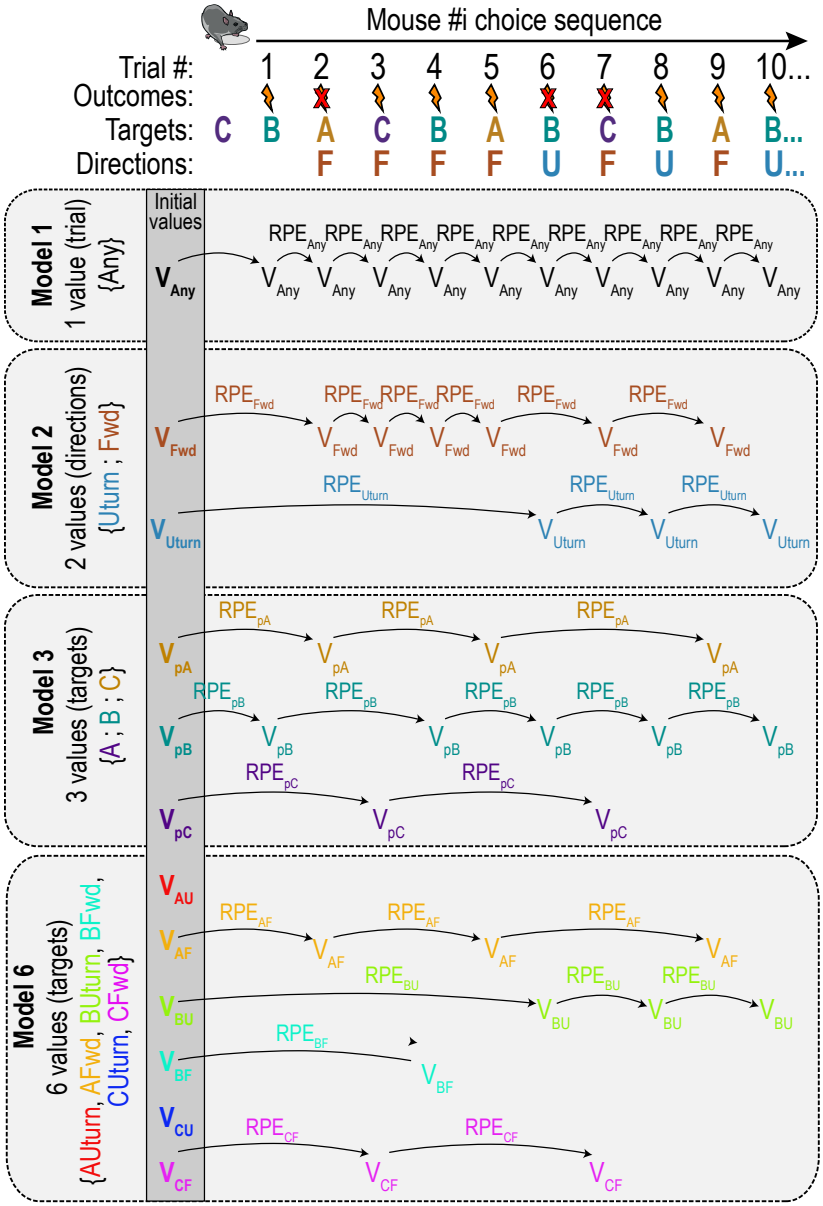
1058 **Fig. S5: DA fiber photometry signals in various configurations. A.** DA response to expected (Task)
1059 vs unexpected (Restcage) ICSS (same current intensity) in Det Last session. **B.** Same as A but during
1060 first session (S1) of conditioning in Det. **C.** DA response to expected (On-target) vs unexpected (Off-
1061 target) ICSS (same current intensity) in Det Last session. **D.** Same as C but during second session (S2)
1062 of conditioning in Det. **E.** Comparison of DA response for short (<3s), mid ([3s:5s]) and long (>5s) intervals
1063 of stimulation in restcage. Curves are shown as mean ±sem for session-wise average of several mice, N
1064 is the number of mice in each condition. Bar plots are shown as mean ±sem, in addition to individual data
1065 points. **F.** GLM of the amplitude of DA response as a function of interval of stimulation in the restcage, the
1066 same day as Last Det. Weight value of each variable compared to zero.
1067
1068



Supplementary figure 6

1069 **Fig. S6: Additional analyses of NAc DA release regarding task behavioral features. A.** Individual
1070 data corresponding to transient peak of Fig 4E. Comparison between Fwd and Uturn. **B.** Target effect on
1071 DA transients in Det End. Comparison between pA, pB and pC. **C.** Previous outcome effect on DA
1072 transients in Cplx End. Individual data corresponding to Fig 4F. **(Top)** Reward-induced transients
1073 comparison between previous reward and previous omission. **(Bottom)** Omission-induced transients
1074 comparison between previous reward and previous omission. **D.** Target effect on DA transients in Cplx
1075 End. (Top) Reward-induced transients comparison between pA, pB and pC. **(Bottom)** Omission-induced
1076 transients comparison between pA, pB and pC. **E.** Trajectory effect on DA transients in Cplx End. **(Top)**
1077 Reward-induced transients comparison between Uturn and Fwd. **(Bottom)** Omission-induced transients
1078 comparison between Uturn and Fwd. **F.** Target effect on DA transients in Proba End. Individual data
1079 corresponding to Fig 4G **(Top)** Reward-induced transients comparison between p100, p50 and p25.
1080 **(Bottom)** Omission-induced transients comparison between p50 and p25. **G.** Trajectory effect on DA
1081 transients in Proba End. **(Top)** Reward-induced transients comparison between Uturn and Fwd. **(Bottom)**
1082 Omission-induced transients comparison between Uturn and Fwd. **H.** Previous outcome effect on DA
1083 transients in Proba End. **(Top)** Reward-induced transients comparison between previous reward and
1084 previous omission. **(Bottom)** Omission-induced transients comparison between previous reward and
1085 previous omission.
1086 (Bar plots are shown as mean ±sem, in addition to individual data points. Curves are shown as mean
1087 ±sem. N is always the number of mice in each context.)
1088
1089

A

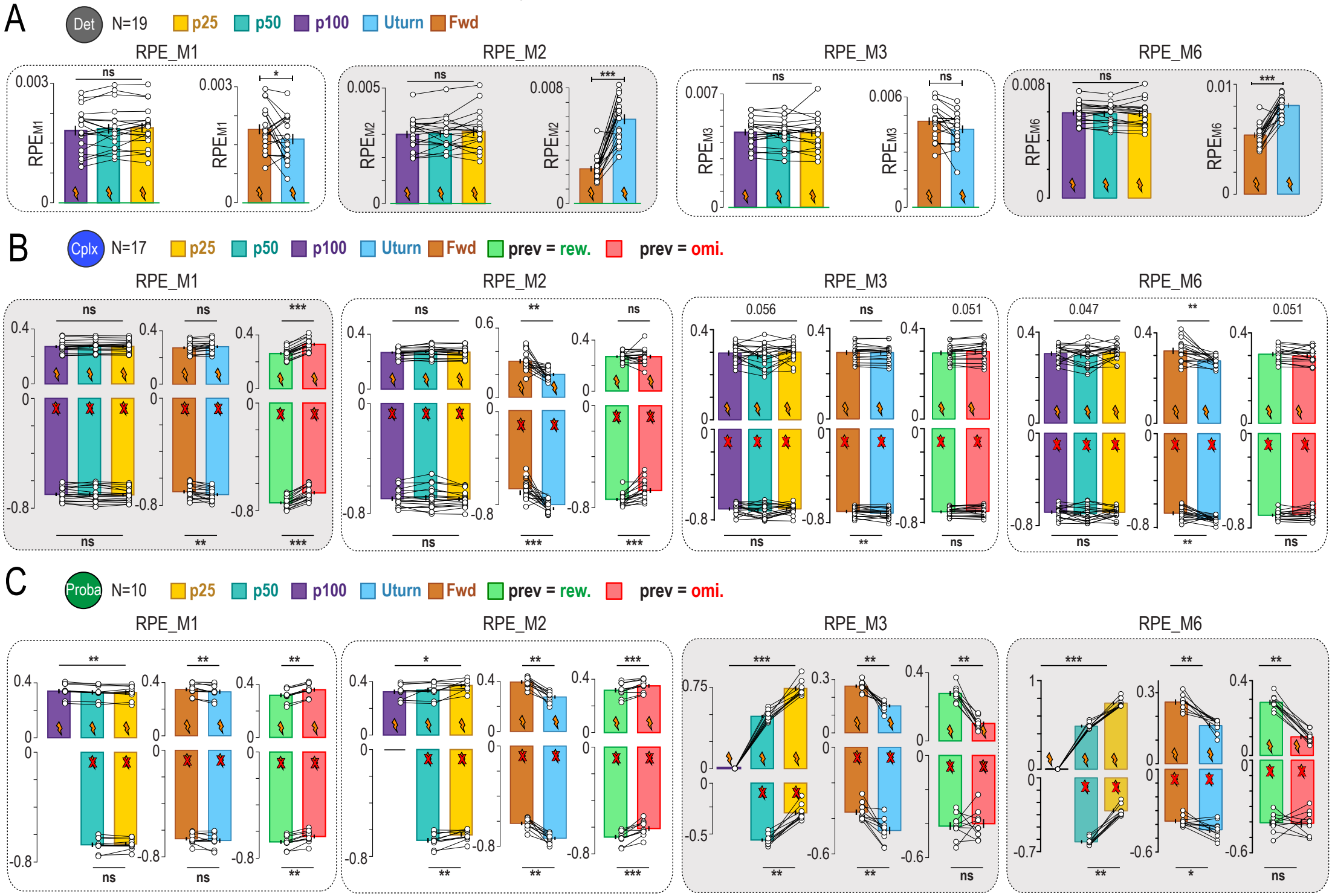


Supplementary Figure 7

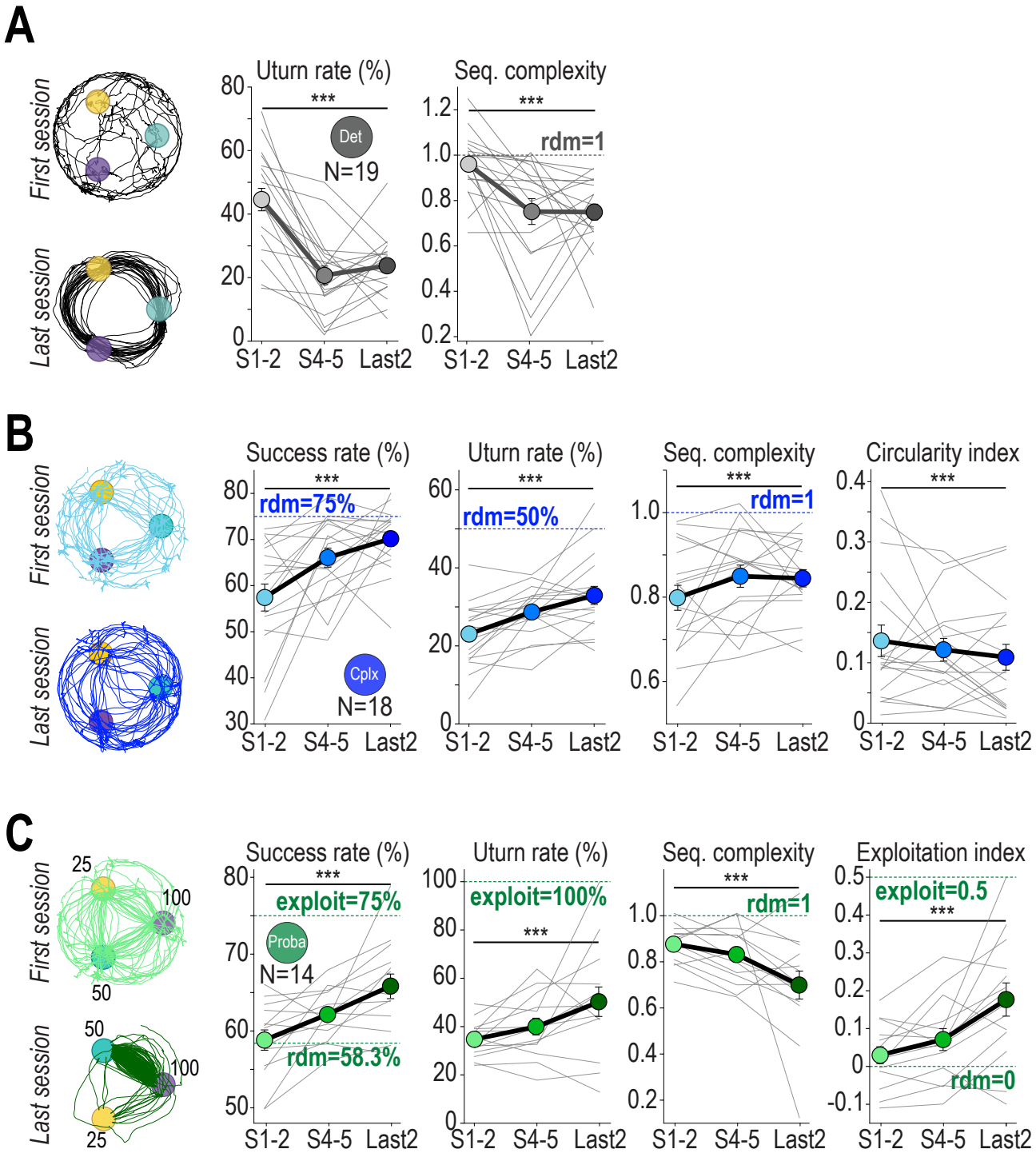
1090 **Fig. S7: Detailed schematic of RL modelling for each of the four models.** From actual mice choice
1091 sequences, we applied RL models and computed corresponding RPEs. The first model consists in single
1092 value representation "going to any target" or "performing any trial" to get a reward, where we simply
1093 compute $V_{expected}$ = { V_{Any} } and RPE_{Any} at each trial. The second model consists in two value
1094 representations depending on chosen trajectory $V_{expected}$ = { V_{Fwd} ; V_{Uturn} }. In this case, RPE_{Uturn} and
1095 RPE_{Fwd} are specific and computed separately for each of those two actions. The third model consists in
1096 three value representations depending on chosen target $V_{expected}$ = { V_{pA} ; V_{pB} ; V_{pC} }. Again, RPE_{pA}, RPE_{pB}
1097 and RPE_{pC} are computed for each target independently. The fourth model consist in six value considering
1098 forward and Uturn for each position A, B or C $V_{expected}$ = { V_{AF} ; V_{AU} ; V_{BF} ; V_{BU} ; V_{CF} ; V_{CU} }. Again, RPE_{AF},
1099 RPE_{AU}, RPE_{BF}, RPE_{BU} RPE_{CF}, and RPE_{CU} are computed for each target independently.
1100
1101

Supplementary figure 8

Avg RPE modelled from mice choice seq.



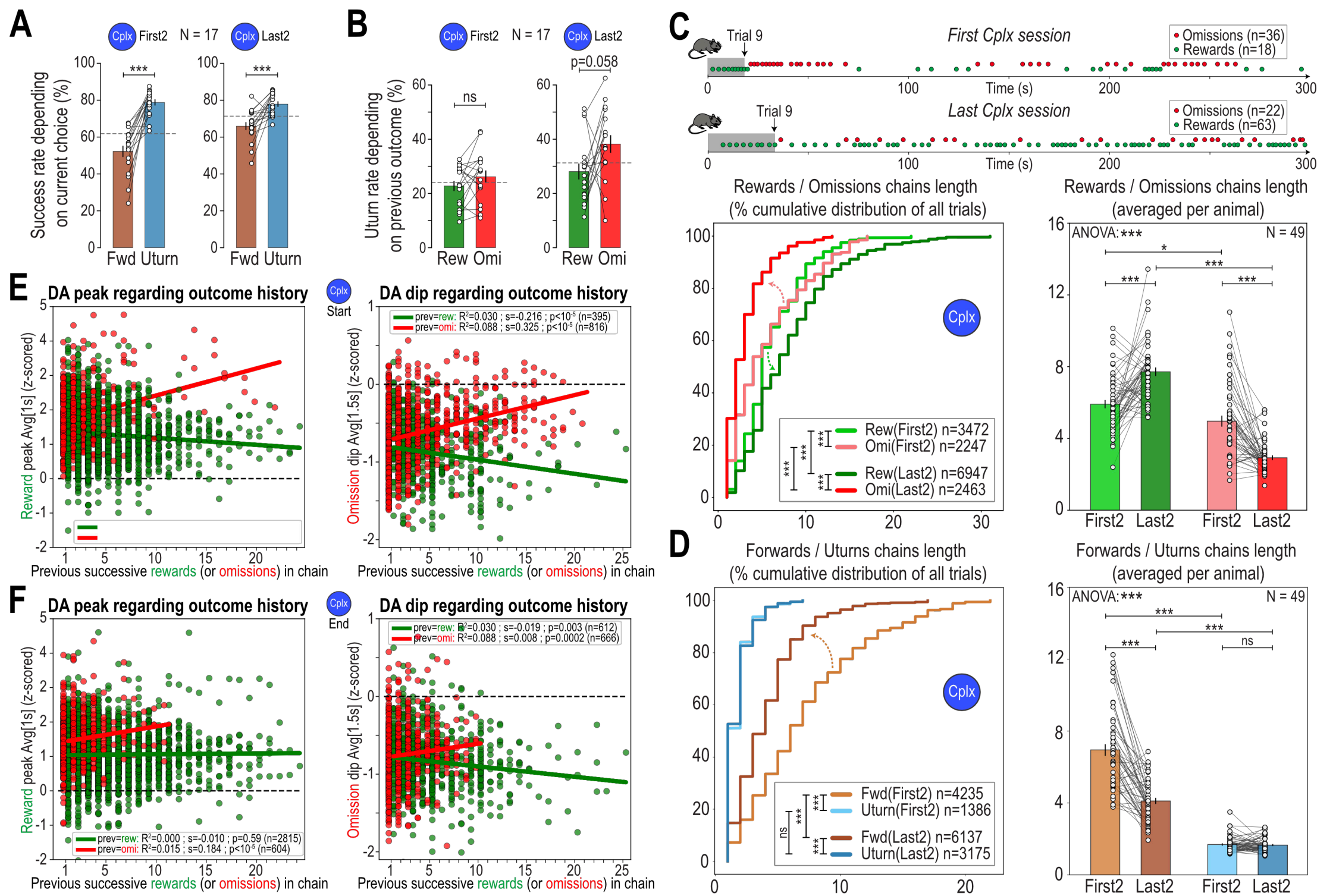
1102 **Fig. S8: Additional information on the four RL models, comparison of computed RPEs in various**
1103 **behavioral scenarios, and results in Proba Change context.**
1104 **A-C. For each model (M_1 to M_6), computed RPEs were averaged over mice sessions in the same**
1105 **scenarios used to characterize DA responses (regarding target, trajectory, and previous outcome).**
1106 The model that qualitatively reproduces best DA responses in all scenarios in given context is supposed
1107 to be the best value representation that mice are using in this context. **A. End Det context.** Average M1,
1108 M2, M3 and M6-computed RPE comparison between targets, and trajectories. **B. End Cplx context.**
1109 **From top to bottom:** Average M1, M2, M3 and M6-computed RPE comparison between targets,
1110 trajectories and previous outcome. **C. End Proba context. From top to bottom:** Average M1, M2, M3
1111 and M6-computed RPE comparison between targets, trajectories and previous outcome. Bar plots are
1112 shown as mean ± sem, in addition to individual data points. N is always the number of mice in each
1113 context. Some of this plots are also shown in Figure 5
1114
1115



Supplementary figure 9

Fig. S9: Evolution of choice parameters across rules and sessions in GRAB DA mice.

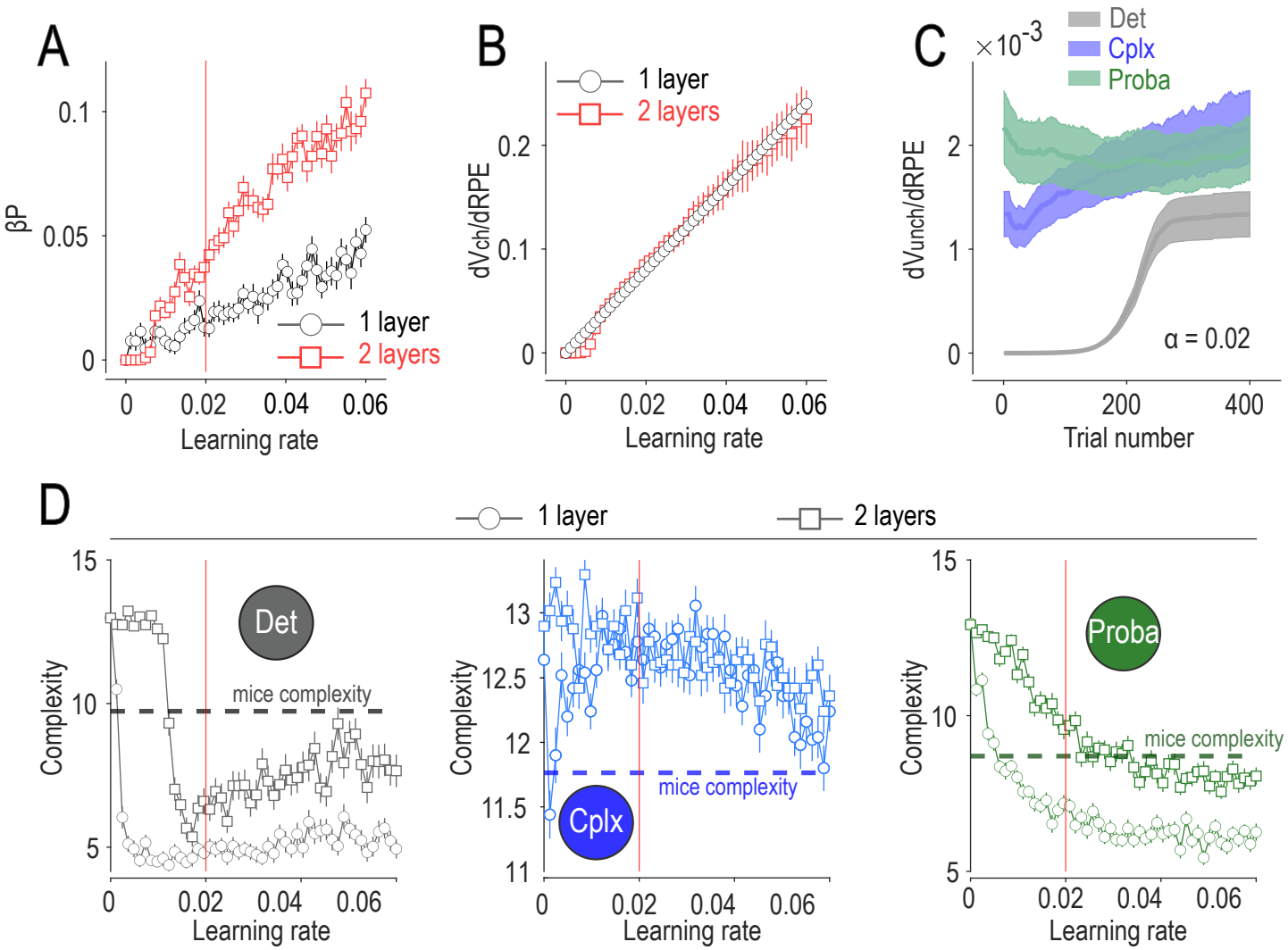
A. Evolution of choice parameters across Det sessions. Left: Example trajectories of first and last Det session. Comparison of **(middle)** Uturn rate and **(right)** sequence complexity between sessions 1&2, sessions 4&5 and last 2 sessions in Grab-DA mice. **B. Evolution of choice parameters across Cplx sessions. Left:** Example trajectories of first (cyan) and last (blue) Cplx sessions. Comparison of **(middle-left)** Success rate, **(middle)** Uturn rate, **(middle-right)** sequence complexity and **(right)** circularity index between sessions 1&2, sessions 4&5 and last 2 sessions in Grab-DA mice. **C. Evolution of choice parameters across Proba sessions. Left:** Example trajectories of first (light green) and last (dark green) Proba sessions. Comparison of **(middle-left)** Success rate, **(middle)** Uturn rate, **(middle-right)** sequence complexity and **(right)** exploitation index between sessions 1&2, sessions 4&5 and last 2 sessions in Grab-DA mice. (data are shown as mean ±sem, in addition to individual data points. In D, each data point is one animal at one time point. signal curves are shown as mean ±sem. In A, data are shown as mean ±sem for clarity. Individual data points are available upon requests. Due to multiple corrections generating dilutions in p-values, **#** symbol has been used in the figure to highlight $p<0.12$ after correction. N is always the number of mice in each context.)



Supplementary figure 10

Fig. S10: Additional analyses of DA transients and choice behavior in Cplx.

A. Success rate depending on current Uturn/Fwd choice in Cplx. Left: First 2 sessions. **Right:** Last 2 sessions. **B. Uturn rate depending on previous outcome in Cplx. Left:** First 2 sessions. **Right:** Last 2 sessions. **C. Analysis of chains of successive rewards and omissions in Cplx. Top:** In early Cplx, mice tend to keep repeating circular patterns and therefore get long series of omissions. In late Cplx, omissions are regularly distributed, generating smaller chains, as expected from a random agent. **Bottom-left:** Cumulative distribution of reward and omission chain lengths during first 2 and last 2 Cplx sessions. **Bottom-right:** Average chain lengths per mouse. **D. Same for chains of successive forwards and Uturns. Left:** Cumulative distribution of forward and Uturn chains length during first 2 and last 2 Cplx sessions. **Right:** Average chains length per mouse. For regressions in A, B, each dot is a trial of one mouse. In C, D, E, F, Bar plots are shown as mean±sem, in addition to individual data points. In E, F, cumulative distribution are computed for all trials of all mice together. n is always the number of trials, and N the number of mice, in each context.) **E. Linear regressions of DA transients depending on the number of successive previous rewards or omissions in chains in Cplx Start. Left:** Reward-induced DA peak amplitudes regarding length of successive previous rewards chains (green) or omissions chains (red). **Right:** Same for omission-induced DA dip amplitudes regarding length of successive previous rewards chains (green) or omissions chains (red). **F. Same for Cplx End. Left:** Reward-induced DA peak amplitudes regarding length of successive previous rewards chains (green) or omissions chains (red). **Right:** Same for omission-induced DA dips amplitude regarding length of successive previous rewards chains (green) or omissions chains (red).



Supplementary figure 11

Fig. S11: Property and performance comparison between 1-layer and 2-layers models

A. Sensitivity of model's RPE to previous outcome in Cplx (βP of the full linear model, mean±SEM) according to imposed learning rate (α parameter). The vertical red bar shows the learning rate used in the rest of the paper (α = 0.02). **B.** Derivate of Value at t+1 of the action chosen at t according to RPE at t, calculated for different learning rate (mean±SEM). This figure shows that the difference in between the 2 models in A is not explained by a difference in the amplitude of the update for a given RPE between the 2 models. **C.** Evolution of the derivate of Value at t+1 of the actions unchosen at t according to RPE at t, for the learning rate used in the paper (α = 0.02), for the 2 layers model. The derivate evolves to become strictly positive through training schedule, while this derivate is 0 by definition for the 1 layer model, thus providing explanation for the difference seen in A. **D.** Mean complexity of behavior of the 2 models, calculated on a sequence of 50 trials, accross the 3 rules. This measure provides an estimation of how stereotyped (repetition of the same choices) the behavior of each model is. The dotted line shows the same estimation on the actual mice data. The 1-layer model tends to produce very stereotyped behavior (in Det and Proba) while the 2-layers model is closer to the actual data.