Misclassification in memory modification in *App*$^{NL\text{-}G\text{-}F}$ knock-in mouse model of Alzheimer's disease

Mei-Lun Huang[1], Yusuke Suzuki[1,2,3*], Hiroki Sasaguri[4,5], Takashi Saito[6], Takaomi C Saido[4], Itaru Imayoshi[1,2,3*]

[1]Laboratory of Brain Development and Regeneration, Division of Systemic Life Science, Graduate School of Biostudies, Kyoto University, Kyoto, Japan

[2]Center for Living Systems Information Science, Graduate School of Biostudies, Kyoto University, Kyoto, Japan

[3]Laboratory of Deconstruction of Stem Cells, Institute for Life and Medical Sciences, Kyoto University, Kyoto, Japan

[4]Laboratory for Proteolytic Neuroscience, RIKEN Center for Brain Science, Saitama, Japan

[5]Dementia Pathophysiology Collaboration Unit, RIKEN Center for Brain Science, Saitama, Japan

[6]Department of Neurocognitive Science, Institute of Brain Science, Graduate School of Medical Sciences, Nagoya City University, Aichi, Japan

***Correspondence:**

Yusuke Suzuki

suzuki.yusuke.7n@kyoto-u.ac.jp

Itaru Imayoshi

imayoshi.itaru.2n@kyoto-u.ac.jp

Abstract

Alzheimer's disease (AD), the leading cause of dementia, could potentially be mitigated through early detection and interventions. However, it remains challenging to assess subtle cognitive changes in the early AD continuum. Computational modeling is a promising approach to explain a generative process underlying subtle behavioral changes with a number of putative variables. Nonetheless, internal models of the patient remain underexplored in AD. Determining the states of an internal model between measurable pathological states and behavioral phenotypes would advance explanations about the generative process in earlier disease stages beyond assessing behavior alone. In this study, we assumed the latent cause model as an internal model and estimated internal states defined by the model parameters being in conjunction with measurable behavioral phenotypes. The 6- and 12-month-old *App*$^{NL\text{-}G\text{-}F}$ knock-in AD model mice and the age-matched control mice underwent memory modification learning, which consisted of classical fear conditioning, extinction, and reinstatement. The results showed that *App*$^{NL\text{-}G\text{-}F}$ mice exhibited a lower extent of reinstatement of fear memory. Computational modeling revealed that the deficit in the *App*$^{NL\text{-}G\text{-}F}$ mice would be due to their internal states being biased toward overgeneralization or overdifferentiation of their observations, and consequently the competing memories were not retained. This deficit was replicated in another type of memory modification learning in the reversal Barnes maze task. Following reversal learning, *App*$^{NL\text{-}G\text{-}F}$ mice, given spatial cues, failed to infer coexisting memories for two goal locations during the trial. We concluded that the altered internal states of *App*$^{NL\text{-}G\text{-}F}$ mice illustrated their misclassification in the memory modification process. This novel approach highlights the potential of investigating internal states to precisely assess cognitive changes in early AD and multidimensionally evaluate how early interventions may work.

Keywords: Alzheimer's disease, *App*$^{NL\text{-}G\text{-}F}$ knock-in mice, fear conditioning, reinstatement, latent cause model, Barnes maze task

Introduction

Alzheimer's disease (AD) is the leading cause of dementia, which severely disturbs daily lives in the aging population and causes a tremendous social burden (Nandi et al., 2022). Neuropathological hallmarks, including amyloid-β (Aβ) plaques and neurofibrillary tangles (NFT), define AD biologically, and the advance in biomarkers have enabled early diagnosis (Jack Jr. et al., 2018, 2024). Abnormal biomarkers can appear years before symptoms arise, a stage referred to as preclinical AD (Sperling et al., 2011, 2014). As the disease progresses, the cognitive symptom gets severe, and its degree defines mild cognitive impairment (MCI) and the onset of AD dementia. Cognitive decline has been reported in a wide range of domains, including memory, language, visuospatial function, and executive function, leading to behavioral change as early as a decade before dementia (Amieva et al., 2008; Knopman et al., 2021). The early stage of AD provides a time window to intervene in disease progression, yet there is an explanatory gap and inconclusive consensus on what particular behavioral phenotype can be directly linked to neuropathology in early AD. Given the present circumstance, the subtle and heterogeneous cognitive change before the onset of dementia poses a challenge to assess clinical outcomes with traditional measurement (Jutten et al., 2023).

To precisely assess clinical symptoms and effects of early intervention for AD pathology, determining universal behavioral phenotypes as well as understanding how such behaviors change from asymptomatic to symptomatic is needed. Memory impairment is the predominant first symptom of AD (J. Barnes et al., 2015). The symptoms are commonly identified as a deficit in recalling previously learned or recognized items (Grober et al., 2018). Previous studies suggested that the performance of prioritized recall, in which memories are recalled in order of assigned value, is a sensitive measure in the early stages of AD (for review, see Knowlton & Castel, 2022). It remains elusive how memory impairments may arise from biased information processing that could affect the encoding, maintaining, or retrieval process of memory function. Using genetic mouse models is a common approach to dissecting the effects of pathological changes on cognitive symptoms measured in behavioral tasks while controlling for potential confounding factors (Webster et al., 2014; Zhong et al., 2024). However, the symptoms at the early stage of AD are relatively mild in traditional measurements, and phenotypes may be inconsistent among studies even when using the same AD mouse model (Jankowsky & Zheng, 2017). As behavioral phenotypes in conjunction with brain pathology state may not be captured by current assessments robustly, it is crucial to seek more sensitive and comprehensive approaches.

Recently proposed computational models provide generative processes underlying observed behavior, potentially bridging the gap between behavioral phenotype and AD pathology. Indeed, such

models not only replicated the memory task performance of AD patients in different disease stages (M. D. Lee et al., 2020; M. D. Lee & Stark, 2023; Pooley et al., 2011) but also were valuable in detecting cognitive decline in preclinical AD patients (Bock et al., 2021; Vanderlip et al., 2024). However, previous approaches did not suppose an internal model of the world to predict the future from current observation given prior knowledge. Internal models with sufficient explanatory power can replicate various behavioral phenotypes in specific tasks through a common generative process. Estimating an internal model in the early stages of AD patients would elucidate how AD alters their internal states and thereby their behaviors along with disease progression, which is informative for precise diagnosis and assessment of early intervention effects. To our knowledge, only a few studies tried to estimate internal states in amnestic MCI and AD patients and reported that a certain retention rate underlying motor learning (Sutter et al., 2024) but not learning rate in associative learning (Wessa et al., 2016) differed from healthy control. This circumstance demands us to investigate internal models illustrating not only learning but also its interaction with memory with higher explanatory power.

In this study, we assumed the latent cause model as an internal model explaining the memory modification process to solve a problem such that a novel observation should be classified into previously acquired memory or novel memory (Gershman et al., 2017). We hypothesized that the internal state of the latent cause model is altered along with disease progression from the preclinical stage of AD. To test this, we used the $App^{NL\text{-}G\text{-}F}$ knock-in AD mouse model (Saito et al., 2014). This model mice carry the Swedish KM670/671NL, Iberian I716F, and Arctic E693G mutations on the gene of amyloid-β precursor protein (APP) and display Aβ plaque accumulation and neuroinflammation across age, reproducing neuropathological change in preclinical AD without artifacts caused by APP overexpression (Saito et al., 2014; Sasaguri et al., 2017). In this study, the 6- and 12-month-old $App^{NL\text{-}G\text{-}F}$ knock-in mice and the age-matched control mice were engaged in a memory modification task consisting of Pavlovian fear conditioning, extinction, and reinstatement, Internal states across trials were estimated for each mouse, simultaneously with determining a set of model parameters such that the simulated behavior under the latent cause model given the parameters fits enough to the observed behavior. Because we found a deficit in the $App^{NL\text{-}G\text{-}F}$ mice as overgeneralization or overdifferentiation of their observations into certain memories indicated by parameters in the latent cause model, we further confirmed whether this deficit could be replicated in another kind of memory modification task, the reversal Barnes maze paradigm. In the latent cause framework, we discussed how the internal states in the $App^{NL\text{-}G\text{-}F}$ mice diverged from those in the control mice during the memory modification process.

Results

The latent cause model, an internal model, explains memory modification processes in the classical fear conditioning paradigm.

The latent cause model proposed by Gershman et al. (2010) is an internal model such that an agent infers a latent cause as a source of the co-occurrence of events in the environment. It was later adapted to model the memory modification process, a problem to infer previously acquired latent causes or a novel latent cause for every observation in the Pavlovian fear conditioning paradigm in rodents and humans (Gershman et al., 2017). In the latent cause model, an agent learning an association between a conditioned stimulus (CS) and an unconditioned stimulus (US) has an internal model such that stimuli $\mathbf{x}$ (i.e., d-dimensional vector involving the CS or context) and outcome r (i.e., US) at time t are generated at an associative strength $\mathbf{w}$ (i.e., d-dimensional vector) from k-th latent cause z_{tk} of total K latent causes by a likelihood $p(\mathbf{x}|z_{tk})$ and $p(r|z_{tk})$, respectively (Fig. 1A). $p(\mathbf{x}|z_{tk})$ and $p(r|z_{tk})$ follow Gaussian distributions in which each mean is calculated from previous observations, while each variance (σ_x^2 and σ_r^2) was *a priori* defined as a hyperparameter. Since latent causes are unobservable, the agent given an observation infers which latent cause is likely to generate current observation by Bayes' rule. The prior of z follows the Chinese restaurant process, that is

$$P(z_t = k|\mathbf{z}_{1:t-1}) \propto \begin{cases} \sum_{t'<t} \mathcal{K}(\tau(t)-\tau(t'))\mathbb{I}[z_{t'} = k], & \text{if } k \leq K \text{(i.e., } k \text{is an old cause)} \\ \alpha, & \text{otherwise (i.e., } k \text{is a new cause)} \end{cases}$$

Prior of z_{tk} refers to the frequency that z_k was inferred so far, where $\mathbb{I}[\cdot] = 1$ if its argument is true; otherwise, 0. This prior is biased by a temporal kernel $\mathcal{K}$, as an analogy of the power law of the forgetting curve,

$$\mathcal{K}(\tau(t)-\tau(t')) = (\tau(t)-\tau(t'))^{-g}$$

to exponentially decay the probability that z_k inferred at previous trial t' is inferred again at current trial t by a temporal scaling factor g if the time interval between trial t' and t increases. A new latent cause is inferred at a probability governed by concentration parameter α. Higher α or lower likelihood of the current observation biases the agent to infer a new latent cause given an observation (i.e., differentiation) rather than to infer an old one (i.e., generalization) (Fig. 1B). Although the latent cause model allows for having an infinite number of latent causes, agents with appropriate α and g favor to infer a smaller number of latent causes for observations instead of inferring a unique latent cause for every observation.

According to the latent cause model, the agent solves two computational problems during Pavlovian fear conditioning learning. One is to compute the posterior probability of each latent cause given its associative weight between stimuli $\mathbf{x}$ and shock r with a history of observations, and the other

is to compute the associative weight of each latent cause given its posterior. Both computations are achieved through the expectation-maximization (EM) algorithm, as the former and the latter computation correspond to E-step and M-step, respectively. In the E-step, the posterior of each latent cause is computed from its prior and the likelihood of current observation given the latent cause following Bayes' rule, consequently determining the most likely latent cause at the current trial as the one with the maximal posterior, then moves to the M-step. In the M-step, the agent predicts shock by a linear combination of stimuli and associative weight in each latent cause. The weight is updated based on the Rescorla-Wagner rule (Rescorla & Wagner, 1972) to minimize the prediction error between observed and predicted shock. The amount of the weight change depends on learning rate η and prediction error, which is biased by the posterior probability of the latent causes. The amount of weight change will be smaller even in larger prediction errors if the latent cause is unlikely. The agent estimates weights for all latent causes and then returns to E-step. This procedure is iterated a preset number of times, and thereby, the posterior probability over latent causes and their associative weights are determined. Finally, the agent concludes the expected shock in the current trial by a linear combination of observed stimuli and an expectation of weight, which is weighted by the posterior distribution over latent causes. The conditioned response (CR) for the expected shock is determined as an integral of Gaussian distribution (mean is the expected shock, variance is λ) above threshold θ, where λ and θ are hyperparameters. Thus, these parameters of the latent cause model determine an internal state of the agent leading to particular behavioral outcomes given an observation.

Reinstatement exemplifies the memory modification process and indeed is well explained by the latent cause model. Reinstatement occurs when agents who have learned and then extinguished an association between a CS and a US exhibit CR again after being exposed to the US alone following the extinction. The latent cause model explains the reinstatement of fear memory with the changes of agent's internal state as follows. First, an agent repeatedly observed a CS accompanied with a US in a Pavlovian fear conditioning procedure (Fig. 1C, "Acquisition"). Because these observations are novel for the agent, it inferred a set of latent causes $\mathbf{z}_A$ (z_1, z_2, and z_3 in Fig. 1D) where the CS is associated with the US by associative weight $\mathbf{w}_A$ (w_1, w_2, and w_3 in Fig. 1E) during the acquisition. Considering the effect of the context, $\mathbf{w}$ contains both w_{cs} and $w_{context}$ (Fig. 1E). After the acquisition, the agent inferred $\mathbf{z}_A$ and then predicted the upcoming US, leading to increased CR when it observed the CS (Fig. 1C, "Test 1" after "Acquisition"). Second, the agent repeatedly observed the CS alone in an extinction procedure. Early in the extinction, the agent would infer $\mathbf{z}_A$ and then exhibit higher CR to the CS, while the US was absent (Fig. 1C, left "Extinction"). This state increases prediction error, then prompts the agent either to decrease $\mathbf{w}_A$ toward 0 or to infer a novel set of latent causes $\mathbf{z}_B$ where the CS is not accompanied with the US, that is, associative strength $\mathbf{w}_B = 0$ (Fig. 1B); both processes are

run in the mouse in Fig. 1CDE. Note that the sets A and B were introduced arbitrarily for explanation here. Later in the extinction, $\mathbf{z}_B$ rather than $\mathbf{z}_A$ is inferred, leading to the decreased CR (Fig. 1C, right "Extinction" or "Test 2"). Third, when the agent observes the US alone after the extinction (Fig. 1C, "Unsignaled shock"), the probability of $\mathbf{z}_A$ or associative weight of some latent causes increases, hence it predicts the US and thereby the CR increases again, that is reinstatement (Fig. 1C, "Reinstatement" at "Test 3"); in the mouse in Fig. 1CDE, the unsignaled shock increases not only the posterior probability of $\mathbf{z}_A$ that has maintained higher w_{cs} but also $w_{context}$ in $\mathbf{w}_6$ interpreted as a modification in the association between context and US in z_6. Thus, it is rational to expect that certain parameters of the latent cause model are in conjunction with particular behaviors in the reinstatement paradigm (see Appendix 2 for the relationship between parameters, CR, and internal states). As demonstrated in Gershman & Hartley (2015), the likelihood that $\alpha > 0$ in human participants correlates with the degree of their spontaneous recovery, that is a return of the CR after a long hiatus following an extinction procedure. Furthermore, a recent study has shown that the latent cause model explains maladaptive decision-making in post-traumatic stress disorder patients as a perseverated inference of previously acquired latent causes even under situations where they should infer novel latent causes (Norbury et al., 2022). By estimating such parameters from the behaviors of each experimental mouse, how the internal state of the AD group diverges from that of the control group could be determined.

$App^{NL\text{-}G\text{-}F}$ mice showed unimpaired associative fear and extinction learning but a lower extent of reinstatement.

To examine the age and $App^{NL\text{-}G\text{-}F}$ knock-in effects on the reinstatement, all cohorts of control and homozygous $App^{NL\text{-}G\text{-}F}$ mice were tested at the age of either 6 months or 12 months in the reinstatement paradigm based on auditory-cued fear conditioning (Fig. 2A and Table S1). According to previous studies, the $App^{NL\text{-}G\text{-}F}$ mice displayed an age-dependent Aβ deposition from 4 months old and memory impairments could be detected at 6 months of age (Masuda et al., 2016; Saito et al., 2014; Sakakibara et al., 2018). The wide spread of Aβ accumulation was confirmed in this study, and results were provided in Figure 2 – figure supplement 1. Given that age affects the Aβ accumulation caused by $App^{NL\text{-}G\text{-}F}$ knock-in, a one-way ANCOVA was used to evaluate the $App^{NL\text{-}G\text{-}F}$ knock-in effects on behavioral measures controlling for age.

In the acquisition phase of fear conditioning, the freezing rate gradually increased with the number of CS-US pairings (Fig. 2B, leftmost red panel). In the test 1 phase, 24 h after the acquisition, a significant genotype effect was observed in freezing rates during CS presentation [$F(1,95) = 9.679$, $p = 0.002$], where the $App^{NL\text{-}G\text{-}F}$ mice showed a higher freezing rate than in the control (Fig. 2C and Table S2), indicating the associative fear memory was established in $App^{NL\text{-}G\text{-}F}$ mice. As the increasing

214 trials of CS alone presented, both groups showed a gradually decreased freezing rate in the extinction
215 phase (Fig. 2B, blue panel). In the test 2 phase, 24 h after the last trial of the extinction, there was a
216 significant age effect in freezing rates during CS presentation [$F(1,95) = 4.698$, $p = 0.033$], where 12-
217 month-old mice showed a higher freezing rate than in the 6-month-old mice (Fig. 2D and Table S2).
218 In the test 3 phase, 24 h after given an unsignaled shock, mice showed an elevated freezing rate, and
219 no significant genotype or age effect was detected (Fig. 2B, 2E and Table S2). To evaluate the effect
220 of the context on the CR, we measured the preCS freezing rate during 1 min before the tone
221 presentation (Fig. 2 – figure supplement 2). Generally, the preCS freezing rate was lower than CS
222 freezing rate and gradually reduced but kept positive across trials after acquisition (Fig. 2 – figure
223 supplement 2A and 2B), suggesting the context has a certain effect on CR. When we compared the
224 preCS freezing rate (Figure 2 – figure supplement 2E) and the CS freezing rate (Fig. 2E) in test 3
225 within group by the paired samples *t*-tests, the CS freezing rate was significantly higher than the preCS
226 freezing rate in all groups: 6-month-old control, $t(23) = -6.344$, $p < 0.001$, $d = -1.295$; 6-month-old
227 $App^{NL\text{-}G\text{-}F}$, $t(24) = -4.679$, $p < 0.001$, $d = -0.936$; 12-month-old control, $t(23) = -4.512$, $p < 0.001$, $d = -$
228 0.921; 12-month-old $App^{NL\text{-}G\text{-}F}$, $t(24) = -2.408$, $p = 0.024$, $d = -0.482$. Although the associative strength
229 between the context and the US would be increased by the unsignaled shock, these results suggest that
230 reinstatement was not solely explained by the CR for context, but also reactivation acquisition memory.
231 To evaluate the magnitude of how each mouse can discriminate between different phases, we
232 calculated the discrimination index (DI) using freezing rates during the CS presentation. If the mouse
233 shows the same freezing rate at the two test phases, the value of DI will be 0.5, while if the mouse
234 shows a higher freezing rate in one phase compared to the other, DI will be far from 0.5. The results
235 showed that the DIs between test 2 and test 1 were significantly lower than 0.5 in all groups (Fig. 2F
236 and Table S3), indicating the successful establishment of extinction memory. The genotype effect was
237 detected [$F(1,95) = 5.013$, $p = 0.027$] (Fig. 2F and Table S4), where $App^{NL\text{-}G\text{-}F}$ mice showed lower
238 values due to higher freezing in the test 1 phase. These results suggested that there was no markable
239 fear and extinction learning deficit in $App^{NL\text{-}G\text{-}F}$ mice at the behavioral level. The DIs between the test
240 3 and 2 phases in four groups were significantly higher than 0.5, indicating that the reinstatement was
241 successfully induced after the unsignaled shock (Fig. 2G and Table S3). Six-month-old mice showed
242 higher DIs, as a significant age effect was detected [$F(1,95) = 7.480$, $p = 0.007$] (Fig. 2G and Table
243 S4). The DIs between the test 3 and 1 phases did not deviate from 0.5 for both 6- and 12-month-old
244 control mice, suggesting that they displayed comparable freezing rates in tests 1 and 3 (Fig. 2H and
245 Table S3). In contrast, the DIs between test 3 and 1 phases were significantly lower than 0.5 in 6-
246 month-old $App^{NL\text{-}G\text{-}F}$ mice [$t(24) = -3.245$, $p = 0.003$] and 12-month-old $App^{NL\text{-}G\text{-}F}$ mice [$t(24) = -6.165$,
247 $p < 0.001$] (Fig. 2H and Table S3). Moreover, a significant genotype effect was detected [$F(1,95) =$

248 15.393, $p < 0.001$] (Fig. 2H and Table S4), indicating $App^{NL\text{-}G\text{-}F}$ mice exhibited a lower extent of
249 reinstatement. Thus, these results suggest that mice retrieved fear memory in the test 1 phase and
250 retrieved extinction memory in the test 2 phase, respectively. The fear memory and the extinction
251 memory compete with each other: the US is present in the former, while it is absent in the latter. In the
252 test 3 phase, after they observed the unsignaled shock, mice had to decide which memory was more
253 relevant given the CS. The same level of freezing rate in tests 1 and 3 observed in control mice suggests
254 that they could retain both fear and extinction memory and thereby infer the fear memory more
255 preferentially. In contrast, $App^{NL\text{-}G\text{-}F}$ mice displayed a lower freezing rate in test 3 compared to test 1,
256 suggesting that they might still infer extinction memory even after the unsignaled shock or a
257 completely new memory, while the initial fear memory might be suppressed or eliminated.
258
259 **The internal state underlying reinstatement simulated with the latent cause model differs**
260 **between the ages and *APP* genotype.**
261 To seek the internal state of each mouse generating the behaviors in the reinstatement paradigm,
262 we estimated the parameters of the latent cause model, minimizing prediction errors between the
263 observed CR and simulated CR in the latent cause model given a set of parameters (see also Material
264 and Methods). Figure 3 demonstrates the traces of observed CR, simulated CR, and the changes in the
265 internal state in each group. We initially confirmed that the DIs in the simulated CRs in each group
266 well replicated those in the observed CRs, except that the DI between test 3 and test 1 in the 6-month-
267 old $App^{NL\text{-}G\text{-}F}$ mice did not significantly deviate from 0.5 in simulated results (c.f. Fig. 2F to 2H and
268 Fig. 3 – figure supplement 1A to 1C). To check if the latent cause model has a certain bias to simulate
269 CR in either group, we compared the residual (i.e., observed CR minus simulated CR) between groups.
270 The fit was similar between control and $App^{NL\text{-}G\text{-}F}$ mice groups in the test trials, except test 3 in the 12-
271 month-old group (Figure 3 – figure supplement 1D and 1E). In test 3, the residual was significantly
272 higher in the 12-month-old control mice than $App^{NL\text{-}G\text{-}F}$ mice, indicating the model underestimated the
273 reinstatement in the control mice. These results suggest that the latent cause model fits our data with
274 little systematic bias such as an overestimation of CR for the control group in the reinstatement,
275 supporting the validity of the comparisons in estimated parameters between groups.
276 In the 6-month-old group, two latent causes (z_1 and z_2) were generated in the acquisition phase
277 with increasing associative weight (rows 2 to 4 in Fig. 3A, 3B). During the extinction phase, the same
278 latent causes were inferred, and the corresponding associative weights declined, indicating the
279 devaluation of previously acquired fear memory (rows 2 to 4 in Fig. 3A and 3B). Notably, the w_{cs} in
280 $\mathbf{w}_2$ in control mice remained at a higher value compared to that in $App^{NL\text{-}G\text{-}F}$ mice. At trial 35 given an
281 unsignaled shock, $w_{context}$ in $\mathbf{w}_1$ was elevated in control mice (row 4 in Fig. 3A), while both $w_{context}$ in

282 $\mathbf{w}_1$ and $\mathbf{w}_2$ were elevated in $App^{NL\text{-}G\text{-}F}$ mice (row 4 in Fig. 3B). At trial 36 (test 3 phase), in control mice,
283 higher values in $w_{context}$ in $\mathbf{w}_1$ and w_{cs} in $\mathbf{w}_2$ contributed to increased CR (Fig. 3A), whereas $w_{context}$ in
284 $\mathbf{w}_1$ and $\mathbf{w}_2$ were the main components for reinstatement in $App^{NL\text{-}G\text{-}F}$ mice (Fig. 3B). This discrepancy
285 suggests that the expectation of the US is different between 6-month-old control and $App^{NL\text{-}G\text{-}F}$ mice.
286 The transition of internal state in the 12-month-old group was similar to that in the 6-month-old
287 group until the test 1 phase. Unlike the 6-month-old group, new latent causes (z_3 and z_4) were inferred
288 during the extinction phase in the 12-month-old group (rows 2 to 4 in Fig. 3C, 3D). At trial 35 given
289 an unsignaled shock, the control mice inferred the latent causes z_3 and z_4 acquired in the extinction
290 phase so that their weights $w_{context}$ in $\mathbf{w}_3$ and $\mathbf{w}_4$ were updated (row 4 in Fig. 3C). In contrast, there was
291 virtually no update of weights in $App^{NL\text{-}G\text{-}F}$ mice (row 4 in Fig. 3D). At trial 36 (test 3 phase), the
292 elevated CR in control mice was attributed to a successful update of associative weight after the
293 unsignaled shock (Fig. 3C), whereas the CR in $App^{NL\text{-}G\text{-}F}$ mice did not increase due to a redundant
294 latent cause z_5, which was newly inferred and reduced the posterior probability of latent causes having
295 higher associative weight (Fig. 3D).
296
297 **The unsuccessful state inference at reinstatement was caused by misclassification in $App^{NL\text{-}G\text{-}F}$**
298 **mice.**
299 Due to the different evolution of latent causes between ages (Fig. 3), we separately discuss the
300 contribution of differences in estimated parameters within the same age and their correlations with DI
301 between test 3 (after unsignaled shock) and test 1 (after acquisition) (Fig. 2H). We first investigated
302 the individual internal state differences in the 12-month-old group where the impairment was apparent
303 at the behavioral level (Fig. 2H). $App^{NL\text{-}G\text{-}F}$ mice had significantly lower α (Mann-Whitney $U = 100$, p
304 $= 0.002$) and lower σ_x^2 (Mann-Whitney $U = 88.5$, $p = 0.008$) than control mice (Fig. 4A). α and σ_x^2
305 were significantly correlated each other (Fig. 4 – figure supplement 1). Both were also positively
306 correlated with the DI (Fig. 4B), and the trend was preserved after adjusting for the genotype effect
307 (Fig. 4 – figure supplement 2). While the maximal number of inferable latent causes (K) was
308 significantly higher in $App^{NL\text{-}G\text{-}F}$ mice than in control mice (Table S7), the total number of acquired
309 latent causes was comparable between the groups (K_{total} in Table S9). K was negatively correlated with
310 α and σ_x^2 (Fig. 4 – figure supplement 1).
311 We assume that the effect of K on the memory modification process together with α and σ_x^2 could
312 come to the surface in extremely artificial conditions, but not in natural conditions as in the empirical
313 data (see Appendix 2 – table 6). Statistical results for remaining parameters are shown in Table S7 and
314 S8. It should be noted that these estimated parameters and internal state might not be unique but one
315 of the possible solutions.

316 We further characterized how the latent causes acquired in different phases contribute to the
317 distinct internal states between control and $App^{NL\text{-}G\text{-}F}$ mice. Lower α and lower σ_x^2 would bias $App^{NL\text{-}G\text{-}F}$ mice toward overgeneralization as they favor inferring a small number of latent causes when facing
319 the same cue. Indeed, the number of latent causes initially inferred at the acquisition phase was
320 significantly lower in $App^{NL\text{-}G\text{-}F}$ mice, where most of them inferred two latent causes, and more than
321 half of the control mice inferred three latent causes (Fig. 4C and Table S9). Furthermore, when they
322 initially observed the CS in the absence of the US in test 1, the posterior probabilities of acquisition
323 latent causes in test 1 were significantly higher in the $App^{NL\text{-}G\text{-}F}$ mice than in the control mice, leading
324 to the higher CRs in test 1 in the simulation (Fig. 4 – figure supplement 3), consistent with the observed
325 CR (Fig. 2C).

326 According to the latent cause model, lower σ_x^2 would bias $App^{NL\text{-}G\text{-}F}$ mice toward
327 overdifferentiation unless the cue exactly matches to previously observed ones. In other words, they
328 prefer to infer new potential causes when similar observations are presented instead of interpreting
329 them generated from the same distribution. The likelihood of current stimuli given latent causes was
330 calculated from a Gaussian distribution with the mean of stimulus values under the latent cause so far
331 and fixed variance σ_x^2. Repeated observation of the CS alone during the extinction would sufficiently
332 increase the likelihood of stimuli given the latent causes inferred during the extinction in both $App^{NL\text{-}G\text{-}F}$ and control mice. This attenuates the need to infer new latent causes and the number of latent causes
334 acquired at the extinction phase was comparable between groups (Fig. 4D and Table S9). However,
335 the CS was absent in the unsignaled shock and then presented again in test 3. These observations would
336 decrease the likelihood of stimuli given past latent causes, and consequently increase the probability
337 to infer the new latent causes. This volatility would be more obvious in $App^{NL\text{-}G\text{-}F}$ mice with lower σ_x^2.
338 As a result, the number of latent causes acquired after the extinction (i.e., the unsignaled shock and
339 test 3) as well as the sum of posterior probabilities of them were greater in $App^{NL\text{-}G\text{-}F}$ mice than those
340 in the control mice, despite that the mice received either the US or the CS alone that were the same
341 with those they observed so far (Fig. 4E and Table S9).

342 Such internal states in $App^{NL\text{-}G\text{-}F}$ mice would diverge the update of associative weight from those
343 in the control mice after extinction(Fig. 4F, 4G and Table S9). Both $App^{NL\text{-}G\text{-}F}$ and control mice would
344 still infer the extinction latent causes at the unsignaled shock with higher posterior probabilities. Since
345 the update of the associative weight of each latent cause is modulated by both prediction error and
346 posterior probability of the latent cause, the associative weights in the extinction latent causes were
347 increased by the unsignaled shock in the control mice (Fig. 3C, 4G and Table S9). In $App^{NL\text{-}G\text{-}F}$ mice,
348 however, they inferred new latent causes with low associative weight after the extinction (Fig. 4E, 4H
349 and Table S9), which decreased the posterior probabilities of other latent causes and impeded update

of their weight (Fig. 3D, 4G). As the initial weight of new latent causes showed a trend toward significance that control mice showed a higher value (Table S7), this would explain the significant higher in w_{CS} in extinction latent cause (Fig. 4G). This distinct internal state therefore led to the lower CR and DI in test 3 in $App^{NL\text{-}G\text{-}F}$ mice (Fig. 2E, 2H). These results suggest that the α and σ_x^2 would differentiate the internal state and thereby behavioral phenotypes between $App^{NL\text{-}G\text{-}F}$ and control mice. In the Chinese restaurant process, the probability that a new latent cause is inferred decreases with trials, preventing that too many latent causes are inferred (Gershman & Blei, 2012). This could be one of the reasons why the overgeneralization by lower α was less effective than the overdifferentiation by lower σ_x^2 in test 3 in $App^{NL\text{-}G\text{-}F}$ mice.

As shown previously in Fig. 2 and Fig. 2 – figure supplement 1, Aβ aggregation has already widely spread in the brain of 6-month-old $App^{NL\text{-}G\text{-}F}$ mice, but no clear behavioral impairments were detected by the conventional analysis. Similarly to the 12-month-old group, we investigated the individual internal state differences in the 6-month-old group and tested whether the discrepancy in α and σ_x^2 emerged earlier without behavioral impairment. We found that $App^{NL\text{-}G\text{-}F}$ mice had significantly lower α (Mann-Whitney $U = 38$, $p = 0.012$), and it was significantly correlated with DI (Fig. 5A, 5B and Table S10, S11). The σ_x^2 was comparable between control and $App^{NL\text{-}G\text{-}F}$ mice, but its significant correlation with DI was found (Fig. 5A, 5B and Table S10, S11). Significant positive correlation between α and σ_x^2 was also found in the 6-month-old group (Fig. 5 – figure supplement 1). Unlike in the 12-moth-old group, the significant correlation between α and DI, as well as that between σ_x^2 and DI was not preserved within subgroups (Fig. 5 – figure supplement 2), suggesting genotype would be confounder for α, σ_x^2, and DI. The hyperparameter θ, a threshold to emit CR for stimulus inputs, was significantly higher in $App^{NL\text{-}G\text{-}F}$ mice than in control mice. Since the λ was comparable between control and $App^{NL\text{-}G\text{-}F}$ mice (Table S10), the CR in $App^{NL\text{-}G\text{-}F}$ mice could rapidly decrease with expected US during the extinction phase, even if their CR at the test 1 was higher than those in the control. The effect of higher θ was subtle when the expected US was far from its value, and therefore had little contribution to the DI between test 3 and test 1, which is in line with the correlation result in Table S11. Statistical results for the remaining parameters are shown in Tables S10 and S11.

Similar to the 12-month-old group, the influence of α was observed in the acquisition phase, where most $App^{NL\text{-}G\text{-}F}$ mice inferred two latent causes, while more than half of control mice inferred three latent causes (Fig. 5C). In addition, lower α in $App^{NL\text{-}G\text{-}F}$ mice affect the posterior probabilities of acquisition latent cause and CR in test 1 phase (Fig. 5 – figure supplement 3). No significant difference in the internal states was found between control and $App^{NL\text{-}G\text{-}F}$ mice (Fig. 5C to 5H and Table S12).

In summary, α and σ_x^2 are the main contributors to differential internal states and memory modification processes between control and $App^{NL\text{-}G\text{-}F}$ mice. Regardless of age, $App^{NL\text{-}G\text{-}F}$ mice have a defect in forming a new memory even if the same cue comes with a new outcome as in test 1, suggesting overgeneralization associated with-lower α (Fig. 4A, 5A). With a more severe phenotype, 12-month-old $App^{NL\text{-}G\text{-}F}$ mice were biased to classify similar cues into different causes, which is associated with lower variance of the stimulus, σ_x^2 (Fig. 4A). This overdifferentiation would eventuate the lower reinstatement of fear memory (Fig. 2H). Thus, these results suggest that $App^{NL\text{-}G\text{-}F}$ mice failed to retain competing memories because of the misclassification of observation into memories, even though each memory accounts for past and present observations.

$App^{NL\text{-}G\text{-}F}$ mice failed to infer coexisting memories in the reversal Barnes maze task.

To explore the explanatory power of the latent cause framework formalizing mechanisms of associative learning and memory modification, cohorts 4, 5, 6, and 7 were subjected to a reversal Barnes maze task (C. A. Barnes, 1979) two weeks after the reinstatement experiment (Table. S1). Barnes maze is a commonly used behavioral task to examine spatial learning and memory in Alzheimer's disease model mice (Webster et al., 2014). Associative learning is fundamental in spatial learning (Leising & Blaisdell, 2009; Pearce, 2009). Although we did not make any specific assumptions of what kind of associations were learned in the Barnes maze, the trial-and-error updates of these associations involving sensory preconditioning or secondary conditioning could underlie the learning performance. The first stage contains a 6-day training and 1-day probe test to establish the first spatial memory of the target hole (Fig. 6A). During the training phase, the mice explored a circular field and were reinforced to find an escape box under a target hole as any one of 12 holes equally spaced around the perimeter of the field (Fig. 6A). In the 12-month-old group, the number of errors, latency, and travel distance from start to goal decreased across days in the training phase, suggesting successful initial learning (Fig. 6 – figure supplement 1A, and Table S13), A significant interaction between day and genotype [$F(5, 170) = 2.447$, $p = 0.036$] was observed in the latency where control mice took a longer time to reach the target hole on training days 4, 5, and 6 (Fig. 6 – figure supplement 1A, and Table S13). In order to estimate what kind of strategies mice used to solve the task relying on their memory, we performed algorithm-based strategy analysis (Suzuki & Imayoshi, 2017; Tachiki et al., 2023). A moving trajectory in each trial was categorized into one of five spatial strategies following the definition (Fig. 6 – figure supplement 2). As in the conventional analysis above, the usages of the strategies were virtually comparable between $App^{NL\text{-}G\text{-}F}$ and control mice (Fig. 6B and Table S14).

In the probe test phase, the mice explored the field for 5 min without the escape box, which generated a certain prediction error (Fig. 6A). While the mice retrieved the spatial memory of the target

hole in spite of the novel observation, they would stay around the target hole. Consequently, *App*$^{NL\text{-}G\text{-}F}$ mice spent significantly less time around the target hole than control mice did in the probe test 1, as a significant interaction between genotype and holes [$F(11, 374) = 4.22$, $p < 0.001$] was detected (Fig. 6C and Table S15). These results suggest that *App*$^{NL\text{-}G\text{-}F}$ mice could successfully form a spatial memory of the target hole, while they did not retrieve the spatial memory of the target hole as strongly as control mice when they observed the absence of the escape box during the probe test.

In the first reversal learning phase, the position of the target hole with the escape box was moved to the opposite from day 9 to 11, then returned to the original position from day 12 to day 14 (Fig. 6A). Through the first reversal learning, we expected that the second spatial memory of the target hole is formed that competes with the spatial memory of the first target hole. In other words, this reversal training phase allowed the mice to acknowledge that both two holes were possible to be the target hole. The improved performance across reversal training days in the conventional analysis was similar to that of the initial training phase, suggesting intact behavioral flexibility (Brown & Tait, 2010; Gawel et al., 2019) (Fig. 6 – figure supplement 1B and Table S13). Unexpectedly, 12-month-old *App*$^{NL\text{-}G\text{-}F}$ mice reached the goal significantly faster than the control mice [Genotype effect from day 9 to 11, $F(1,68) = 10.53$, $p = 0.003$; Genotype effect from day 12 to 14, $F(1,68) = 10.27$, $p = 0.003$] (Fig. 6 – figure supplement 1B, and Table S13). This would be due to the difference in the strategies primarily taken in each group of mice. During the first reversal training (days 9 to 11), more than half of *App*$^{NL\text{-}G\text{-}F}$ mice used the perimeter strategy, which is significantly higher than control mice (Fig. 6B and Table S14). The mice taking this strategy could reach the novel target hole by exploring holes sequentially without spending time around the previous target hole (Fig. 6 – figure supplement 2). On the other hand, the use of perimeter strategy decreased, and spatial strategy increased in control mice from days 9 to 11 (Fig. 6B). Moreover, the use of the confirmatory strategy was around 30% in control mice, which is significantly higher than *App*$^{NL\text{-}G\text{-}F}$ mice (Fig. 6B and Table S14). This strategy is the one such that the mice visit the first and second target holes as if they relied on the two memories of the target holes where the escape box should exist (Fig. 6 – figure supplement 2). During the second reversal training (days 12 to 14), *App*$^{NL\text{-}G\text{-}F}$ mice constantly used the perimeter strategy more than the control mice. Meanwhile, the control mice used the confirmatory strategy more than *App*$^{NL\text{-}G\text{-}F}$ mice, as in the first reversal learning phase (Fig. 6B and Table S14). These results suggest that the search strategy diverged between *App*$^{NL\text{-}G\text{-}F}$ mice and control mice since the reversal learning. The control mice could retrieve two competing spatial memories of the target hole during each trial. In contrast, *App*$^{NL\text{-}G\text{-}F}$ mice were less likely to retrieve these two memories within a trial even though the use of spatial strategy was comparable with control mice on the same day.

450 As expected, the *App*$^{NL\text{-}G\text{-}F}$ mice explored significantly shorter time for the second target hole as
451 well as the first target hole, compared to the control mice (Fig. 6D and Table S15). A significant
452 interaction between genotype and hole was detected [Genotype × Hole, $F(11, 374) = 3.24$, $p < 0.001$].
453 The exploration time was uniform over holes in *App*$^{NL\text{-}G\text{-}F}$ mice, while control mice showed hole
454 preferences around two target holes as if the *App*$^{NL\text{-}G\text{-}F}$ mice and the control mice took the perimeter
455 and the confirmatory strategy, respectively. At last, we calculated the discrimination index (DI) from
456 the exploration time for the first and second target holes in probe test 2 and test 1. As in the
457 reinstatement experiment, we expected that if the mice did not retrieve a spatial memory of the second
458 target hole during probe test 2, the exploration time of this hole would be the same between probe tests
459 1 and 2, resulting in DI around 0.5. The DI for the second target hole was significantly higher than 0.5
460 in control mice, indicating that the mice successfully modified the spatial memories from probe tests
461 1 to 2 at a discriminable level (Fig. 6E and Table S16). On the contrary, the DI in *App*$^{NL\text{-}G\text{-}F}$ mice was
462 comparable to 0.5 and significantly lower than that of control mice, indicating that their spatial memory
463 for the second target hole was not modified after the reversal learning (Fig. 6E and Table S16). We
464 confirmed that the DIs for the first target hole were close to 0.5 and comparable between groups,
465 suggesting that the exploration time for the first target hole remained similarly long in both probe tests
466 (Fig. 6 – figure supplement 1C, and Table S16).
467 The distinct learning process observed in the 12-month-old group was also present in the 6-
468 month-old group. The performance shown in the conventional analysis was comparable between the
469 6-month-old control and *App*$^{NL\text{-}G\text{-}F}$mice (Fig. 6 – figure supplement 2A, 2B, and Table S17). The use
470 of perimeter strategy was over 30% throughout two reversal training phases in 6-month-old *App*$^{NL\text{-}G\text{-}}$
471 ^{F}mice, while control mice used significantly higher confirmatory strategy in the second reversal
472 training (Fig. 6F, and Table S18). During the probe tests, the difference in exploration time around the
473 target hole was less prominent in the 6-month-old group, while *App*$^{NL\text{-}G\text{-}F}$mice uniformly explored the
474 holes as in the 12-month-old *App*$^{NL\text{-}G\text{-}F}$ mice, especially in the probe test 2 (Fig. 6G, 6H, and Table
475 S19). Unlike 12-month-old control mice, the 6-month-old control mice did not show a strong
476 preference for the second target hole in probe test 2 (Fig. 6H). However, their DIs for the first and
477 second target holes were significantly higher than 0.5, suggesting an increased confidence of spatial
478 memory for each target hole (Fig. 6I, Fig. 6 – figure supplement 3C, and Table S20). In *App*$^{NL\text{-}G\text{-}F}$
479 mice, the DIs for the first and second target hole were comparable to 0.5, as those in 12-month-old
480 *App*$^{NL\text{-}G\text{-}F}$ mice, suggesting few modifications of the spatial memories (Fig. 6I, Fig. 6 – figure
481 supplement 3C, and Table S20).
482 In summary, these results suggest that the control mice retained competing spatial memories for
483 the first and the second target holes following reversal learning. In contrast, the *App*$^{NL\text{-}G\text{-}F}$ mice retained

484 at most one spatial memory during a trial. In other words, the *App*$^{NL\text{-}G\text{-}F}$ mice, given the spatial cues,
485 struggled to infer the spatial memory of either the first or second target hole, leading to the rare usage
486 of the confirmatory strategy. Thus, we confirmed the misclassification in the *App*$^{NL\text{-}G\text{-}F}$ mice in the
487 reversal Barnes maze paradigm. Although both DIs in the reinstatement (Fig. 2H) and the Barnes maze
488 experiment (Fig. 6E, 6I) suggest that inferred memories in each time point would diverge between the
489 *App*$^{NL\text{-}G\text{-}F}$ mice and the control mice, no significant correlations were detected (12-month-old group:
490 Spearman's correlation coefficient = 0.062, p = 0.738, N = 32; 6-month-old group: Spearman's
491 correlation coefficient = -0.029, p = 0.862, N = 39), potentially due to differential effects of Aβ
492 accumulation on associative and instrumental learning.

Discussion

Recent advances in disease-modifying intervention have shed light on AD treatment and prevention. Identifying the earliest manifestation of cognitive decline is critical for intervention. However, the behavioral phenotypes and altering trajectories are heterogeneous in predementia AD patients (Duara & Barker, 2022; Jutten et al., 2023), yet standard neuropsychological tests have been optimized for detecting stereotypical symptoms near or after the onset of dementia (Snyder et al., 2014). Recently proposed generative models successfully contributed to predicting nature-nurture factors conjugate with Aβ and behavioral outcomes from the preclinical to severe stage of AD (Hwang et al., 2023; Petrella et al., 2019; Yada & Honda, 2023), while only a few studies explicitly assume an internal model of the patients (Sutter et al., 2024). Assuming an internal model would broaden and deepen interpretations of various cognitive deficits in terms of common computational account (Kocagoncu et al., 2021).

In line with this, our perspective is that states of an internal model would be one candidate of factors to generate a wide variety of behavioral phenotypes in AD, and hence probing the internal state would contribute not only to detecting early signs of AD but also to multidimensionally evaluating how the disease gets worse and how early intervention works. In this study, we assumed the latent cause model proposed by Gershman et al. (2017) as an internal model and unveiled the internal state in 6- and 12-month-old *App*$^{NL\text{-}G\text{-}F}$ mice, a preclinical mouse model of AD, from their behavioral data by estimating the parameters of the latent cause model well replicating the behaviors.

The misclassification in the memory modification process underlies the impairment of *App*$^{NL\text{-}G\text{-}F}$ mice in the reinstatement of conditioned fear memory.

Initially, we confirmed a significant accumulation of Aβ increased along aging in the *App*$^{NL\text{-}G\text{-}F}$ mice (Fig. 2 – figure supplement 1). We then tested whether the memory modification process was behaviorally impaired in the *App*$^{NL\text{-}G\text{-}F}$ mice across the auditory-cued fear conditioning, extinction, and reinstatement. First, the mice learned an association between a tone as a CS and an electrical shock as a US, as they exhibited increased CR in test 1 (Fig. 2B, 2C). Next, they observed the CS alone repeatedly through the extinction, and then they exhibited decreased CR in test 2 (Fig. 2B, 2D). After they observed the US alone, the mice exhibited increased CR again in test 3, suggesting that the reinstatement was established (Fig. 2B, 2E). However, the *App*$^{NL\text{-}G\text{-}F}$ mice exhibited a lower extent of reinstatement compared to the control (Fig. 2H). The DI between test 1 and 3 in the *App*$^{NL\text{-}G\text{-}F}$ mice was significantly lower than 0.5, while that in the control was comparable with 0.5 (Fig. 2H), suggesting that the *App*$^{NL\text{-}G\text{-}F}$ mice would have lower expectations of upcoming shock in the

526 reinstatement than immediately after the acquisition. These results were the first to report the deficits
527 in the reinstatement in the *App*$^{NL\text{-}G\text{-}F}$ mice, suggesting that Aβ accumulation induces the defect in the
528 memory modification process after the extinction.

529 Although our results support that the Aβ accumulation in *App*$^{NL\text{-}G\text{-}F}$ mice would have little effect
530 on associative fear learning and extinction, the results have been mixed in studies using fear
531 conditioning tasks. Two studies found normal contextual or auditory-cued fear conditioning responses
532 up to 18 months old (Kundu et al., 2021; Sakakibara et al., 2018), which aligns with our results. On
533 the contrary, another two studies reported that *App*$^{NL\text{-}G\text{-}F}$ mice have defects in both contextual and cued
534 fear conditioning from the age of five months (Emre et al., 2022; Mehla et al., 2019). The inconsistency
535 suggests the impact of Aβ may be complicated and cannot be simply detected by common behavioral
536 tasks, highlighting the need for novel approaches to probe in-depth behavioral changes.

537 Recent advances in computational psychiatry offer a wealth of internal models to interpret a
538 patient's internal states (Hauser et al., 2022; Huys et al., 2016; Montague et al., 2012). For instance,
539 Norbury et al. (2022) simulated a memory modification process in severe post-traumatic stress disorder
540 (PTSD) patients by the latent cause model and found that they were biased to infer old memories rather
541 than to acquire new memories for novel observations, as they have lower concentration parameter α.
542 In this study, we found that *App*$^{NL\text{-}G\text{-}F}$ mice exhibited lower α both at 6 months and 12 months of age
543 (Fig. 4A, 5A). As the Aβ accumulation increased with age (Fig. 2 – figure supplement 1), lower σ_x^2
544 was observed in 12-month-old *App*$^{NL\text{-}G\text{-}F}$ mice (Fig. 4A). As α and σ_x^2 significantly correlated with DI
545 between test 3 and 1 (Fig. 4B, 5B), we considered both as critical parameters that shaped the diverged
546 internal states underlying reinstatement between control and *App*$^{NL\text{-}G\text{-}F}$ mice. Although we have
547 considered alternative internal models, the latent cause model offered better explanatory power to our
548 data (see Appendix 1), therefore we stood on the latent cause model to explain reinstatement.

549 *App*$^{NL\text{-}G\text{-}F}$ mice with lower α favored to generalize the observation as if it arose from the old cause
550 instead of differentiating it as a new memory. Indeed, their posterior probabilities of acquisition latent
551 causes were higher than those of the control mice when a novel observation was given in test 1 (Fig.
552 4 – figure supplement 3B), contributing to the higher CR in *App*$^{NL\text{-}G\text{-}F}$ mice (Fig. 2C, Fig. 4 – figure
553 supplement 3A). Such overgeneralization associated with lower α may, in part, explain the deficits in
554 discriminating highly similar visual stimuli reported in AD mouse model (Ding et al., 2023; Saifullah
555 et al., 2020; Zhu et al., 2017), preclinical AD patients (Leal et al., 2019; H. Lee et al., 2020), and MCI
556 patients (Ally et al., 2013; Belliart-Guérin & Planche, 2023; Laczó et al., 2021; Parizkova et al., 2020;
557 Wesnes et al., 2014; Yassa et al., 2010), as they classify the lure as previous learned object.

558 Following the Chinese restaurant process, the impact of α on internal states decreases with trials,
559 and meanwhile that of σ_x^2 relatively increases in our reinstatement paradigm. *App*$^{NL\text{-}G\text{-}F}$ mice in 12-

560 month-old with lower σ_x^2 favored to differentiate similar observations into unique memories,
561 especially in later phases, due to the decreased posterior probabilities of existing latent causes. Indeed,
562 *App*$^{NL\text{-}G\text{-}F}$ mice had a larger number of latent causes acquired after the extinction with posterior
563 probabilities (Fig. 4E), resulting in a limited update of the associative weight of the existing latent
564 causes by unsignaled shock (Fig. 4D), and thereby lower CR in test 3 (Fig. 2E, 2H). These results may
565 explain the impaired ability to transfer previously learned association rules could be due to
566 misclassified them as separated knowledge in preclinical autosomal dominant AD mutation carriers
567 and mild AD patients (Bódi et al., 2009; Petok et al., 2018). Such overdifferentiation associated with
568 lower σ_x^2 might underlie delusions, a common neuropsychiatric symptom reported in AD dementia
569 (Kumfor et al., 2022), as an extended latent cause model could simulate the emergence of delusional
570 thinking (Erdmann & Mathys, 2022). Thus, the deficit in reinstatement of conditioned fear memory in
571 the 12-month-old *App*$^{NL\text{-}G\text{-}F}$ mice could be due to their overgeneralization or overdifferentiation of
572 observations into memories, and consequently the competing memories were not retained through the
573 memory modification process (Fig. 4 and 5).
574 Our study demonstrated that estimating internal states with the parameters of the latent cause
575 model provided additional explanations for the cognitive differences between control and *App*$^{NL\text{-}G\text{-}F}$
576 mice. It is undoubted that conventional behavioral tests for AD patients or AD mouse models have
577 provided evidence of cognitive decline and are useful for stratification in clinical trials. Nonetheless,
578 it is also possible that certain deficits have antecedently emerged in internal states, even if behavioral
579 deficits are not obvious as in 6-month-old *App*$^{NL\text{-}G\text{-}F}$ mice. The computational phenotype was proposed
580 by Montague et al. (2012) and defined as parameters of a computational model being in conjunction
581 with measurable behavioral or biological phenotypes. In this sense, α and σ_x^2 in the latent cause model
582 might satisfy this definition in this study. Computational phenotype was established to fill explanatory
583 gaps between biological and psychological evidence in psychiatry, which would be informative
584 likewise in AD studies.
585
586 **The misclassification prevents *App*$^{NL\text{-}G\text{-}F}$ mice from retaining competing memories.**
587 We further tested whether the deficit in the *App*$^{NL\text{-}G\text{-}F}$ mice observed in the reinstatement
588 experiments can be replicated in the reversal Barnes maze paradigm, where the reversal training phases
589 could resemble the extinction phase in the reinstatement paradigm, in that "counterexample" of the
590 previously established memory is observed. The *App*$^{NL\text{-}G\text{-}F}$ and control mice exhibited similar learning
591 curves in the initial training phase regardless of age (Fig. 6 – figure supplement 1A and 3A). These
592 data align with previous studies that did not find a significant spatial learning deficit in *App*$^{NL\text{-}G\text{-}F}$ mice
593 up to 10 months old by using the Morris water maze, a common spatial learning task relevant to the

Barnes maze (Latif-Hernandez et al., 2019; Saifullah et al., 2020; Whyte et al., 2018). Although *App*$^{NL-G-F}$ mice were previously reported to have longer latency and make more errors during training in the Barnes maze (Broadbelt et al., 2022; Sakakibara et al., 2018), these studies haven't described the mice's navigation strategies to explain such deficits.

During the reversal training, *App*$^{NL-G-F}$ mice displayed behavioral flexibility as control mice did based on the conventional analysis (Fig. 6 – figure supplement 2B and 3B), similar to findings in Sakakibara et al. (2018). However, our strategy analysis revealed that this comparable performance was achieved in a sub-optimal way. The usage of confirmatory strategy such that mice selectively explored the first and the second target hole was significantly higher in the control mice during the reversal training phase, especially in the 12-month-old (Fig. 6B, 6F). After the reversal learning, the confirmatory strategy would become the optimal strategy at the first trial in each day to know today's goal location. In contrast, the *App*$^{NL-G-F}$ mice primarily took the perimeter search strategy such that mice explored the target hole in a "brute-force" manner (Fig. 6B, 6F), consistent with deficits in navigation ability found in preclinical AD patients (Allison et al., 2016; Coughlan et al., 2018). These results suggest that the control mice solved the task by inferring the two competing spatial memories representing the two possible target holes, whereas the *App*$^{NL-G-F}$ mice solved the task not relying on selective spatial memories but the perimeter strategy, while both mice were given the same spatial cues. Thus, our hypothesis that the *App*$^{NL-G-F}$ mice have a deficit in classifying observed cues into two or more competing memories in the same context would be supported by the results of the reversal Barnes maze experiment.

In the perspective from the latent cause framework, we posit that the mice would learn a larger number of association rules between stimuli in the maze compared to those in the reinstatement of the fear conditioning, and either infer new memories or modify existing memories for the unexpected observations in the Barnes maze (e.g., changed location or absence of escape box) as in the reinstatement paradigm. In the reversal Barnes maze paradigm, the animals would infer that a latent cause generates the stimuli in the maze at certain associative weights in each trial and would adjust behavior by retaining competing memories. Both overgeneralization and overdifferentiation could explain the lower exploration time of the target hole in the *App*$^{NL-G-F}$ mice. In the case of overgeneralization, the mice would overwrite the existing spatial memory of the target hole with a memory that the escape box is absent. In the case of overdifferentiation, the mice would infer a new memory such that the goal does not exist in the novel field, in addition to the old memory where the goal exists in the previous field. In both cases, the *App*$^{NL-G-F}$ mice would not infer that the location of the goal is fixed at a particular point and failed to retain competing spatial memories of the goal, leading to relying on a less precise, non-spatial strategy to solve the task.

As the latent cause model was built on the fear conditioning paradigm and does not explicitly assume the contribution of other cognitive functions such as attention or working memory, this study provides no direct evidence as to which model parameters are in conjunction with the behavioral outcomes in the reversal Barnes maze and which parameters involve other the cognitive functions that are presumably affected by AD pathology (Finke et al., 2013; Kirova et al., 2015; Malhotra, 2019). To evaluate them, further studies would be required to integrate computational models encompassing different cognitive functions, such as multiple successor representations (Madarasz & Behrens, 2019) or hidden state inference (Sanders et al., 2020) for spatial learning, upon more fundamental principles, such as the predictive coding in a hierarchical neural network (Kocagoncu et al., 2021). We believe that such models would not only explain a wide variety of symptoms in dementia and AD in a common framework but also predict disease progression or effects of interventions for various cognitive functions.

Putative brain regions involving the altered internal states in the *App*$^{NL\text{-}G\text{-}F}$ mice

This study did not explicitly provide evidence about neural mechanisms underlying the defect in memory classification in *App*$^{NL\text{-}G\text{-}F}$ mice. Nonetheless, we speculated that possible neural circuits are presumably involved in the results reported here according to the findings of the previous studies.

It is well known that the amygdala, the hippocampus, and the medial prefrontal cortex are critical in regulating fear memory and its extinction (Maren et al., 2013). In addition, the ventral tegmental area (VTA) plays a modulatory role with its wide projection to these brain regions (Beier et al., 2015; Cai & Tong, 2022). The dysfunction in the abovementioned brain regions may cause the defect of reinstatement. Indeed, previous studies have shown that the inhibition of dopaminergic signal from VTA to the infralimbic cortex (Hitora-Imamura et al., 2015) or inhibition of glutamatergic signal from VTA to the dorsal hippocampus (Han et al., 2020) during unsignaled shock reduced the reinstatement of fear memory in mice.

Although the neural implementation of the latent cause model has not been demonstrated yet, contributions of certain neuronal populations are expected. It is well established that prediction error in classical or operant conditioning learning is computed in the VTA (Lerner et al., 2021). The prediction error is signaled from the VTA to the hippocampal CA1 via the dopaminergic projections, while the VTA receives a feedback signal from the CA1 (Lisman & Grace, 2005). The pattern separation and completion were implemented in the dentate gyrus (DG) and the CA3 in the hippocampus, respectively (Neunuebel & Knierim, 2014). As Sanders et al. (2020) pointed out, one might find an analogy between the process of pattern completion/separation and the inference of old/new latent cause, in the sense that both consist of a generalization/differentiation process to assign

662 a current observation to either the same class with the previous observations or a novel class. The
663 dysfunction in this circuit has long been implicated in memory problems in aging and AD (Palmer &
664 Good, 2011; Wilson et al., 2006). Gershman et al. (2017) predicted that activities of newborn neurons
665 in the DG and neurons in the CA3 correlate with the prior probability over latent causes, while the
666 dopaminergic projection from the VTA to the hippocampal CA1 represents the posterior probability
667 over latent causes. In other words, prediction error signals from the VTA to the CA1 might serve to
668 compute the likelihood of the observation given the latent causes. The feedback from the CA1 to the
669 VTA might signal posterior probability over latent causes and then modulate the extent of associative
670 weight update.

671 Different lines of evidence suggest that hippocampus functions were affected in $App^{NL\text{-}G\text{-}F}$ mice,
672 including lower neuronal firing rates in the CA1 region (Inayat et al., 2023), degradation of gamma
673 oscillation power in the CA3 region (Arroyo-García et al., 2021), and disrupted place cell remapping
674 in CA1 region (Jun et al., 2020). While VTA dysfunction was not reported in $App^{NL\text{-}G\text{-}F}$ mice, studies
675 using Tg2576 mice, which overexpress mutated APP, have shown that neurodegeneration in the VTA
676 occurred at an early age and consequent dysfunction was correlated with impairments in the
677 hippocampus (Nobili et al., 2017; Spoleti et al., 2024). Given the evidence above, dysfunctions caused
678 by pathological Aβ in the hippocampus and VTA are potential candidates underlying unsuccessful
679 latent cause inference in $App^{NL\text{-}G\text{-}F}$ mice, leading to a lower extent of reinstatement. Although this study
680 considered limited disease factors, cognitive impairments in AD patients result from a combination of
681 neuropathological factors and cannot be ascribed solely to Aβ deposition in clinical practice.
682 Examining whether the interaction of neuropathology can be seen in the internal state with a mouse
683 model including tau pathology, e.g., $App^{NL\text{-}G\text{-}F}$/*MAPT* double knock-in mice (Hashimoto et al., 2019;
684 Saito et al., 2019), would improve the translational interpretation in this study.
685

686 **Conclusion**

687 Uncovering the underdeveloped symptoms in the AD continuum is critical to early detection and
688 intervention by filling the explanatory gap between the pathological states and the cognitive decline in
689 AD. We expect that seeking internal state as well as parameters in an internal model in conjunction
690 with behavioral or biological phenotypes could provide in-depth explanations about behavioral
691 outcomes in AD beyond assessing behavior alone. Our study is the first to report that the alteration in
692 the internal state of $App^{NL\text{-}G\text{-}F}$ mice biased their memory modification process toward
693 overgeneralization or overdifferentiation.
694

695 # Materials and Methods

696 ## Animals

697 Male *App*$^{NL\text{-}G\text{-}F}$ homozygous knock-in mice (abbreviated as *App*$^{NL\text{-}G\text{-}F}$ mice in this study) and age-
698 matched control mice were subjected to behavioral tests at 6 and 12 months old. Cohort and genotype
699 information was provided in Table S1. We decided to assign at least 20 mice for each group (6-month-
700 old *App*$^{NL\text{-}G\text{-}F}$ mice, 6-month-old control mice, 12-month-old *App*$^{NL\text{-}G\text{-}F}$ mice, and 12-month-old control
701 mice) on average when designing the reinstatement experiment (see below). Animals were housed
702 under a 12-h light/dark cycle in a temperature-controlled environment (23–24°C temperature and
703 40%–50% relative humidity) with food pellets (Japan SLC) and water provided *ad libitum*. Behavioral
704 experiments were performed during the dark cycle (8 am–8 pm). After the behavioral experiments,
705 mice were sacrificed by cervical dislocation unless their brains were sampled. All animal procedures
706 were performed in accordance with the Kyoto University animal care committee's regulations
707 (approval number: Lif-K24014).
708

709 ## Behavioral test

710 ### *Reinstatement paradigm in fear conditioning*

711 The reinstatement of conditioned fear response after a reminder of unconditioned stimulus is
712 observed in different species (Hermans et al., 2005; Hitora-Imamura et al., 2015; Monfils et al., 2009).
713 In the present study, the four-chamber fear conditioning testing system (O'HARA & CO., LTD, Tokyo,
714 Japan) was used to deliver tone and electrical foot shock, and the top view camera in the chamber
715 recorded mouse behavior at 2 Hz, as previously described (Tachiki et al., 2023). The procedure
716 contained the following phases (Fig. 2A): acquisition (day 1), test 1 (day 2), extinction (day 3 and day
717 4), test 2 (day 5), unsignaled shock (day 8), and test 3 (day 9). The context (acrylic box) was the same
718 throughout the test. Mice were transferred to the experimental room from the breeding room and then
719 habituated for 30 min in their homecage in a rack until the phase started. In each phase, mice were
720 moved between their homecage and the chamber via a transporting cage filled with woodchips. In the
721 acquisition phase, a 60-s tone (65 dB, 10 kHz) as the conditioned stimulus (CS) was presented 300 s
722 later since the mice were released in the chamber. The last 2 seconds of the CS were accompanied by
723 an electrical foot shock (0.3 mA) as the unconditioned stimulus (US). This CS-US association was
724 repeated 3 times with a 90-s interval. After the last US, the mice were left in the chamber for 140 s and
725 then returned to the homecage. In the test phases, a 60-s tone was given 120 s after the mice entry. In
726 the extinction phase, only a 60-s tone was given 19 times on day 3 and 10 times on day 4 with a 200-
727 s interval. In the unsignaled shock phase, an electrical foot shock (0.7 mA) was given 10 min after the

728 mice were placed into the chamber. After the electrical shock, the mice were left in the chamber for 5
729 min and then transferred back to the homecage. The chamber and grid were cleaned with 70% propanol
730 before trials started. The freezing behavior was measured as a conditioned response (CR), and the
731 criterion for freezing is defined as the change of detected mouse area being less than 10 pixels within
732 1 second. The freezing rate was calculated as the percentage of time spent freezing 1 min before (preCS)
733 and during the tone presentation (CS). Discrimination index (DI) has been widely used as a normalized
734 measure to evaluate the extent to which the system can distinguish between two conditions. To evaluate
735 the magnitude of how each mouse can discriminate between different phases, DI is calculated as CR(t)
736 / [CR(t) + CR(t')], where CR indicates freezing rate during CS presentation; t and t' indicate any two
737 test phases.

738 ***Barnes maze***

739 Barnes maze is a commonly used behavioral task to examine spatial learning and reference
740 memory in Alzheimer's disease model mice (Webster et al., 2014). The apparatus and setting were the
741 same as previously reported (Suzuki & Imayoshi, 2017; Tachiki et al., 2023)(Bio-Medica, Osaka,
742 Japan). The maze consisted of a circular open arena (98 cm in diameter) with 12 holes equally spaced
743 around the perimeter. A black iron escape box was placed under one of the holes as a goal during the
744 training phase. The location of the goal stayed constant for a given mouse but was randomized across
745 mice. Four unique visual cues were located at the outside of the arena. Behavior during the trials was
746 recorded using a GigE Vision camera (UI-5240SE-NIR; IDS Imaging Development Systems GmbH,
747 Obersulm, Germany) mounted on the top of the maze to estimate the mouse's position in each frame.
748 All programs used for data acquisition, processing, saving, and synchronized device controls were
749 written in LabVIEW 2013 (National Instruments, TX, USA). The procedure contained the following
750 phases (Fig. 6A): habituation (day 0), initial training (day 1 to day 6), probe test 1 (day 7), retraining
751 (day 8), reversal training 1 (day 9 to day 11), reversal training 2 (day 12 to day 14) and probe test 2
752 (day15). Mice were transferred to the experimental room from the breeding room and then waited for
753 30 min in their homecage in a rack by the phase started. In the habituation phase, mice were released
754 to the center of the field and could freely explore the field for 5 min. After the trial, the mice were
755 moved into the escape box for another 5 min, then back to the homecage. In the training phase, the
756 mice had to explore the field until they successfully got into the escape box within 10 min. Then, they
757 were returned to the homecage and on standby for the next trial. In the probe test, the procedure was
758 the same as the habituation phase except that the mouse was returned to the homecage directly after
759 the field exploration. The field was cleaned with 70% ethanol thoroughly, and the paper bedding in
760 the escape box was changed per trial to reduce the effect of possible olfactory cues in the maze.

The analysis of the Barnes maze was performed on the custom code based on our previous study (Suzuki & Imayoshi, 2017) with some modifications in strategy analysis. The code for strategy analysis is freely available online at https://github.com/suzuki-yusuke/BM_analysis.git/. In brief, the conventional analysis included the number of errors (count to visit holes other than the goal), latency, and travel distance from start to goal in each trial in the training phase, and time spent around each hole in the probe test. The strategy analysis was performed to qualitatively evaluate what kind of exploratory strategies were used by mice based on their belief in each trial (Suzuki & Imayoshi, 2017; Tachiki et al., 2023). For each mouse, the trajectory from start to goal in each trial was automatically assigned to any one of the exploration strategies following the definition described in Figure 6 – figure supplement 2. In this study, confirmatory and perimeter strategies were assumed in addition to the strategies previously reported. The former is the one such that the mice visit two holes, the current or past goal located opposite in the field, as if they have two memories such that the goal exists in each of the two holes. The latter is the one such that the mice find the goal simply running along with the holes, as if they do not have memories that the goal exists in a specific hole. To evaluate how the belief in the mice changed after the reversal training, the discrimination index (DI) is calculated as $\mathrm{ER}(t_2)$ / $[\mathrm{ER}(t_1) + \mathrm{ER}(t_2)]$, where ER indicates the exploration rate; t_1 and t_2 indicate probe test 1 and probe test 2. The ER was calculated as time spent around the hole divided by the maximum exploration time across all holes in each mouse. The DI is calculated for the first target hole (hole 1) and the second target hole (hole 7).

Simulating reinstatement and estimating parameters in the latent cause model

As the latent cause model serves as a generative model, it allows us to simulate the reinstatement experiment while exploring a set of model parameters, minimizing the prediction error between traces of simulated and observed CR for each mouse. The observed trace of CR across all trials in each mouse was initially normalized. In the simulation, the reinstatement experiment was replicated as follows. The number of trials was the same with our experiment: trials 1 to 3 was the acquisition phase, trial 4 was the test 1 phase, trials 5 to 23 was the first extinction phase, trials 24 to 33 was the second extinction phase, trial 34 was the test 2 phase, trial 35 was the unsignaled shock phase, and trial 36 was the test 3 phase. All trials contained stimuli, while the acquisition and the unsignaled shock trials contained the US. The stimulus given a trial was represented as a two-dimensional vector, where the first and the second dimensions indicated the CS and the context, respectively. If the CS was presented at the phases except for the unsignaled shock, the value got 1; otherwise, 0. Because the context was always presented, the value of context was given 0.2 for all trials. The US was represented as a binary

value, either 1 at the acquisition and the unsignaled shock or 0 otherwise. The temporal distance of 24 h between each phase was set to be 20, the same as in the original study (Gershman et al., 2017).

Ten model parameters to be estimated were concentration parameter (α), temporal scaling parameter (g), learning rate (η), maximum number of iterations in the EM algorithm (max. no. of iteration), initial associative weight (w_0), variance of the US (σ_r^2) and the stimulus (σ_x^2), response threshold (θ), response gain (λ), and maximum number of latent causes (K), where g is not explicitly described in the original model (Gershman et al., 2017) while this is implemented in their source code (https://github.com/sjgershm/memory-modification.git). They were estimated through the slice sampling algorithm in the Markov chain Monte Carlo method (Neal, 2003). In implementation, we used the MATLAB function, *slicesample*. The initial guess and upper and lower bounds of each parameter were listed in Appendix 2 – table 1. The initial size of the sampling window for each parameter was the difference between the upper and lower bounds of the parameter divided by 50. In each sampling step, a set of parameters was proposed, and then a CR trace was simulated in the virtual reinstatement experiment. For each parameter, one sample was drawn for every 100 steps in 200000 times sampling, then the first 1000 samples were discarded as burn-in, and eventually 1000 samples were left. The convergence of the sample chain was determined if the Gelman-Rubin statistic was smaller than 1.1. The expected value of each parameter was considered to be the maximum bin in the sample histogram, for which the bin width was statistically optimized by the method proposed by Shimazaki & Shinomoto (2007). As max. no. of iteration and K are integers, each expected value was rounded toward infinity. Given the expected parameters, the likelihood of squared errors between the simulated and the observed CR trace was evaluated on a half-normal distribution ($\mu = 0$, $\sigma = 0.3$) so that increasing error decreased the likelihood. If the likelihoods of errors were smaller than the thresholds (1 for acquisition trials; 0.25 for tests 1, 2, and 3 trials; 0.5 for remaining trials) in all trials and the Gelman-Rubin statistic < 1.1, the sampling procedure above was terminated. Otherwise, it was continued while keeping the sample chain and setting the initial guess as the last value of the sampling chain. If the procedure was not terminated within 20 iterations, the observed CR trace was classified as an anomaly. The code described above is freely available online at https://github.com/suzuki-yusuke/comp_modeling_lcm.git.

Given the estimated parameters for each mouse, their internal state and CR were simulated for every trial. The internal states consisted of latent causes, their posterior probabilities, and associative weight. The latent causes were categorized based on which phase they were initially inferred: acquisition trials, extinction trials (i.e., test 1, extinction, test 2), and trials after extinction (i.e., the unsignaled shock and test 3).

828 **Immunohistochemistry**

829 After the designated behavioral test, mice were anesthetized by intraperitoneal injection with a
830 cocktail of 0.3 mg/kg medetomidine (Nippon Zenyaku Kogyo, Koriyama, Japan), 4.0 mg/kg
831 midazolam (Sandoz, Minato, Japan), and 5.0 mg/kg butorphanol (Meiji Seika Pharma, Chuo, Japan)
832 and transcardially perfused phosphate buffer saline (PBS) followed by 4% paraformaldehyde (PFA).
833 Brain samples were preserved in 4% PFA overnight, and the solution was replaced with 30% sucrose
834 in PBS. Then, the sample was embedded in the OCT compound (Tissue TEK, Sakura finetek Japan,
835 Chuo, Japan) and sectioned at 30 μm (CM1950, Leica Biosystems, Tokyo, Japan). Every six sections
836 were collected. Ten coronal brain sections from the anterior part (AP = +2.2 ~ +1 mm from Bregma)
837 and the posterior part (AP = -1.5~-2.5 mm from Bregma) were subjected to staining in a 12-well plate.
838 Sections were washed in PBS with 0.3% Triton X-100 (PBST) to remove excessive OCT compound,
839 then blocked in 5% normal donkey serum (NDS) in PBST for an hour at room temperature and
840 incubated with primary antibodies on a shaker at 4°C overnight. The following day, sections were
841 washed in PBST and reacted with second antibodies and DAPI for 2 hours at room temperature. Finally,
842 the sections were washed in PBST, mounted on slides, and coverslipped with antifade reagent
843 (FluoroMount-G®, SoutherBiotech, Cat#0100-01). Primary antibodies, including mouse anti-β-
844 Amyloid (1:1000, Biolegend, Cat#803001, RRID:AB_2564653), rabbit anti-Iba1 (1:1000, FUJIFILM
845 Wako Pure Chemical Corporation, Cat#019-19741, RRID:AB_839504), and rat anti-GFAP (1:1000,
846 Invitrogen, Cat#13-0300, RRID:AB_2532994), were prepared in PBST with 1% NDS. Second
847 antibodies against mouse Alexa Fluor-568nm (Invitrogen, Cat#A10037, RRID:AB_11180865), rabbit
848 Alexa Fluor-647nm (Invitrogen, Cat#A31573, RRID:AB_2536183), and rat Alexa Fluor-488nm
849 (Invitrogen, Cat#A21208, RRID:AB_2535794) were used in a dilution of 1:500 in PBST with 1%
850 NDS. Fluorescent immunohistochemistry slides were imaged on a fluorescence microscope (BZ-X800,
851 Keyence, Japan) with a 10x objective, and the stitching function was used to cover the entire section.
852 Images were stitched in built-in analysis software (Keyence, Japan). The Aβ plaque area of each brain
853 was quantified under QUINT workflow (Yates et al., 2019). First, a series of brain section images were
854 registered to Allen Brain Atlas CCFv3 on QuickNII (Puchades et al., 2019) and fine-adjusted on
855 VisuAlign (https://www.nitrc.org/projects/visualign). Next, the Aβ staining channel images were
856 segmented in Ilastik (Berg et al., 2019). Finally, the integration of registration and segmented results
857 was performed in Nutil (Groeneboom et al., 2020) to quantify Aβ plaque in selected brain regions. The
858 Aβ plaque area (%) is calculated as the percentage of positive Aβ staining area covering the whole
859 coronal brain section.
860

861 **Statistical analysis**

862 For the fear conditioning, extinction, and reinstatement results, samples with freezing rates in test
863 phases exceeding the 1.5 times interquartile range of the group were treated as outliers and excluded
864 for further statistical analysis. The one-way ANCOVA was performed to analyze genotype effects on
865 the freezing rate and DI, with age as a covariate. The Student's *t*-test was performed to compare the
866 freezing rate and the DI between control and $App^{NL\text{-}G\text{-}F}$ mice within the same age. To test whether the
867 DI differs from 0.5, the one-sample *t*-test was performed within the group. To compare the estimated
868 parameters and properties of latent causes between control and $App^{NL\text{-}G\text{-}F}$ mice within the same age,
869 the Mann-Whitney *U* test was performed, as the Shapiro-Wilk normality test reported a violation of
870 normality assumption in the data. The Spearman's rank correlation coefficient between the estimated
871 model parameters and the DIs between test 3 and test 1 phase were calculated.

872 For the reversal Barnes maze results, a mixed-designed two-way [Genotype (control, $App^{NL\text{-}G\text{-}F}$)
873 × Day (1-6, 9-11, or 12-14)] ANOVA was performed for the number of errors, the latency, and the
874 travel distance in the initial and reversal training. A mixed-designed two-way [Genotype (control,
875 $App^{NL\text{-}G\text{-}F}$) × Hole (1-12)] ANOVA was performed for the time spent around each hole in the probe
876 test 1 and 2. If a significant interaction between genotype and hole was detected, Tukey's HSD test at
877 a specific hole between control and $App^{NL\text{-}G\text{-}F}$ was performed. The Wilcoxon rank sum test was
878 performed to compare the use of each strategy on each training day between control and $App^{NL\text{-}G\text{-}F}$
879 within the same age. To test whether the DI between probe tests 1 and 2 differs from 0.5, the Wilcoxon
880 signed-rank test was performed within the group, as the Shapiro-Wilk normality test reported a
881 violation of the normality assumption in the data. The Mann-Whitney *U* test was performed to compare
882 the DI between control and $App^{NL\text{-}G\text{-}F}$ mice within the same age.

883 All statistical analyses were done in JASP 0.18.3.0 (JASP team, 2024) and MATLAB R2022b
884 (MathWorks Inc., MA, USA). The significance level for all statistical tests was set at 0.05. The power
885 analysis table was provided in Table S21 and Table S22 by using G*Power 3.1 (Faul et al., 2009). Not
886 all of the multiple comparisons were corrected for, and the sample sizes written in the figure legends
887 do not include outliers detected through the exclusion procedure mentioned above.

888

Author Contributions

M.H.: Writing – original draft, Conceptualization, Formal analysis, Investigation, Methodology, Visualization

Y.S.: Writing – review & editing, Conceptualization, Formal analysis, Methodology, Software, Supervision

H.S.: Writing – review & editing, Resource

T.S.: Writing – review & editing, Resource

T.C.S.: Resource

I.I.: Writing – review & editing, Funding acquisition, Resource, Supervision

Conflict of interests

None declared.

Funding sources

This work was supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Early-Career Scientists JSPS 19K16292 (to Y.S.), Grant-in-Aid for Scientific Research (B) JSPS 21H02485 (to I.I.) from the Ministry of Education, Culture, Sports, Science and the Technology of Japan (MEXT); by Japan Science and Technology Agency (JST) CREST program Grants JPMJCR1921 (to I.I.) and JPMJCR1752 (to I.I.); and by the Program for Technological Innovation of Regenerative Medicine Grant 21bm0704060h0001 and 24bm1123049h0001 (to I.I.) and Brain/MINDS Grant 21dm0207090h0003 (to I.I.) and Moonshot Research & Development Program JP22zf0127007 (to I.I.) from the Japanese Agency for Medical research and Development (AMED).

Acknowledgments

We thank Adam T. Guy, Ph.D. and Shigeru Shinomoto, Ph.D. for informative comments on the manuscript. We also thank all the members of the Imayoshi lab for their support. We acknowledge ChatGPT-4 and Google Gemini for their contributions, which were selectively referenced during the Python code refactoring process for visualization and drawing mice artwork in Figure 1.

References

Allison, S. L., Fagan, A. M., Morris, J. C., & Head, D. (2016). Spatial Navigation in Preclinical Alzheimer's Disease. *Journal of Alzheimer's Disease : JAD*, *52*(1), 77–90. https://doi.org/10.3233/JAD-150855

Ally, B. A., Hussey, E. P., Ko, P. C., & Molitor, R. J. (2013). Pattern separation and pattern completion in Alzheimer's disease: Evidence of rapid forgetting in amnestic mild cognitive impairment. *Hippocampus*, *23*(12), 1246–1258. https://doi.org/10.1002/hipo.22162

Amieva, H., Le Goff, M., Millet, X., Orgogozo, J. M., Pérès, K., Barberger-Gateau, P., Jacqmin-Gadda, H., & Dartigues, J. F. (2008). Prodromal Alzheimer's disease: Successive emergence of the clinical symptoms. *Annals of Neurology*, *64*(5), 492–498. https://doi.org/10.1002/ana.21509

Arroyo-García, L. E., Isla, A. G., Andrade-Talavera, Y., Balleza-Tapia, H., Loera-Valencia, R., Alvarez-Jimenez, L., Pizzirusso, G., Tambaro, S., Nilsson, P., & Fisahn, A. (2021). Impaired spike-gamma coupling of area CA3 fast-spiking interneurons as the earliest functional impairment in the AppNL-G-F mouse model of Alzheimer's disease. *Molecular Psychiatry*, *26*(10), 5557–5567. https://doi.org/10.1038/s41380-021-01257-0

Barnes, C. A. (1979). Memory deficits associated with senescence: A neurophysiological and behavioral study in the rat. *Journal of Comparative and Physiological Psychology*, *93*(1), 74–104. https://doi.org/10.1037/h0077579

Barnes, J., Dickerson, B., Frost, C., Jiskoot, L. C., Wolk, D., & van der Flier, W. M. (2015). Alzheimer's disease first symptoms are age dependent: Evidence from the NACC dataset. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, *11*(11), 1349–1357. https://doi.org/10.1016/j.jalz.2014.12.007

Beier, K. T., Steinberg, E. E., DeLoach, K. E., Xie, S., Miyamichi, K., Schwarz, L., Gao, X. J., Kremer, E. J., Malenka, R. C., & Luo, L. (2015). Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input–Output Mapping. *Cell*, *162*(3), 622–634. https://doi.org/10.1016/j.cell.2015.07.015

Belliart-Guérin, G., & Planche, V. (2023). Mnemonic Discrimination Performance in a Memory Clinic: A Pilot Study. *Journal of Alzheimer's Disease*, *94*(4), 1527. https://doi.org/10.3233/JAD-230221

Berg, S., Kutra, D., Kroeger, T., Straehle, C. N., Kausler, B. X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., Eren, K., Cervantes, J. I., Xu, B., Beuttenmueller, F., Wolny, A., Zhang, C., Koethe, U., Hamprecht, F. A., & Kreshuk, A. (2019). ilastik: Interactive machine learning

for (bio)image analysis. *Nature Methods*, *16*(12), 1226–1232. https://doi.org/10.1038/s41592-019-0582-9

Bock, J. R., Hara, J., Fortier, D., Lee, M. D., Petersen, R. C., Shankle, W. R., & The Alzheimer's Disease Neuroimaging Initiative. (2021). Application of Digital Cognitive Biomarkers for Alzheimer's Disease: Identifying Cognitive Process Changes and Impending Cognitive Decline. *The Journal of Prevention of Alzheimer's Disease*, *8*(2), 123–126. https://doi.org/10.14283/jpad.2020.63

Bódi, N., Csibri, É., Myers, C. E., Gluck, M. A., & Kéri, S. (2009). Associative Learning, Acquired Equivalence, and Flexible Generalization of Knowledge in Mild Alzheimer Disease. *Cognitive and Behavioral Neurology*, *22*(2), 89. https://doi.org/10.1097/WNN.0b013e318192ccf0

Broadbelt, T., Mutlu-Smith, M., Carnicero-Senabre, D., Saido, T. C., Saito, T., & Wang, S.-H. (2022). Impairment in novelty-promoted memory via behavioral tagging and capture before apparent memory loss in a knock-in model of Alzheimer's disease. *Scientific Reports*, *12*, 22298. https://doi.org/10.1038/s41598-022-26113-1

Brown, V. J., & Tait, D. S. (2010). Behavioral Flexibility: Attentional Shifting, Rule Switching, and Response Reversal. In I. P. Stolerman & L. H. Price (Eds.), *Encyclopedia of Psychopharmacology* (pp. 1–7). Springer. https://doi.org/10.1007/978-3-642-27772-6_340-2

Cai, J., & Tong, Q. (2022). Anatomy and Function of Ventral Tegmental Area Glutamate Neurons. *Frontiers in Neural Circuits*, *16*, 867053. https://doi.org/10.3389/fncir.2022.867053

Cochran, A. L., & Cisler, J. M. (2019). A flexible and generalizable model of online latent-state learning. *PLOS Computational Biology*, *15*(9), e1007331. https://doi.org/10.1371/journal.pcbi.1007331

Coughlan, G., Laczó, J., Hort, J., Minihane, A.-M., & Hornberger, M. (2018). Spatial navigation deficits—Overlooked cognitive marker for preclinical Alzheimer disease? *Nature Reviews Neurology*, *14*(8), 496–506. https://doi.org/10.1038/s41582-018-0031-x

Ding, Y., Li, L., Wang, S., Cao, Y., Yang, M., Dai, Y., Lin, H., Li, J., Liu, Y., Wang, Z., Liu, W., & Tao, J. (2023). Electroacupuncture promotes neurogenesis in the dentate gyrus and improves pattern separation in an early Alzheimer's disease mouse model. *Biological Research*, *56*(1), 65. https://doi.org/10.1186/s40659-023-00472-z

Duara, R., & Barker, W. (2022). Heterogeneity in Alzheimer's Disease Diagnosis and Progression Rates: Implications for Therapeutic Trials. *Neurotherapeutics*, *19*(1), 8–25. https://doi.org/10.1007/s13311-022-01185-z

Emre, C., Arroyo-García, L. E., Do, K. V., Jun, B., Ohshima, M., Alcalde, S. G., Cothern, M. L., Maioli, S., Nilsson, P., Hjorth, E., Fisahn, A., Bazan, N. G., & Schultzberg, M. (2022).

Intranasal delivery of pro-resolving lipid mediators rescues memory and gamma oscillation impairment in AppNL-G-F/NL-G-F mice. *Communications Biology*, *5*, 245. https://doi.org/10.1038/s42003-022-03169-3

Erdmann, T., & Mathys, C. (2022). A generative framework for the study of delusions. *Schizophrenia Research*, *245*, 42–49. https://doi.org/10.1016/j.schres.2020.11.048

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Finke, K., Myers, N., Bublak, P., & Sorg, C. (2013). A biased competition account of attention and memory in Alzheimer's disease. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1628), 20130062. https://doi.org/10.1098/rstb.2013.0062

Gawel, K., Gibula, E., Marszalek-Grabska, M., Filarowska, J., & Kotlinska, J. H. (2019). Assessment of spatial learning and memory in the Barnes maze task in rodents—Methodological consideration. *Naunyn-Schmiedeberg's Archives of Pharmacology*, *392*(1), 1–18. https://doi.org/10.1007/s00210-018-1589-y

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1), 1–12. https://doi.org/10.1016/j.jmp.2011.08.004

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197–209. https://doi.org/10.1037/a0017808

Gershman, S. J., & Hartley, C. A. (2015). Individual differences in learning predict the return of fear. *Learning & Behavior*, *43*(3), 243–250. https://doi.org/10.3758/s13420-015-0176-z

Gershman, S. J., Monfils, M.-H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *eLife*, *6*, e23763. https://doi.org/10.7554/eLife.23763

Grober, E., Veroff, A. E., & Lipton, R. B. (2018). Temporal unfolding of declining episodic memory on the Free and Cued Selective Reminding Test in the predementia phase of Alzheimer's disease: Implications for clinical trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *10*(1), 161–171. https://doi.org/10.1016/j.dadm.2017.12.004

Groeneboom, N. E., Yates, S. C., Puchades, M. A., & Bjaalie, J. G. (2020). Nutil: A Pre- and Post-processing Toolbox for Histological Rodent Brain Section Images. *Frontiers in Neuroinformatics*, *14*. https://doi.org/10.3389/fninf.2020.00037

Han, Y., Zhang, Y., Kim, H., Grayson, V. S., Jovasevic, V., Ren, W., Centeno, M. V., Guedea, A. L., Meyer, M. A. A., Wu, Y., Gutruf, P., Surmeier, D. J., Gao, C., Martina, M., Apkarian, A. V., Rogers, J. A., & Radulovic, J. (2020). Excitatory VTA to DH projections provide a valence

signal to memory circuits. *Nature Communications*, *11*(1), 1466. https://doi.org/10.1038/s41467-020-15035-z

Hashimoto, S., Matsuba, Y., Kamano, N., Mihira, N., Sahara, N., Takano, J., Muramatsu, S., Saido, T. C., & Saito, T. (2019). Tau binding protein CAPON induces tau aggregation and neurodegeneration. *Nature Communications*, *10*(1), 2394. https://doi.org/10.1038/s41467-019-10278-x

Hauser, T. U., Skvortsova, V., Choudhury, M. D., & Koutsouleris, N. (2022). The promise of a model-based psychiatry: Building computational models of mental ill health. *The Lancet Digital Health*, *4*(11), e816–e828. https://doi.org/10.1016/S2589-7500(22)00152-2

Hermans, D., Dirikx, T., Vansteenwegenin, D., Baeyens, F., Van den Bergh, O., & Eelen, P. (2005). Reinstatement of fear responses in human aversive conditioning. *Behaviour Research and Therapy*, *43*(4), 533–551. https://doi.org/10.1016/j.brat.2004.03.013

Hitora-Imamura, N., Miura, Y., Teshirogi, C., Ikegaya, Y., Matsuki, N., & Nomura, H. (2015). Prefrontal dopamine regulates fear reinstatement through the downregulation of extinction circuits. *eLife*, *4*, e08274. https://doi.org/10.7554/eLife.08274

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

Hwang, U., Kim, S.-W., Jung, D., Kim, S., Lee, H., Seo, S. W., Seong, J.-K., & Yoon, S. (2023). Real-world prediction of preclinical Alzheimer's disease with a deep generative model. *Artificial Intelligence in Medicine*, *144*, 102654. https://doi.org/10.1016/j.artmed.2023.102654

Inayat, S., McAllister, B. B., Chang, H., Lacoursiere, S. G., Whishaw, I. Q., Sutherland, R. J., & Mohajerani, M. H. (2023). Weak-hyperactive hippocampal CA1 neurons in the prodromal stage of Alzheimer's disease in hybrid *App*NL-G-F/NL-G-F × *Thy1*−GCaMP6s+/− mice suggest disrupted plasticity. *Neurobiology of Aging*, *130*, 154–171. https://doi.org/10.1016/j.neurobiolaging.2023.06.002

Jack Jr., C. R., Andrews, J. S., Beach, T. G., Buracchio, T., Dunn, B., Graf, A., Hansson, O., Ho, C., Jagust, W., McDade, E., Molinuevo, J. L., Okonkwo, O. C., Pani, L., Rafii, M. S., Scheltens, P., Siemers, E., Snyder, H. M., Sperling, R., Teunissen, C. E., & Carrillo, M. C. (2024). Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimer's & Dementia*, *20*(8), 5143–5169. https://doi.org/10.1002/alz.13859

Jack Jr., C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T., Phelps, C.,

Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M., … Silverberg, N. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535–562. https://doi.org/10.1016/j.jalz.2018.02.018

Jankowsky, J. L., & Zheng, H. (2017). Practical considerations for choosing a mouse model of Alzheimer's disease. *Molecular Neurodegeneration*, *12*(1), 89. https://doi.org/10.1186/s13024-017-0231-7

Jun, H., Bramian, A., Soma, S., Saito, T., Saido, T. C., & Igarashi, K. M. (2020). Disrupted Place Cell Remapping and Impaired Grid Cells in a Knockin Model of Alzheimer's Disease. *Neuron*, *107*(6), 1095-1112.e6. https://doi.org/10.1016/j.neuron.2020.06.023

Jutten, R. J., Papp, K. V., Hendrix, S., Ellison, N., Langbaum, J. B., Donohue, M. C., Hassenstab, J., Maruff, P., Rentz, D. M., Harrison, J., Cummings, J., Scheltens, P., & Sikkes, S. A. M. (2023). Why a clinical trial is as good as its outcome measure: A framework for the selection and use of cognitive outcome measures for clinical trials of Alzheimer's disease. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, *19*(2), 708–720. https://doi.org/10.1002/alz.12773

Kirova, A.-M., Bays, R. B., & Lagalwar, S. (2015). Working Memory and Executive Function Decline across Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease. *BioMed Research International*, *2015*(1), 748212. https://doi.org/10.1155/2015/748212

Knopman, D. S., Amieva, H., Petersen, R. C., Chételat, G., Holtzman, D. M., Hyman, B. T., Nixon, R. A., & Jones, D. T. (2021). Alzheimer disease. *Nature Reviews Disease Primers*, *7*(1), 1–21. https://doi.org/10.1038/s41572-021-00269-y

Knowlton, B. J., & Castel, A. D. (2022). Memory and Reward-Based Learning: A Value-Directed Remembering Perspective. *Annual Review of Psychology*, *73*(Volume 73, 2022), 25–52. https://doi.org/10.1146/annurev-psych-032921-050951

Kocagoncu, E., Klimovich-Gray, A., Hughes, L. E., & Rowe, J. B. (2021). Evidence and implications of abnormal predictive coding in dementia. *Brain*, *144*(11), 3311. https://doi.org/10.1093/brain/awab254

Kumfor, F., Liang, C. T., Hazelton, J. L., Leyton, C. E., Kaizik, C., Devenney, E., Connaughton, E., Langdon, R., Mioshi, E., Kwok, J. B., Dobson-Stone, C., Halliday, G. M., Piguet, O., Hodges, J. R., & Landin-Romero, R. (2022). Examining the presence and nature of delusions in Alzheimer's disease and frontotemporal dementia syndromes. *International Journal of Geriatric Psychiatry*, *37*(3), 10.1002/gps.5692. https://doi.org/10.1002/gps.5692

Kundu, P., Torres, E. R. S., Stagaman, K., Kasschau, K., Okhovat, M., Holden, S., Ward, S., Nevonen, K. A., Davis, B. A., Saito, T., Saido, T. C., Carbone, L., Sharpton, T. J., & Raber, J. (2021).

Integrated analysis of behavioral, epigenetic, and gut microbiome analyses in AppNL-G-F, AppNL-F, and wild type mice. *Scientific Reports*, *11*(1), 4678. https://doi.org/10.1038/s41598-021-83851-4

Laczó, M., Lerch, O., Martinkovic, L., Kalinova, J., Markova, H., Vyhnalek, M., Hort, J., & Laczó, J. (2021). Spatial Pattern Separation Testing Differentiates Alzheimer's Disease Biomarker-Positive and Biomarker-Negative Older Adults With Amnestic Mild Cognitive Impairment. *Frontiers in Aging Neuroscience*, *13*. https://doi.org/10.3389/fnagi.2021.774600

Latif-Hernandez, A., Shah, D., Craessaerts, K., Saido, T., Saito, T., De Strooper, B., Van der Linden, A., & D'Hooge, R. (2019). Subtle behavioral changes and increased prefrontal-hippocampal network synchronicity in APPNL−G−F mice before prominent plaque deposition. *Behavioural Brain Research*, *364*, 431–441. https://doi.org/10.1016/j.bbr.2017.11.017

Leal, S. L., Ferguson, L. A., Harrison, T. M., & Jagust, W. J. (2019). Development of a mnemonic discrimination task using naturalistic stimuli with applications to aging and preclinical Alzheimer's disease. *Learning & Memory*, *26*(7), 219–228. https://doi.org/10.1101/lm.048967.118

Lee, H., Stirnberg, R., Wu, S., Wang, X., Stöcker, T., Jung, S., Montag, C., & Axmacher, N. (2020). Genetic Alzheimer's Disease Risk Affects the Neural Mechanisms of Pattern Separation in Hippocampal Subfields. *Current Biology*, *30*(21), 4201-4212.e3. https://doi.org/10.1016/j.cub.2020.08.042

Lee, M. D., Bock, J. R., Cushman, I., & Shankle, W. R. (2020). An application of multinomial processing tree models and Bayesian methods to understanding memory impairment. *Journal of Mathematical Psychology*, *95*, 102328. https://doi.org/10.1016/j.jmp.2020.102328

Lee, M. D., & Stark, C. E. L. (2023). Bayesian Modeling of the Mnemonic Similarity Task Using Multinomial Processing Trees. *Behaviormetrika*, *50*(2), 517–539. https://doi.org/10.1007/s41237-023-00193-3

Leising, K. J., & Blaisdell, A. P. (2009). Associative basis of landmark learning and integration in vertebrates. *Comparative Cognition & Behavior Reviews*, 4. https://comparative-cognition-and-behavior-reviews.org/vol4_leising_blaisdell/

Lerner, T. N., Holloway, A. L., & Seiler, J. L. (2021). Dopamine, Updated: Reward Prediction Error and Beyond. *Current Opinion in Neurobiology*, *67*, 123–130. https://doi.org/10.1016/j.conb.2020.10.012

Lisman, J. E., & Grace, A. A. (2005). The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. *Neuron*, *46*(5), 703–713. https://doi.org/10.1016/j.neuron.2005.05.002

Madarasz, T., & Behrens, T. (2019). Better Transfer Learning with Inferred Successor Maps. *Advances in Neural Information Processing Systems*, *32*. https://papers.nips.cc/paper_files/paper/2019/hash/274a10ffa06e434f2a94df765cac6bf4-Abstract.html

Malhotra, P. A. (2019). Impairments of attention in Alzheimer's disease. *Current Opinion in Psychology*, *29*, 41–48. https://doi.org/10.1016/j.copsyc.2018.11.002

Maren, S., Phan, K. L., & Liberzon, I. (2013). The contextual brain: Implications for fear conditioning, extinction and psychopathology. *Nature Reviews. Neuroscience*, *14*(6), 417–428. https://doi.org/10.1038/nrn3492

Masuda, A., Kobayashi, Y., Kogo, N., Saito, T., Saido, T. C., & Itohara, S. (2016). Cognitive deficits in single *App* knock-in mouse models. *Neurobiology of Learning and Memory*, *135*, 73–82. https://doi.org/10.1016/j.nlm.2016.07.001

Mehla, J., Lacoursiere, S. G., Lapointe, V., McNaughton, B. L., Sutherland, R. J., McDonald, R. J., & Mohajerani, M. H. (2019). Age-dependent behavioral and biochemical characterization of single APP knock-in mouse (APPNL-G-F/NL-G-F) model of Alzheimer's disease. *Neurobiology of Aging*, *75*, 25–37. https://doi.org/10.1016/j.neurobiolaging.2018.10.026

Monfils, M.-H., Cowansage, K. K., Klann, E., & LeDoux, J. E. (2009). Extinction-Reconsolidation Boundaries: Key to Persistent Attenuation of Fear Memories. *Science*, *324*(5929), 951–955. https://doi.org/10.1126/science.1167975

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Nandi, A., Counts, N., Chen, S., Seligman, B., Tortorice, D., Vigo, D., & Bloom, D. E. (2022). Global and regional projections of the economic burden of Alzheimer's disease and related dementias from 2019 to 2050: A value of statistical life approach. *eClinicalMedicine*, *51*. https://doi.org/10.1016/j.eclinm.2022.101580

Neunuebel, J. P., & Knierim, J. J. (2014). CA3 Retrieves Coherent Representations from Degraded Input: Direct Evidence for CA3 Pattern Completion and Dentate Gyrus Pattern Separation. *Neuron*, *81*(2), 416–427. https://doi.org/10.1016/j.neuron.2013.11.017

Nobili, A., Latagliata, E. C., Viscomi, M. T., Cavallucci, V., Cutuli, D., Giacovazzo, G., Krashia, P., Rizzo, F. R., Marino, R., Federici, M., De Bartolo, P., Aversa, D., Dell'Acqua, M. C., Cordella, A., Sancandi, M., Keller, F., Petrosini, L., Puglisi-Allegra, S., Mercuri, N. B., … D'Amelio, M. (2017). Dopamine neuronal loss contributes to memory and reward dysfunction in a model of Alzheimer's disease. *Nature Communications*, *8*(1), 14727. https://doi.org/10.1038/ncomms14727

Norbury, A., Brinkman, H., Kowalchyk, M., Monti, E., Pietrzak, R. H., Schiller, D., & Feder, A. (2022). Latent cause inference during extinction learning in trauma-exposed individuals with and without PTSD. *Psychological Medicine*, *52*(16), 3834–3845. https://doi.org/10.1017/S0033291721000647

Palmer, A., & Good, M. (2011). Hippocampal synaptic activity, pattern separation and episodic-like memory: Implications for mouse models of Alzheimer's disease pathology. *Biochemical Society Transactions*, *39*(4), 902–909. https://doi.org/10.1042/BST0390902

Parizkova, M., Lerch, O., Andel, R., Kalinova, J., Markova, H., Vyhnalek, M., Hort, J., & Laczó, J. (2020). Spatial Pattern Separation in Early Alzheimer's Disease. *Journal of Alzheimer's Disease: JAD*, *76*(1), 121–138. https://doi.org/10.3233/JAD-200093

Pearce, J. M. (2009). The 36th Sir Frederick Bartlett Lecture: An associative analysis of spatial learning. *Quarterly Journal of Experimental Psychology*, *62*(9), 1665–1684. https://doi.org/10.1080/17470210902805589

Petok, J. R., Myers, C. E., Pa, J., Hobel, Z., Wharton, D. M., Medina, L. D., Casado, M., Coppola, G., Gluck, M. A., & Ringman, J. M. (2018). Impairment of memory generalization in preclinical autosomal dominant Alzheimer's disease mutation carriers. *Neurobiology of Aging*, *65*, 149–157. https://doi.org/10.1016/j.neurobiolaging.2018.01.022

Petrella, J. R., Hao, W., Rao, A., & Doraiswamy, P. M. (2019). Computational Causal Modeling of the Dynamic Biomarker Cascade in Alzheimer's Disease. *Computational and Mathematical Methods in Medicine*, *2019*, e6216530. https://doi.org/10.1155/2019/6216530

Pooley, J. P., Lee, M. D., & Shankle, W. R. (2011). Understanding memory impairment with memory models and hierarchical Bayesian analysis. *Journal of Mathematical Psychology*, *55*(1), 47–56. https://doi.org/10.1016/j.jmp.2010.08.003

Puchades, M. A., Csucs, G., Ledergerber, D., Leergaard, T. B., & Bjaalie, J. G. (2019). Spatial registration of serial microscopic brain images to three-dimensional reference atlases with the QuickNII tool. *PLOS ONE*, *14*(5), e0216796. https://doi.org/10.1371/journal.pone.0216796

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory: Vol. Vol. 2*.

Saifullah, Md. A. B., Komine, O., Dong, Y., Fukumoto, K., Sobue, A., Endo, F., Saito, T., Saido, T. C., Yamanaka, K., & Mizoguchi, H. (2020). Touchscreen-based location discrimination and paired associate learning tasks detect cognitive impairment at an early stage in an App knock-in mouse model of Alzheimer's disease. *Molecular Brain*, *13*, 147. https://doi.org/10.1186/s13041-020-00690-6

Saito, T., Matsuba, Y., Mihira, N., Takano, J., Nilsson, P., Itohara, S., Iwata, N., & Saido, T. C. (2014). Single App knock-in mouse models of Alzheimer's disease. *Nature Neuroscience*, *17*(5), Article 5. https://doi.org/10.1038/nn.3697

Saito, T., Mihira, N., Matsuba, Y., Sasaguri, H., Hashimoto, S., Narasimhan, S., Zhang, B., Murayama, S., Higuchi, M., Lee, V. M. Y., Trojanowski, J. Q., & Saido, T. C. (2019). Humanization of the entire murine Mapt gene provides a murine model of pathological human tau propagation. *Journal of Biological Chemistry*, *294*(34), 12754–12765. https://doi.org/10.1074/jbc.RA119.009487

Sakakibara, Y., Sekiya, M., Saito, T., Saido, T. C., & Iijima, K. M. (2018). Cognitive and emotional alterations in App knock-in mouse models of Aβ amyloidosis. *BMC Neuroscience*, *19*, 46. https://doi.org/10.1186/s12868-018-0446-8

Sanders, H., Wilson, M. A., & Gershman, S. J. (2020). Hippocampal remapping as hidden state inference. *eLife*, *9*, e51140. https://doi.org/10.7554/eLife.51140

Sasaguri, H., Nilsson, P., Hashimoto, S., Nagata, K., Saito, T., De Strooper, B., Hardy, J., Vassar, R., Winblad, B., & Saido, T. C. (2017). APP mouse models for Alzheimer's disease preclinical studies. *The EMBO Journal*, *36*(17), 2473–2487. https://doi.org/10.15252/embj.201797397

Shimazaki, H., & Shinomoto, S. (2007). A Method for Selecting the Bin Size of a Time Histogram. *Neural Computation*, *19*(6), 1503–1527. https://doi.org/10.1162/neco.2007.19.6.1503

Snyder, P. J., Kahle-Wrobleski, K., Brannan, S., Miller, D. S., Schindler, R. J., DeSanti, S., Ryan, J. M., Morrison, G., Grundman, M., Chandler, J., Caselli, R. J., Isaac, M., Bain, L., & Carrillo, M. C. (2014). Assessing cognition and function in Alzheimer's disease clinical trials: Do we have the right tools? *Alzheimer's & Dementia*, *10*(6), 853–860. https://doi.org/10.1016/j.jalz.2014.07.158

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jack Jr., C. R., Kaye, J., Montine, T. J., Park, D. C., Reiman, E. M., Rowe, C. C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M. C., Thies, B., Morrison-Bogorad, M., … Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 280–292. https://doi.org/10.1016/j.jalz.2011.03.003

Sperling, R. A., Mormino, E., & Johnson, K. (2014). The Evolution of Preclinical Alzheimer's Disease: Implications for Prevention Trials. *Neuron*, *84*(3), 608–622. https://doi.org/10.1016/j.neuron.2014.10.038

Spoleti, E., La Barbera, L., Cauzzi, E., De Paolis, M. L., Saba, L., Marino, R., Sciamanna, G., Di Lazzaro, V., Keller, F., Nobili, A., Krashia, P., & D'Amelio, M. (2024). Dopamine neuron degeneration in the Ventral Tegmental Area causes hippocampal hyperexcitability in experimental Alzheimer's Disease. *Molecular Psychiatry*, 1–16. https://doi.org/10.1038/s41380-024-02408-9

Sutter, K., Wijdenes, L. O., Beers, R. J. van, Claassen, J. A. H. R., Kessels, R. P. C., & Medendorp, W. P. (2024). Early-Stage Alzheimer's Disease Affects Fast But Not Slow Adaptive Processes in Motor Learning. *eNeuro*, *11*(6). https://doi.org/10.1523/ENEURO.0108-24.2024

Suzuki, Y., & Imayoshi, I. (2017). Network analysis of exploratory behaviors of mice in a spatial learning and memory task. *PLoS ONE*, *12*(7), e0180789. https://doi.org/10.1371/journal.pone.0180789

Tachiki, Y., Suzuki, Y., Kurahashi, M., Oki, K., Mavuk, Ö., Nakagawa, T., Ishihara, S., Gyoten, Y., Yamamoto, A., & Imayoshi, I. (2023). Scale Space Calibrates Present and Subsequent Spatial Learning in Barnes Maze in Mice. *eNeuro*, *10*(6). https://doi.org/10.1523/ENEURO.0505-22.2023

Vanderlip, C. R., Lee, M. D., & Stark, C. E. L. (2024). Cognitive modeling of the Mnemonic Similarity Task as a digital biomarker for Alzheimer's disease. *Alzheimer's & Dementia*, *1*(13). https://doi.org/10.1002/alz.14163

Webster, S. J., Bachstetter, A. D., Nelson, P. T., Schmitt, F. A., & Van Eldik, L. J. (2014). Using mice to model Alzheimer's dementia: An overview of the clinical disease and the preclinical behavioral changes in 10 mouse models. *Frontiers in Genetics*, *5*. https://doi.org/10.3389/fgene.2014.00088

Wesnes, K. A., Annas, P., Basun, H., Edgar, C., & Blennow, K. (2014). Performance on a pattern separation task by Alzheimer's patients shows possible links between disrupted dentate gyrus activity and apolipoprotein E ∈4 status and cerebrospinal fluid amyloid-β42 levels. *Alzheimer's Research & Therapy*, *6*(2), 20. https://doi.org/10.1186/alzrt250

Wessa, M., King, A. V., Meyer, P., Frölich, L., Flor, H., Poupon, C., Hoppstädter, M., & Linke, J. (2016). Impaired and preserved aspects of feedback learning in aMCI: Contributions of structural connectivity. *Brain Structure and Function*, *221*(5), 2831–2846. https://doi.org/10.1007/s00429-015-1075-y

Whyte, L. S., Hemsley, K. M., Lau, A. A., Hassiotis, S., Saito, T., Saido, T. C., Hopwood, J. J., & Sargeant, T. J. (2018). Reduction in open field activity in the absence of memory deficits in the AppNL−G−F knock-in mouse model of Alzheimer's disease. *Behavioural Brain Research*, *336*, 177–181. https://doi.org/10.1016/j.bbr.2017.09.006

Wilson, I. A., Gallagher, M., Eichenbaum, H., & Tanila, H. (2006). Neurocognitive aging: Prior memories hinder new hippocampal encoding. *Trends in Neurosciences*, *29*(12), 662–670. https://doi.org/10.1016/j.tins.2006.10.002

Yada, Y., & Honda, N. (2023). Few-shot prediction of amyloid β accumulation from mainly unpaired data on biomarker candidates. *NPJ Systems Biology and Applications*, *9*(1), 59. https://doi.org/10.1038/s41540-023-00321-5

Yassa, M. A., Stark, S. M., Bakker, A., Albert, M. S., Gallagher, M., & Stark, C. E. L. (2010). High-resolution structural and functional MRI of hippocampal CA3 and dentate gyrus in patients with amnestic Mild Cognitive Impairment. *NeuroImage*, *51*(3), 1242–1252. https://doi.org/10.1016/j.neuroimage.2010.03.040

Yates, S. C., Groeneboom, N. E., Coello, C., Lichtenthaler, S. F., Kuhn, P.-H., Demuth, H.-U., Hartlage-Rübsamen, M., Roßner, S., Leergaard, T., Kreshuk, A., Puchades, M. A., & Bjaalie, J. G. (2019). QUINT: Workflow for Quantification and Spatial Analysis of Features in Histological Images From Rodent Brain. *Frontiers in Neuroinformatics*, *13*. https://doi.org/10.3389/fninf.2019.00075

Zhong, M. Z., Peng, T., Duarte, M. L., Wang, M., & Cai, D. (2024). Updates on mouse models of Alzheimer's disease. *Molecular Neurodegeneration*, *19*(1), 23. https://doi.org/10.1186/s13024-024-00712-0

Zhu, H., Yan, H., Tang, N., Li, X., Pang, P., Li, H., Chen, W., Guo, Y., Shu, S., Cai, Y., Pei, L., Liu, D., Luo, M.-H., Man, H., Tian, Q., Mu, Y., Zhu, L.-Q., & Lu, Y. (2017). Impairments of spatial memory in an Alzheimer's disease model via degeneration of hippocampal cholinergic synapses. *Nature Communications*, *8*(1), 1676. https://doi.org/10.1038/s41467-017-01943-0

Figure legends

Figure 1. Internal state in the latent cause model in the Pavlovian fear conditioning, extinction, and reinstatement.

(A) Schematic diagram of latent cause framework in a simple Pavlovian fear conditioning paradigm. An experimental mouse observes a tone, a context, and an electrical shock, which are set down in a parallelogram. Among the stimuli, the tone is designed as the conditioned stimulus (CS), and the electrical shock is designed as the unconditioned stimulus (US) with additional consideration of context effect. The mouse has an internal model such that a latent cause z generates the observation where the tone and the context induce the shock at an associative weight w_{cs} and $w_{context}$, respectively. The latent cause is a latent variable represented by a white square. The posterior probability of z is represented by a black arrow from z to the observation. The stimuli (i.e., tone, context, and shock) are observed variables represented in shaded circles, and the associative weight w is represented by a black arrow between the two. Since latent causes are unobservable, the mouse infers which latent cause is more likely from posterior distribution over latent causes given the observation following Bayes' rule. **(B)** Schematic diagram of internal state posited in the latent cause model demonstrating the memory modification process in the fear conditioning paradigm. The mouse has acquired z_A through two observations of the CS accompanied by the US in the same context. Now, the mouse observes the tone alone, then infers a new latent cause z_B generating the CS and context without US as w_{cs} and $w_{context}$ equal to 0. The internal state at the time consists of z_A with largely decreased posterior probability and thereby slightly decreased w_{cs} and $w_{context}$, in addition to z_B. The mouse in this panel is likely to differentiate the current observation from z_A, as it assigns a higher probability for z_B. The posterior probability of z_A and z_B is represented by the relative size of their diagram. If the mouse generalizes z_A to the current observation without inferring z_B, it would largely decrease w_{cs} and $w_{context}$ of z_A to minimize prediction error. **(C)** Observed and simulated conditioned response (CR) during reinstatement in fear conditioning. In the acquisition phase, a CS was accompanied by a US, inducing associative fear memory in the animal, as it exhibits a higher CR to the CS. In the extinction phase, the CS was presented without the US, inducing extinction memory with decreased CR. After an unsignaled shock without the preceding CS, the CR increases again, and this is called reinstatement. The observed CR was sampled from a 2-month-old male wild-type C57BL/6J mouse (magenta line plot with square markers), where the vertical and horizontal axis indicate CR and cumulative trial, respectively. The latent cause model parameters were estimated so as to minimize the prediction error between the observed and the simulated CR given the parameter (gray line with square markers). The estimated parameter value: $\alpha = 2.4$, $g = 0.316$, $\eta = 0.64$, max. no. of iteration = 9, $w_0 = -0.006$, $\sigma_r^2 =$

4.50, $\sigma_x^2 = 4.39$, $\theta = 0.03$, $\lambda = 0.02$, $K = 6$. Trials with red, blue, and yellow backgrounds correspond to the acquisition, extinction, and unsignaled shock, where the first and second extinction consists of 19 and 10 trials, respectively (see Material and Methods). Three trials with white backgrounds are tests 1, 2, and 3 to evaluate if conditioning, extinction, and reinstatement are established. The vertical dashed lines separate the phases. **(D)** The evolution of the posterior probability and **(E)** the associative weight of each latent cause across the reinstatement experiment, where w_{cs} (top panel) and $w_{context}$ (bottom panel) were computed. **(D and E)** Each latent cause is indicated by a unique subscript number and color, w_{cs} and $w_{context}$. The latent causes acquired during the acquisition phase and their associative weight are termed $\mathbf{z_A}$ and $\mathbf{w_A}$, respectively. Meanwhile, latent causes acquired during the extinction phase and their associative weight are termed $\mathbf{z_B}$ and $\mathbf{w_B}$, respectively.

Figure 2. *App*$^{NL\text{-}G\text{-}F}$ mice exhibited successful associative learning and extinction but a lower extent of reinstatement.

(A) Schematic diagram of reinstatement paradigm in this study. In the acquisition phase, the CS was accompanied with the US 3 times. In the following phase, either the CS or the US was presented at certain times. The context was the same throughout the experiment. **(B)** Freezing rate during CS presentation across reinstatement paradigm. The markers and lines show the median freezing rate of each group in each trial. Red, blue, and yellow backgrounds represent acquisition, extinction, and unsignaled shock in **(A)**. The dashed vertical line separates the extinction 1 and extinction 2 phases. Freezing rate during CS presentation in test 1 **(C)**, test 2 **(D)**, and test 3 **(E)**. Discrimination index (DI) between test 2 and test 1 **(F)**, between test 3 and test 2 **(G)**, and between test 3 and test 1 **(H)** calculated from freezing rate during CS presentation.

(C to E) The data are shown as median with interquartile range. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$ by one-way ANCOVA with age as covariate; $^{\#}p < 0.05$, $^{\#\#}p < 0.01$, and $^{\#\#\#}p < 0.001$ by Student's *t*-test comparing control and *App*$^{NL\text{-}G\text{-}F}$ mice within the same age. Detailed statistical results are provided in **Table S2**. **(F to H)** The dashed horizontal line indicates DI = 0.5, which means no discrimination between the two phases. $^{\dagger}p < 0.05$ by one-sample Student's *t*-test, and the alternative hypothesis specifies that the mean differs from 0.5; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$ by one-way ANCOVA with age as a covariate; $^{\#}p < 0.05$, $^{\#\#}p < 0.01$, and $^{\#\#\#}p < 0.001$ by Student's *t*-test comparing control and *App*$^{NL\text{-}G\text{-}F}$ mice within the same age. Detailed statistical results are provided in **Tables S3 and S4**. **(B to H)** Colors indicate different groups: orange represents 6-month-old control (n = 24), light blue represents 6-month-old *App*$^{NL\text{-}G\text{-}F}$ mice (n = 25), pink represents 12-month-old control (n = 24), and dark blue represents 12-month-old *App*$^{NL\text{-}G\text{-}F}$ mice (n = 25). Each black dot represents one animal.

1344 **Figure 2 – figure supplement 1.** Representative image of immunofluorescence staining using Iba1
1345 (blue), GFAP (green), and Aβ (red) antibodies on anterior **(A)** and posterior **(B)** coronal brain sections.
1346 Quantification results of Aβ plaque area in anterior **(C)** and posterior **(D)** coronal brain section of
1347 $App^{NL\text{-}G\text{-}F}$ mice.
1348 **(A and B)** The scale bar on the images in the first row represents 1 mm. The images in the second row
1349 of the panel are the magnified view of the inlet in the same column, where the scale bar on the images
1350 represents 150 μm in panel A and 250 μm in panel B. **(C and D)** The data are shown as mean with
1351 95% confidence interval. Each black dot represents one animal. $n = 5$ in 6-month-old $App^{NL\text{-}G\text{-}F}$ mice;
1352 $n = 10$ in 12-month-old $App^{NL\text{-}G\text{-}F}$ mice. Student's *t*-test was performed to compare 6- and 12-month-
1353 old $App^{NL\text{-}G\text{-}F}$ mice. The results were $t(14)=2.555$, $p = 0.023$* in **(C)** and $t(14)=3.536$, $p = 0.003$** in
1354 **(D)**.
1355
1356 **Figure 2 – figure supplement 2. Freezing rate before tone presentation in the reinstatement**
1357 **paradigm**
1358 Freezing rate before and during tone presentation across reinstatement paradigm in the 6-month-old
1359 group **(A)** and 12-month-old group **(B)**. The markers and lines show the median of each group in each
1360 trial. Dashed and solid lines indicate the freezing rate for 1 min before the CS presentation (preCS)
1361 and that for 1 min during the tone presentation (CS), respectively. Red, blue, and yellow backgrounds
1362 represent acquisition, extinction, and unsignaled shock as in Figure 2A. The dashed vertical line
1363 separates the extinction 1 and extinction 2 phases. Freezing rate before CS presentation in test 1 **(C)**,
1364 test 2 **(D)**, and test 3 **(E)**. The data are shown as median with interquartile range. Each black dot
1365 represents one animal. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$ by one-way ANCOVA with age as
1366 covariate; $^{\#}p < 0.05$, $^{\#\#}p < 0.01$, and $^{\#\#\#}p < 0.001$ by Student's *t*-test comparing control and $App^{NL\text{-}G\text{-}F}$
1367 mice within the same age.
1368 **(A to E)** Colors indicate different groups: orange represents 6-month-old control ($n = 24$), light blue
1369 represents 6-month-old $App^{NL\text{-}G\text{-}F}$ mice ($n = 25$), pink represents 12-month-old control ($n = 24$), and
1370 dark blue represents 12-month-old $App^{NL\text{-}G\text{-}F}$ mice ($n = 25$).

Figure 3. The divergence of internal states between control and *App*$^{NL\text{-}G\text{-}F}$ mice differed in age.

(A) Simulation of reinstatement in the 6-month-old control mice given a set of estimated parameters of the latent cause model. The estimated parameter value: α = 2.4, g = 0.05, η = 0.42, max. no. of iteration = 2, w_0 = 0.009, σ_r^2 = 2.90, σ_x^2 = 0.49, θ = 0.008, λ = 0.018, K = 10. **(B)** Simulation of reinstatement in 6-month-old *App*$^{NL\text{-}G\text{-}F}$ mice. The estimated parameter value: α = 1.1, g = 0.61, η = 0.51, max. no. of iteration = 2, w_0 = 0.004, σ_r^2 = 1.51, σ_x^2 = 0.32, θ = 0.015, λ = 0.019, K = 30. **(C)** Simulation in 12-month-old control mice. The estimated parameter value: α = 1.1, g = 0.92, η = 0.82, max. no. of iteration = 5, w_0 = 0.009, σ_r^2 = 1.88, σ_x^2 = 0.51, θ = 0.010, λ = 0.017, K = 4. **(D)** Simulation in 12-month-old *App*$^{NL\text{-}G\text{-}F}$ mice. The estimated parameter value for: α = 1.0, g = 1.10, η = 0.04, max. no. of iteration = 2, w_0 = 0.0025, σ_r^2 = 1.51, σ_x^2 = 0.18, θ = 0.011, λ = 0.011, K = 36.

(A to D) The first row shows the trace of observed CR and simulated CR. The observed CR is the median freezing rate during the CS presentation over the mice within each group; the observed CR of each group was divided by its maximum over all trials. The second row shows the posterior probability of each latent cause in each trial. The third and fourth rows show the associative weight of tone to shock (w_{cs}) and that of context to shock ($w_{context}$) in each trial. Each marker and color corresponds to a latent cause up to 5. Each latent cause is represented by the same color as that in the second row and contains w_{cs} and $w_{context}$. The vertical dashed lines indicate the boundaries of phases.

Figure 3 – figure supplement 1. Replication of DI in the test phases and prediction error across trials in the latent cause model.

Discrimination index (DI) between test 2 and test 1 **(A)**, between test 3 and test 2 **(B)**, and between test 3 and test 1 **(C)** calculated from simulated CR. The signed prediction error between observed and simulated CR in the 6-month-old group **(D)** and 12-month-old group **(E)**.

(A to C) Each black dot represents one animal. The dashed horizontal line indicates DI = 0.5, which means no discrimination between the two phases. $^{\dagger}p < 0.05$ by one-sample Student's *t*-test, and the alternative hypothesis specifies that the mean differs from 0.5; $^{*}p < 0.05$, and $^{**}p < 0.01$ by one-way ANCOVA with age as a covariate; $^{\#}p < 0.05$ by Student's *t*-test comparing control and *App*$^{NL\text{-}G\text{-}F}$ mice within the same age. Detailed statistical results are provided in **Tables S5 and S6**. **(D and E)** The gray areas indicate the test phases. $^{*}p < 0.05$ by Mann-Whitney U test comparing control and *App*$^{NL\text{-}G\text{-}F}$ mice within the same age. Statistical results are as follows: U = 93.5, p = 0.883 for test 1, U = 110, p = 0.347 for test 2, U = 67, p = 0.277 for test 3 in the 6-month-old group; U = 229, p = 0.156 for test 1, U = 167, p = 0.718 for test 2, U = 101, p = 0.02 for test 3 in the 12-month-old group. **(A to E)** Colors in the boxplot indicate different groups: orange represents 6-month-old control (n = 15), light blue represents 6-month-old *App*$^{NL\text{-}G\text{-}F}$ mice (n = 12), pink represents 12-month-old control (n = 20), and dark blue represents 12-month-old *App*$^{NL\text{-}G\text{-}F}$ mice (n = 18).

1405 **Figure 4. Individual parameter estimation and internal state in the 12-month-old group.**
1406 **(A)** The estimated latent cause model parameter. **(B)** Correlation between discrimination index (DI)
1407 and estimated parameter values in the 12-month-old group. The count of latent causes initially inferred
1408 during the acquisition trials **(C, left)**, extinction trials (i.e., test 1, extinction, test 2) **(D, left)**, and trials
1409 after extinction (i.e., the unsignaled shock and test 3) **(E, left)**, with the sum of posterior probabilities
1410 **(C, D, E, right)**, and the sum of associative weights at test 3 in these latent causes **(F, G, H)**.
1411 **(A)** Each black dot represents one animal. $*p < 0.05$, and $**p < 0.01$ by the Mann-Whitney U test. **(B)**
1412 The y-axis shows the DI between test 3 and test 1. Spearman's correlation coefficient (ρ) was labeled
1413 with significance, where $*p < 0.05$ and $***p < 0.001$. The blue line represents the linear regression
1414 model fit, and the shaded area indicates the confidence interval. Each dot represents one animal. **(C to**
1415 **E)** In the first column, the histogram of the number of latent causes (z) acquired in each phase is shown.
1416 The maximum number of latent causes that can be inferred is 3, 31, and 2 in panels C, D, and E. In the
1417 second column, the histogram of the sum of the posterior probabilities of the latent causes is shown.
1418 The horizontal axis indicates the proportion of the value in each group. **(F to H)** In the first and second
1419 columns, the sum of w_{cs} and $w_{context}$ of latent causes are shown in the boxplot, respectively. Note that
1420 the initial value of associative weight could take non-zero values, though those were comparable
1421 between groups (**Table S7**). Each black dot represents one animal. $*p < 0.05$, and $**p < 0.01$ by the
1422 Mann-Whitney U test comparing control and $App^{NL\text{-}G\text{-}F}$ mice, and p-values greater than 0.1 were not
1423 labeled on the plot. **(A to H)** Pink represents 12-month-old control ($n = 20$), and dark blue represents
1424 12-month-old $App^{NL\text{-}G\text{-}F}$ mice ($n = 18$). Detailed statistical results are provided in **Tables S7, S8, and**
1425 **S9**.

Figure 4 – figure supplement 1. Correlation matrix of estimated parameters in the 12-month-old group.

The lower triangle shows the scatter plots of pairwise combinations of two parameters where the colors of the dot in the scatterplot indicate groups: pink represents 12-month-old control mice (n = 20), and dark blue represents 12-month-old $App^{NL\text{-}G\text{-}F}$ mice (n = 18). The square outline highlights parameter pairs with statistically significant correlations. The upper triangle shows the heatmap of Spearman's correlation coefficient (ρ), where colors indicate the strength and direction of the correlation, as shown in the color bar. Significant correlations are annotated in the heatmap: *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Figure 4 – figure supplement 2. Correlations between DI and estimated parameters within the 12-month-old control group (A) and $App^{NL\text{-}G\text{-}F}$ group (B).

(A to B) Spearman's correlation coefficient (ρ) was labeled with significance, where *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$. The blue line represents the linear regression model fit, and the shaded area indicates the confidence interval. Each dot represents one animal and its color indicates groups: pink represents 12-month-old control (n = 20), and dark blue represents 12-month-old $App^{NL\text{-}G\text{-}F}$ mice (n = 18).

Figure 4 – figure supplement 3. Simulated CR (A), and posterior probabilities (B), associative weights (C) of latent causes initially inferred during the acquisition trials in the 6-month-old group in test 1 phase.

(A) The data are shown as median with interquartile range. Each black dot represents one animal. Control (Mdn = 0.84, IQR = 0.44) vs. $App^{NL\text{-}G\text{-}F}$ (Mdn = 1, IQR = 0.13), Mann-Whitney $U = 247$, $p = 0.041$. **(B)** The horizontal axis of the histogram indicates the proportion of the value in each group. Control (Mdn = 0.60, IQR = 0.10) vs. $App^{NL\text{-}G\text{-}F}$ (Mdn = 1, IQR = 0), Mann-Whitney $U = 255$, $p = 0.008$. **(C)** Each dot represents one animal. Control (Mdn = 0.308, IQR = 0.482) vs. $App^{NL\text{-}G\text{-}F}$ (Mdn = 0.373, IQR = 0.401), Mann-Whitney $U = 197$, $p = 0.633$ for w_{cs}; Control (Mdn = 0.075, IQR = 0.088) vs. $App^{NL\text{-}G\text{-}F}$ (Mdn = 0.077, IQR = 0.091), Mann-Whitney $U = 179$, $p = 0.988$ for $w_{context}$.

(A to C) Colors in the bar plot or boxplot indicate groups: pink represents 12-month-old control (n = 20), and dark blue represents 12-month-old $App^{NL\text{-}G\text{-}F}$ mice (n = 18).

Figure 5. Individual parameter estimation and internal state in the 6-month-old group.

(A) The estimated parameters of the latent cause model. **(B)** Correlation between discrimination index (DI) and estimated parameter values in the 6-month-old group. The count of latent causes initially inferred during the acquisition trials **(C, left)**, extinction trials (i.e., test 1, extinction, test 2) **(D, left)**, and trials after extinction (i.e., the unsignaled shock and test 3) **(E, left)**, with the sum of posterior probabilities **(C, D, E, right)**, and the sum of associative weights at test 3 in these latent causes **(F, G, H)**.

(A) Each black dot represents one animal. $*p < 0.05$ by the Mann-Whitney *U* test. **(B)** The y-axis shows the DI between test 3 and test 1. Spearman's correlation coefficient (ρ) was labeled with significance, where $*p < 0.05$ and $**p < 0.01$. The blue line represents the linear regression model fit, and the shaded area indicates the confidence interval. Each dot represents one animal. **(C to H)** The configuration of the figure is the same as in Figure 4. All the *p*-values of the Mann-Whitney *U* test comparing control and $App^{NL\text{-}G\text{-}F}$ mice were greater than 0.05 and were not labeled on the plot. **(A to H)** Orange represents 6-month-old control ($n = 15$), and light blue represents 6-month-old $App^{NL\text{-}G\text{-}F}$ mice ($n = 12$). Detailed statistical results are provided in **Tables S10, S11, and S12**.

Figure 5 – figure supplement 1. Correlation matrix of estimated parameters in the 6-month-old group.

The figure format is the same as in Figure 4 - figure supplement 1, except that the colors of the dots in the scatterplot indicate groups: orange represents 6-month-old control (n = 15), and light blue represents 6-month-old $App^{NL\text{-}G\text{-}F}$ mice (n = 12). Significant correlations are annotated in the heatmap: *$p < 0.05$, *$p < 0.01$, and ***$p < 0.001$.

Figure 5 – figure supplement 2. Correlation between DI and selected parameters in 6-month-old control (A) and $App^{NL\text{-}G\text{-}F}$ mice (B).

The figure format is the same as in Figure 4 - figure supplement 2, the colors of the dots indicate groups: orange represents 6-month-old control (n = 15), and light blue represents 6-month-old $App^{NL\text{-}G\text{-}F}$ mice (n = 12).

Figure 5 – figure supplement 3. Simulated CR (A), and posterior probabilities (B), associative weights (C) of latent causes initially inferred during the acquisition trials in the 6-month-old group in test 1 phase.

The figure format is the same as in Figure 4 - Supplement 3. **(A)** The data are shown as median with interquartile range. Each black dot represents one animal. Control (Mdn = 0.73, IQR = 0.34) vs. $App^{NL\text{-}G\text{-}F}$ (Mdn = 1, IQR = 0.18), Mann-Whitney $U = 110$, $p = 0.325$. **(B)** The horizontal axis of the histogram indicates the proportion of the value in each group. Control (Mdn = 0.08, IQR = 0.10) vs. $App^{NL\text{-}G\text{-}F}$ (Mdn = 1, IQR = 0.62), Mann-Whitney $U = 112$, $p = 0.249$. **(C)** Each dot represents one animal. Control (Mdn = 0.196, IQR = 0.414) vs. $App^{NL\text{-}G\text{-}F}$ (Mdn = 0.203, IQR = 0.476), Mann-Whitney $U = 91$, $p = 0.981$ for w_{cs}; Control (Mdn = 0.049, IQR = 0.095) vs. $App^{NL\text{-}G\text{-}F}$ (Mdn = 0.054, IQR = 0.074), Mann-Whitney $U = 94$, $p = 0.867$ for $w_{context}$. **(A to C)** Colors in the bar plot or boxplot indicate groups: orange represents 6-month-old control (n = 15), and light blue represents 6-month-old $App^{NL\text{-}G\text{-}F}$ mice (n = 12).

1499 **Figure 6. *App*$^{NL\text{-}G\text{-}F}$ mice failed to infer coexisting spatial memories in the reversal Barnes maze**
1500 **task.**
1501 **(A)** Schematic diagram of the reversal Barnes maze task. The largest circle represents the Barnes maze
1502 field. The filled and open small circles represent the target hole with the escape box and the remaining
1503 holes in the field, respectively. The small arrow pointing to the hole indicated the position of the target
1504 hole without an escape box in the probe test. The same color of the filled circle and arrow indicates
1505 the same position. **(B)** Strategy usage in initial training (days 1 to 6), the first reversal training (days 9
1506 to 11), and the second reversal training (days 12 to 14) in the 12-month-old group. Time spent around
1507 each hole at probe test 1 **(C)** and probe test 2 **(D)** in the 12-month-old group. **(E)** Discrimination index
1508 (DI) of the second target hole between probe test 2 and probe test 1 in the 12-month-old group. **(F)**
1509 Strategy usage in initial training (days 1 to 6), the first reversal training (days 9 to 11), and the second
1510 reversal training (days 12 to 14) in the 6-month-old group. Time spent around each hole at probe test
1511 1 **(G)** and probe test 2 **(H)** in the 6-month-old group. **(I)** Discrimination index (DI) of the second target
1512 hole between probe test 2 and probe test 1 in the 6-month-old group.
1513 **(B and F)** $^{*}p < 0.05$ of perimeter strategy, $^{\#}p < 0.05$ of confirmatory strategy, $^{\dagger}p < 0.05$ of serial
1514 strategy, $^{\ddagger}p < 0.05$ of spatial strategy by Wilcoxon rank-sum test comparing control and *App*$^{NL\text{-}G\text{-}F}$ mice
1515 at the same age. **(C, D, G, and H)** The data are shown as mean with 95% confidence interval. $^{*}p <$
1516 0.05, $^{**}p < 0.01$, and $^{***}p < 0.001$ by mixed-design two-way [Genotype (control, *App*$^{NL\text{-}G\text{-}F}$) × Hole
1517 (1-12)] ANOVA; $^{\dagger}p < 0.05$ by Tukey's HSD test at the specific hole to compare control and *App*$^{NL\text{-}G\text{-}}$
1518 ^{F} mice at the same age. **(E and I)** The dashed horizontal line indicates DI = 0.5, which means no
1519 discrimination between the two phases. $^{\dagger}p < 0.05$ by the Wilcoxon signed-rank test, and the alternative
1520 hypothesis specifies that the mean differs from 0.5; $^{*}p < 0.05$ by Mann-Whitney *U* test comparing
1521 control and *App*$^{NL\text{-}G\text{-}F}$ mice. **(C-E, G-I)** Each black dot represents one animal. Colors indicate the
1522 different groups: pink represents 12-month-old control (n = 18), dark blue represents 12-month-old
1523 *App*$^{NL\text{-}G\text{-}F}$ mice (n = 18), orange represents 6-month-old control (n = 24), and light blue represents 6-
1524 month-old *App*$^{NL\text{-}G\text{-}F}$ mice (n = 17). Detailed statistical results are provided in **Tables S14 to S16** for
1525 the 12-month-old group results and **Tables S18 to S20** for the 6-month-old group results.
1526

1527 **Figure 6 – figure supplement 1.** Conventional analysis for initial **(A)** and reversal **(B)** training day in
1528 the Barnes maze task and discrimination index of the first target hole between probe test 2 and probe
1529 test 1 **(C)** in the 12-month-old group. Colors indicate different groups: pink represents 12-month-old
1530 control (n = 18), and dark blue represents 12-month-old $App^{NL\text{-}G\text{-}F}$ mice (n = 18).
1531 **(A and B)** The left, middle, and right panels show the number of errors, latency, and travel distance
1532 during training days. The data are shown as mean with 95% confidence interval. *$p < 0.05$, **$p < 0.01$,
1533 and ***$p < 0.001$ by mixed-design two-way [Genotype (control, $App^{NL\text{-}G\text{-}F}$) × Day (1-6)] ANOVA.
1534 Detailed statistical results are provided in **Table S13**. **(C)** Each black dot represents one animal. The
1535 dashed horizontal line indicates DI = 0.5, which means no discrimination between the two phases. $^{\dagger}p$
1536 < 0.05 by the Wilcoxon signed-rank test, and the alternative hypothesis specifies that the mean differs
1537 from 0.5; *$p < 0.05$ by Mann-Whitney U test comparing control and $App^{NL\text{-}G\text{-}F}$ mice. Detailed statistical
1538 results are provided in **Table S16**.
1539
1540 **Figure 6 – figure supplement 2. (A)** Animal trajectory examples and definition of each strategy. The
1541 red solid circle indicates the target hole, and the black solid circle indicates other holes in the field.
1542 The dotted line represents the animal's trajectory, and the color reflects the time in the trial. **(B)** The
1543 algorithm of strategy analysis.
1544
1545 **Figure 6 – figure supplement 3.** Conventional analysis for initial **(A)** and reversal **(B)** training day in
1546 the Barnes maze task and discrimination index of the first target hole between probe test 2 and probe
1547 test 1 **(C)** in the 6-month-old group. The configuration of the figure is the same as in Figure 6 - figure
1548 supplement 1. Colors indicate the different groups: orange represents 6-month-old control (n = 24),
1549 and light blue represents 6-month-old $App^{NL\text{-}G\text{-}F}$ mice (n = 17). Detailed statistical results are provided
1550 in **Table S17** for panels A and B; **Table S20** for panel C.

Figure 1

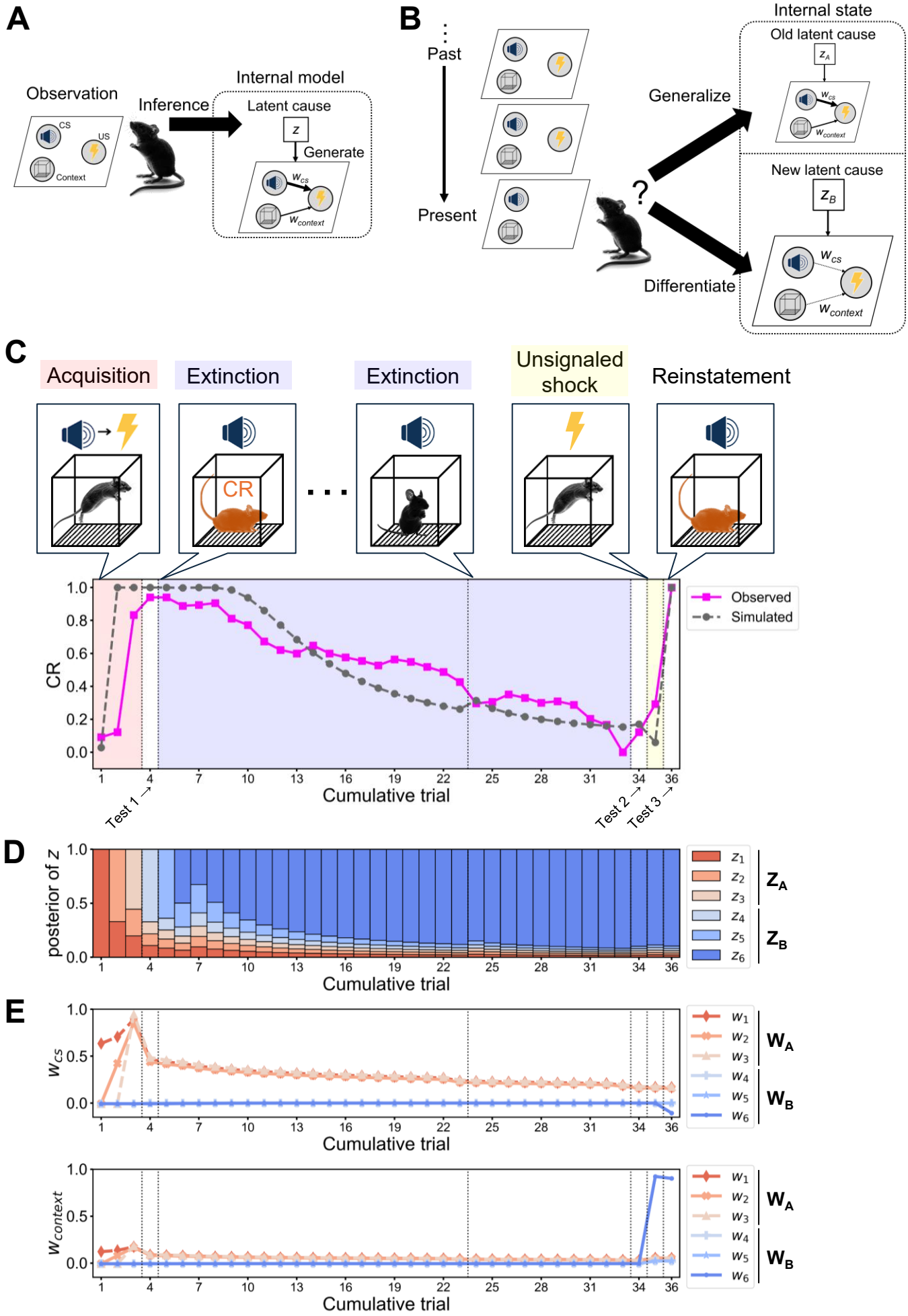

A
Observation
CS
US
Context
Inference
Internal model
Latent cause
z
Generate
w_{cs}
$w_{context}$
B
Past
Present
?
Generalize
Differentiate
Internal state
Old latent cause
z_A
New latent cause
z_B
C
Acquisition
Extinction
Extinction
Unsignaled shock
Reinstatement
CR
Observed
Simulated
Cumulative trial
Test 1
Test 2
Test 3
D
posterior of z
z_1 z_2 z_3 z_4 z_5 z_6
Z_A
Z_B
E
w_{cs}
$w_{context}$
w_1 w_2 w_3 w_4 w_5 w_6
W_A
W_B

Figure 2

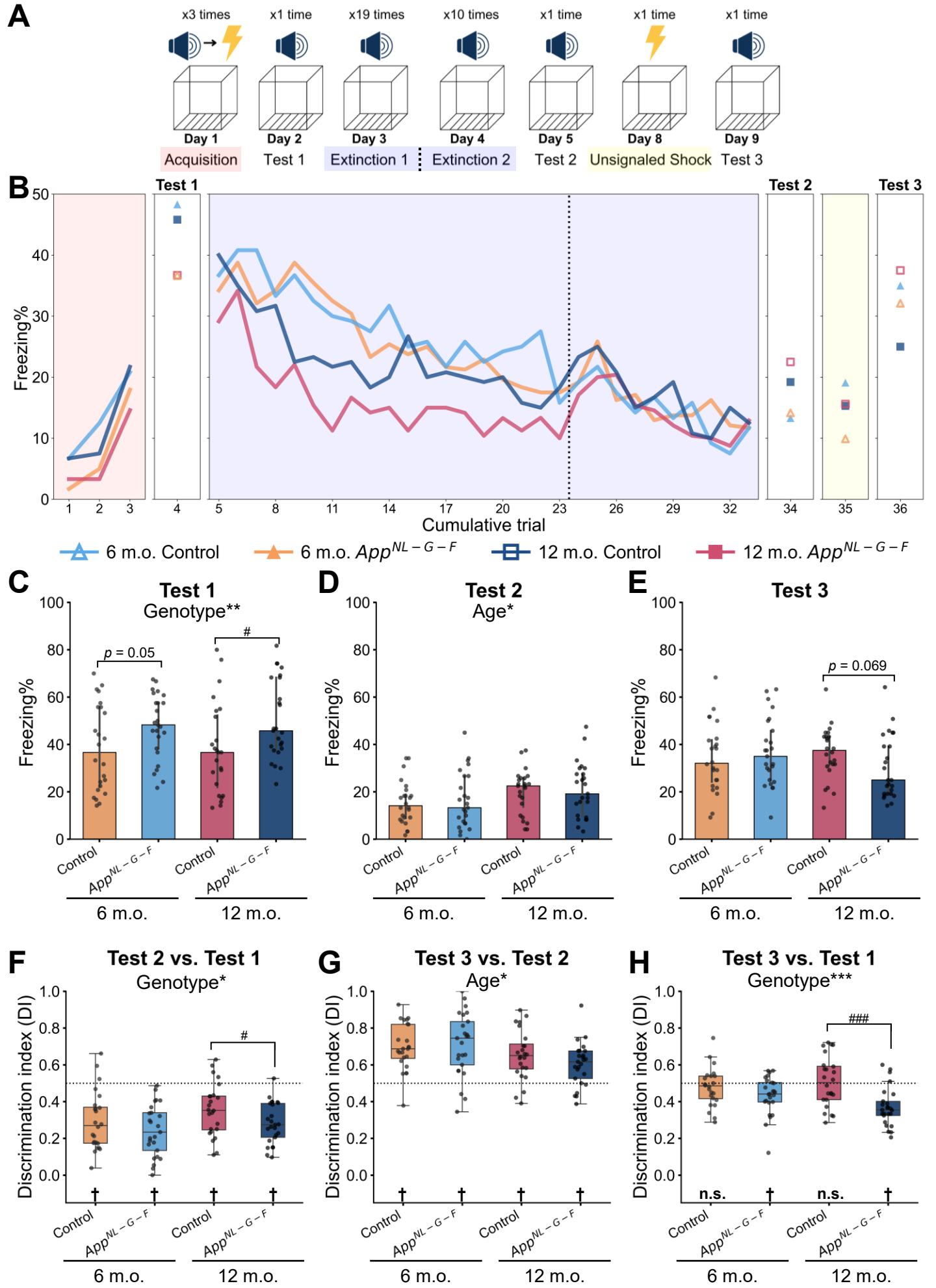

A
x3 times
x1 time
x19 times
x10 times
x1 time
x1 time
x1 time
Day 1
Day 2
Day 3
Day 4
Day 5
Day 8
Day 9
Acquisition
Test 1
Extinction 1
Extinction 2
Test 2
Unsignaled Shock
Test 3
B
Test 1
Test 2
Test 3
Freezing%
Cumulative trial
6 m.o. Control
6 m.o. App^NL−G−F
12 m.o. Control
12 m.o. App^NL−G−F
C
Test 1
Genotype**
p = 0.05
#
Freezing%
Control
App^NL−G−F
6 m.o.
12 m.o.
D
Test 2
Age*
E
Test 3
p = 0.069
F
Test 2 vs. Test 1
Genotype*
Discrimination index (DI)
G
Test 3 vs. Test 2
Age*
H
Test 3 vs. Test 1
Genotype***
###
n.s.

Figure 3

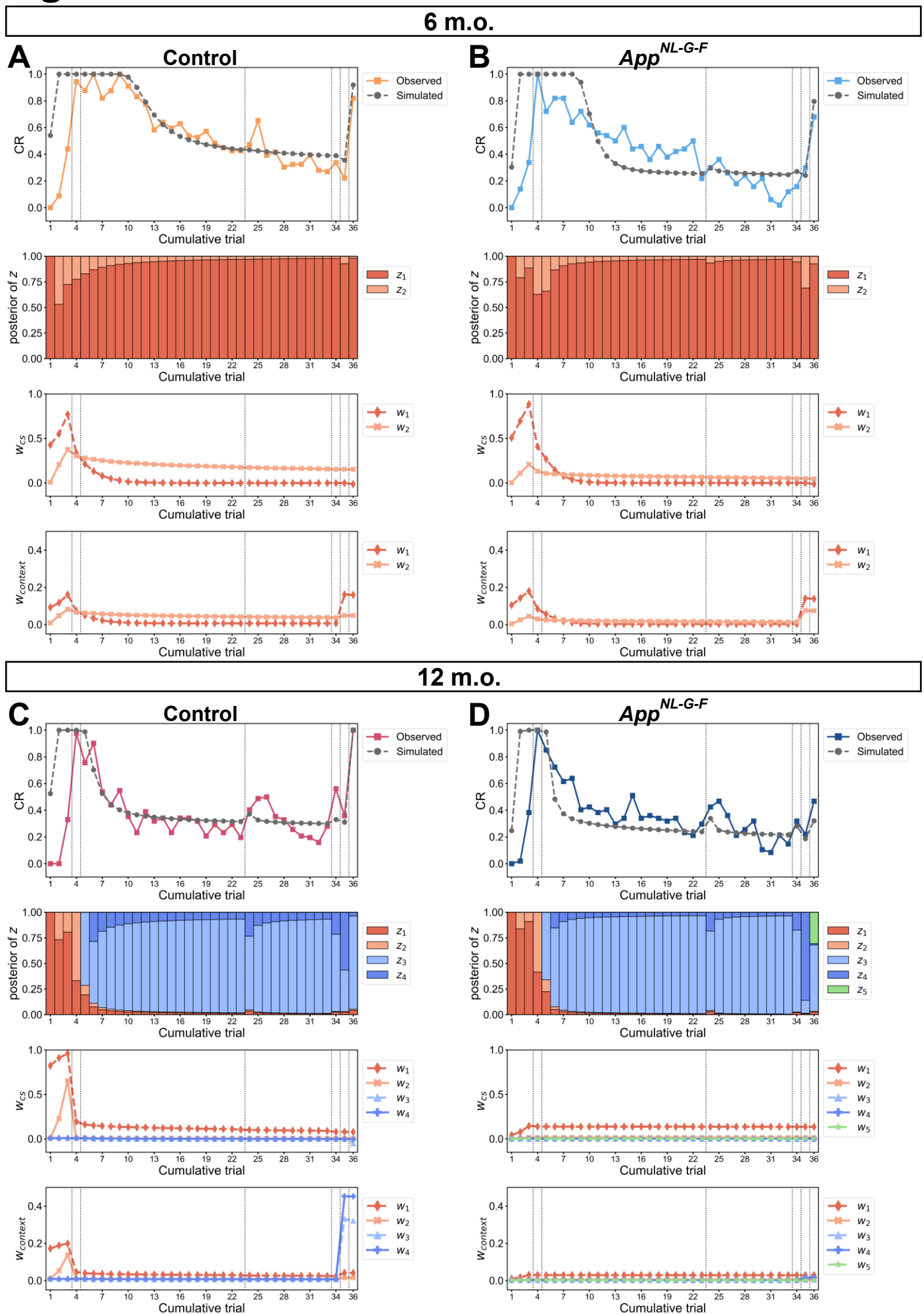

6 m.o.
A
Control
B
App NL-G-F
12 m.o.
C
Control
D
App NL-G-F
Observed
Simulated
CR
Cumulative trial
posterior of z
z_1
z_2
z_3
z_4
z_5
w_{cs}
$w_{context}$
w_1
w_2
w_3
w_4
w_5

Figure 4

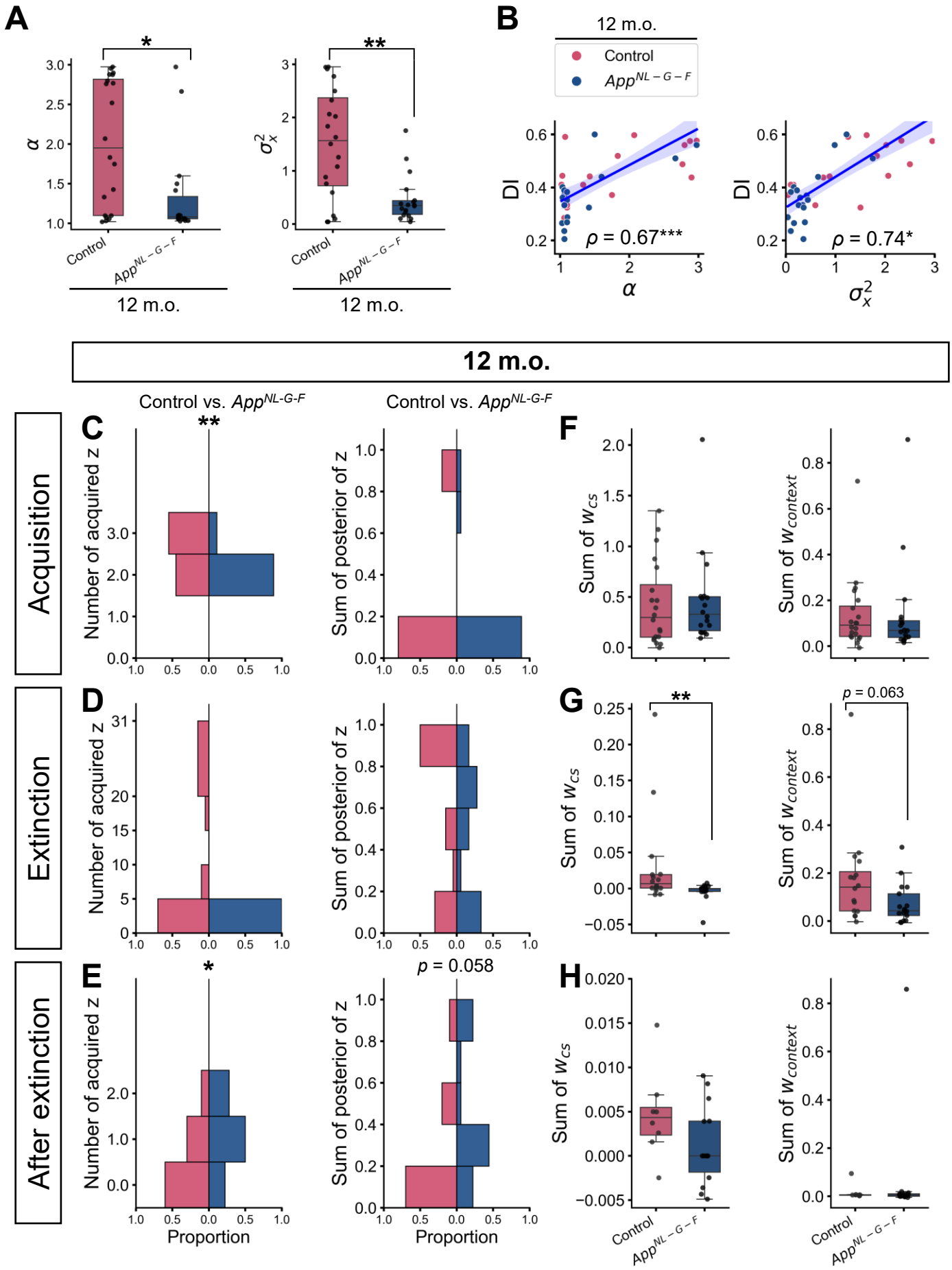

A
B
12 m.o.
Control
App^NL−G−F
α
σ²x
DI
ρ = 0.67***
ρ = 0.74*
12 m.o.
Acquisition
Extinction
After extinction
C
D
E
F
G
H
Control vs. App^NL-G-F
Number of acquired z
Sum of posterior of z
Proportion
Sum of w_cs
Sum of w_context
p = 0.058
p = 0.063

Figure 5

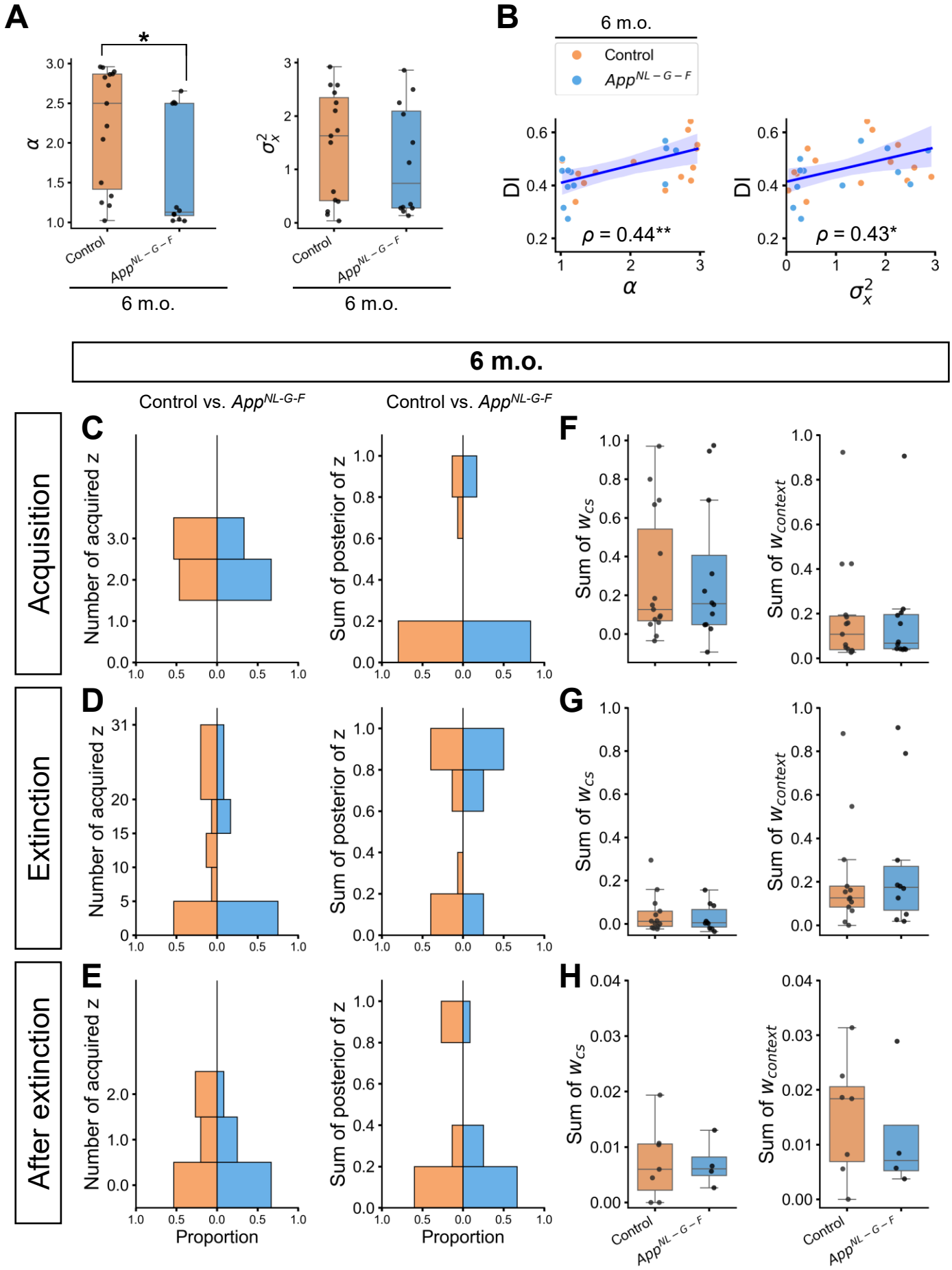

A
*
α
σ_x^2
Control
App^{NL-G-F}
6 m.o.
B
6 m.o.
Control
App^{NL-G-F}
DI
ρ = 0.44**
ρ = 0.43*
α
σ_x^2
6 m.o.
Control vs. App^{NL-G-F}
Control vs. App^{NL-G-F}
Acquisition
Extinction
After extinction
C
D
E
F
G
H
Number of acquired z
Sum of posterior of z
Sum of w_{cs}
Sum of $w_{context}$
Proportion
Proportion

Figure 6

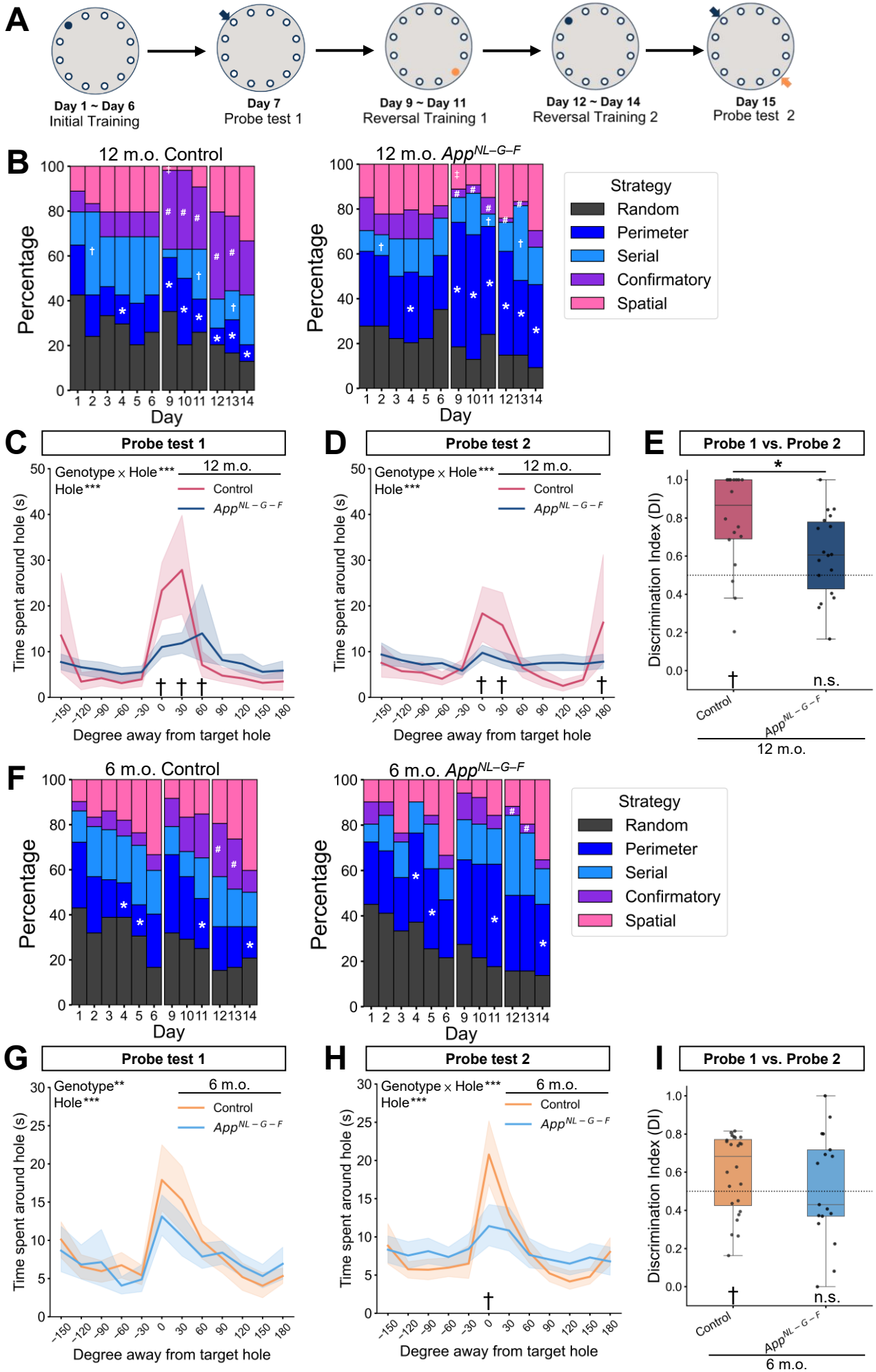
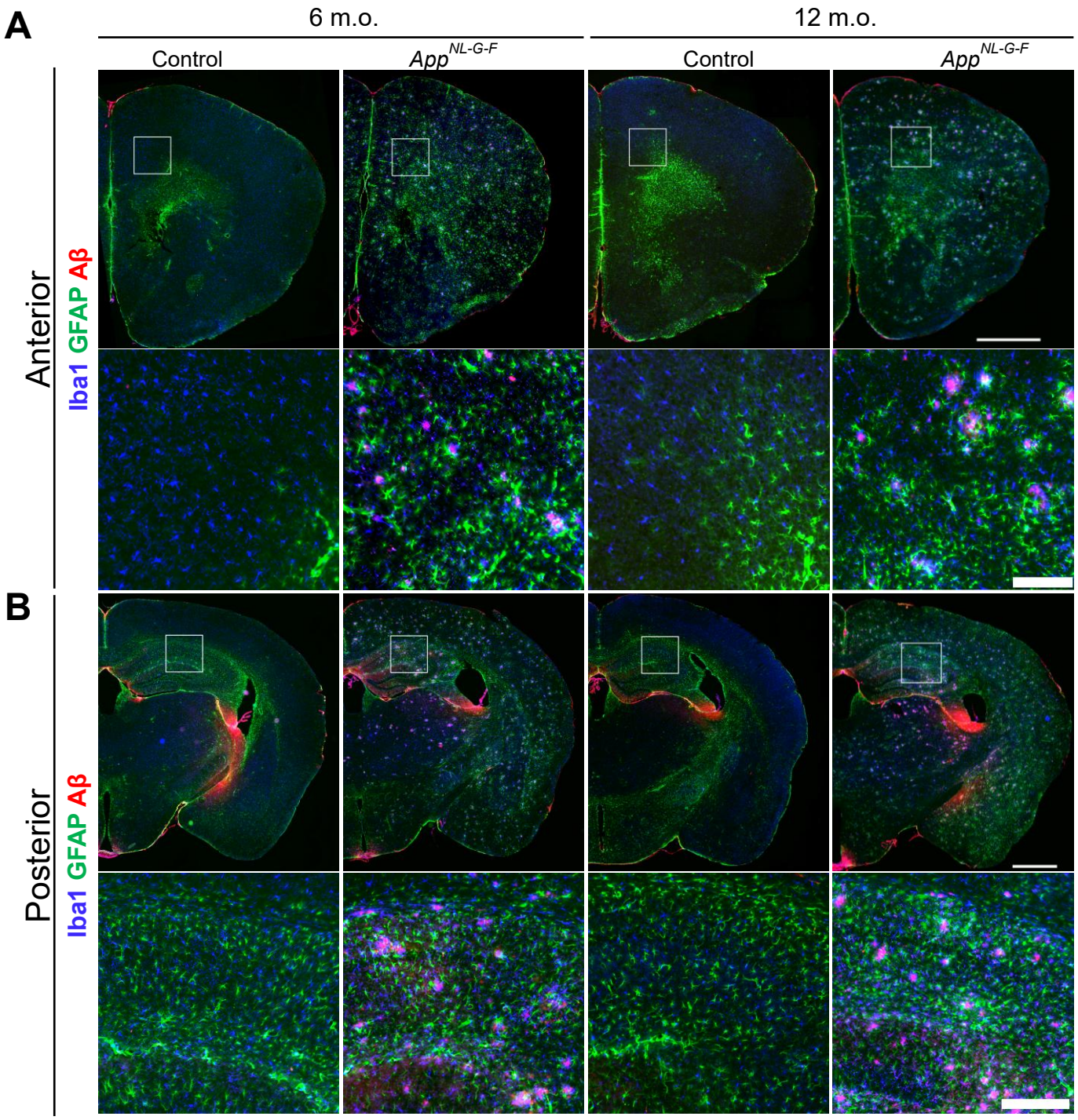

A
Day 1 ~ Day 6
Initial Training
Day 7
Probe test 1
Day 9 ~ Day 11
Reversal Training 1
Day 12 ~ Day 14
Reversal Training 2
Day 15
Probe test 2
B
12 m.o. Control
12 m.o. App^NL−G−F
Percentage
Day
Strategy
Random
Perimeter
Serial
Confirmatory
Spatial
C
Probe test 1
Genotype × Hole***
Hole***
12 m.o.
Control
App^NL−G−F
Time spent around hole (s)
Degree away from target hole
D
Probe test 2
E
Probe 1 vs. Probe 2
Discrimination Index (DI)
n.s.
Control
App^NL−G−F
12 m.o.
F
6 m.o. Control
6 m.o. App^NL−G−F
G
Genotype**
6 m.o.
H
I
6 m.o.

Figure 2 – figure supplement 1



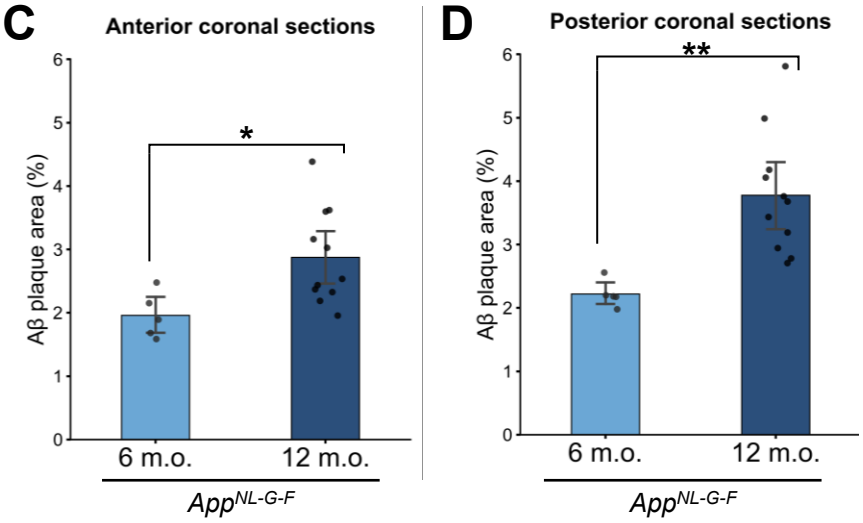


Figure 2 – figure supplement 2

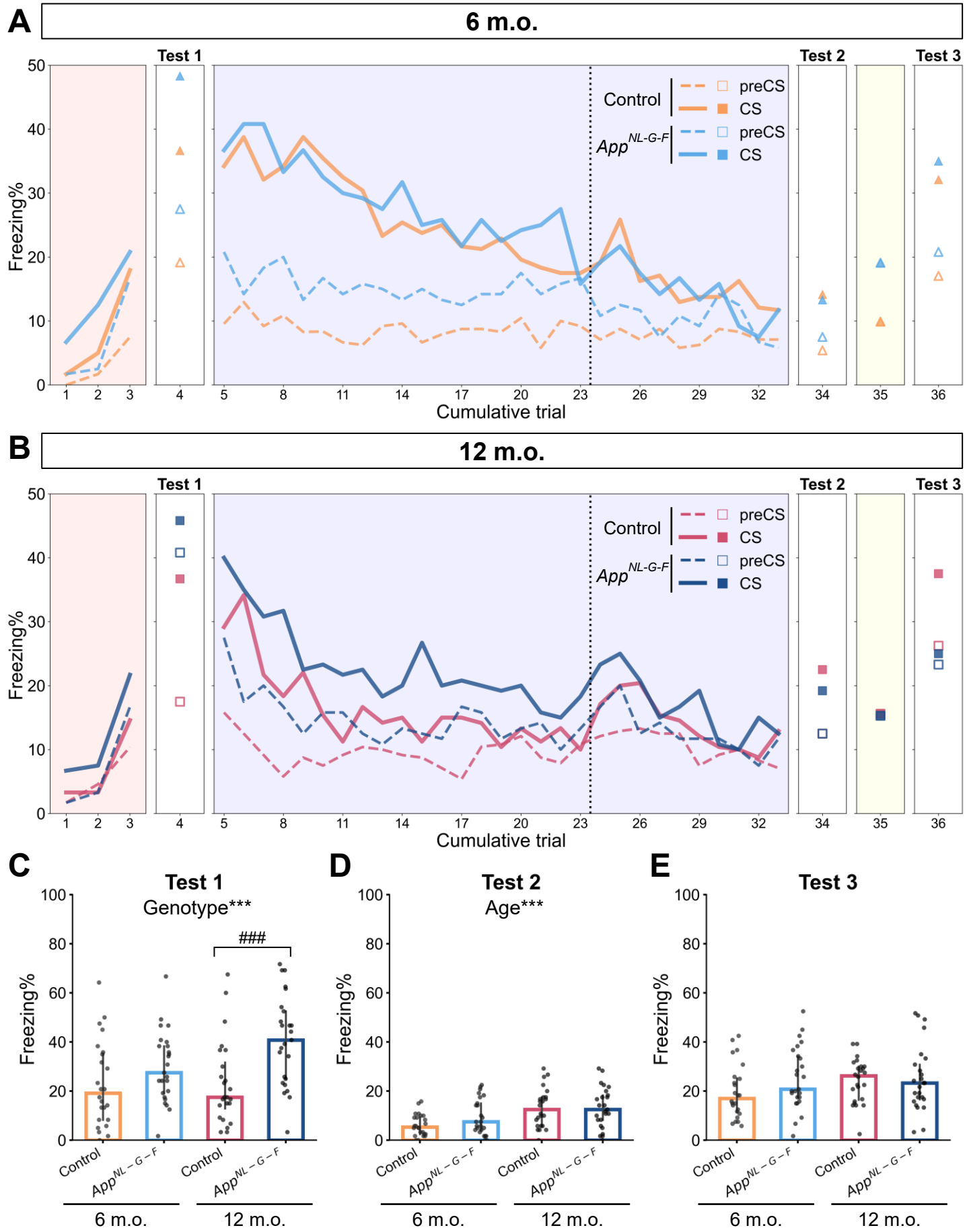


Figure 3 – figure supplement 1

Latent cause model

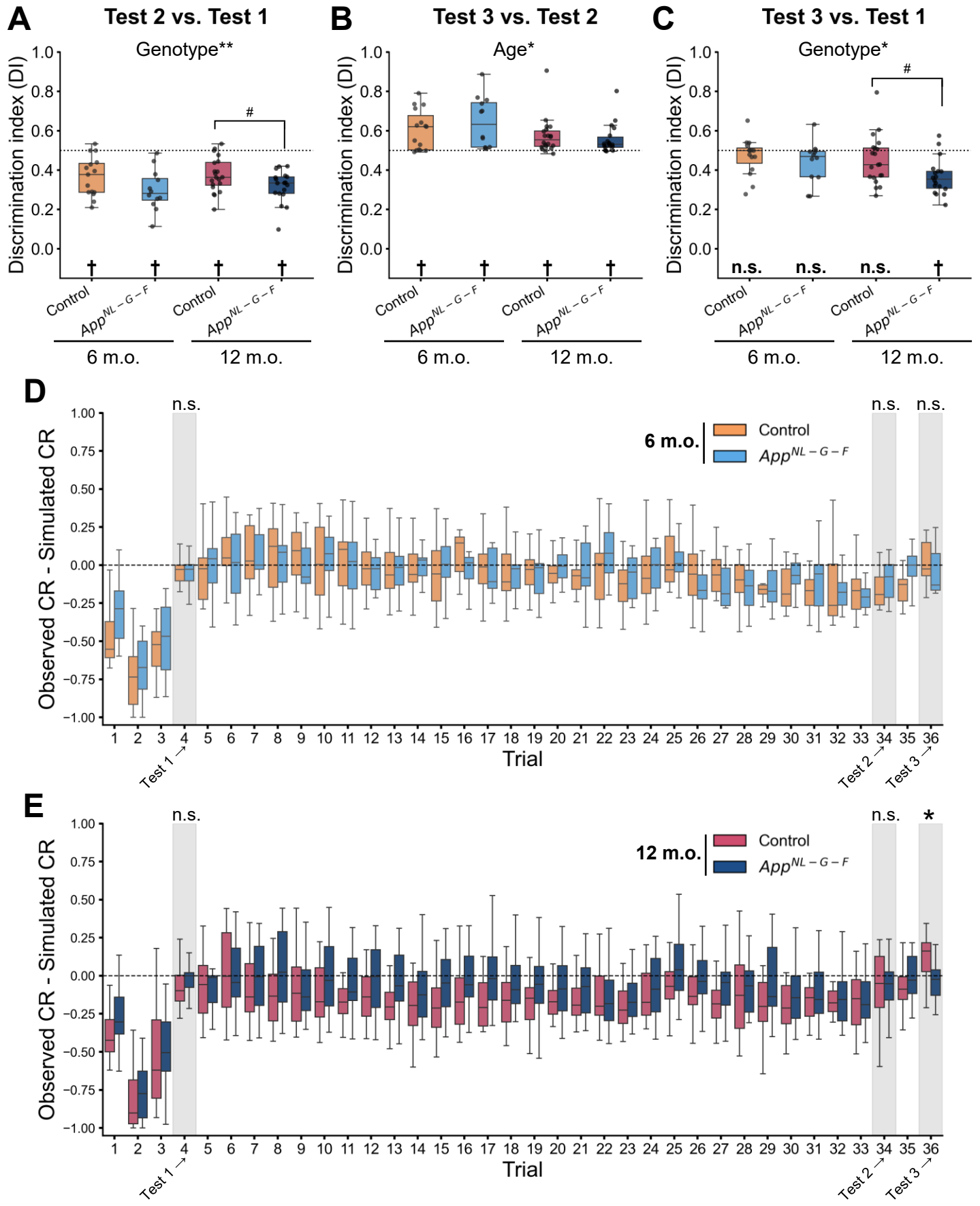


Figure 4 – figure supplement 1

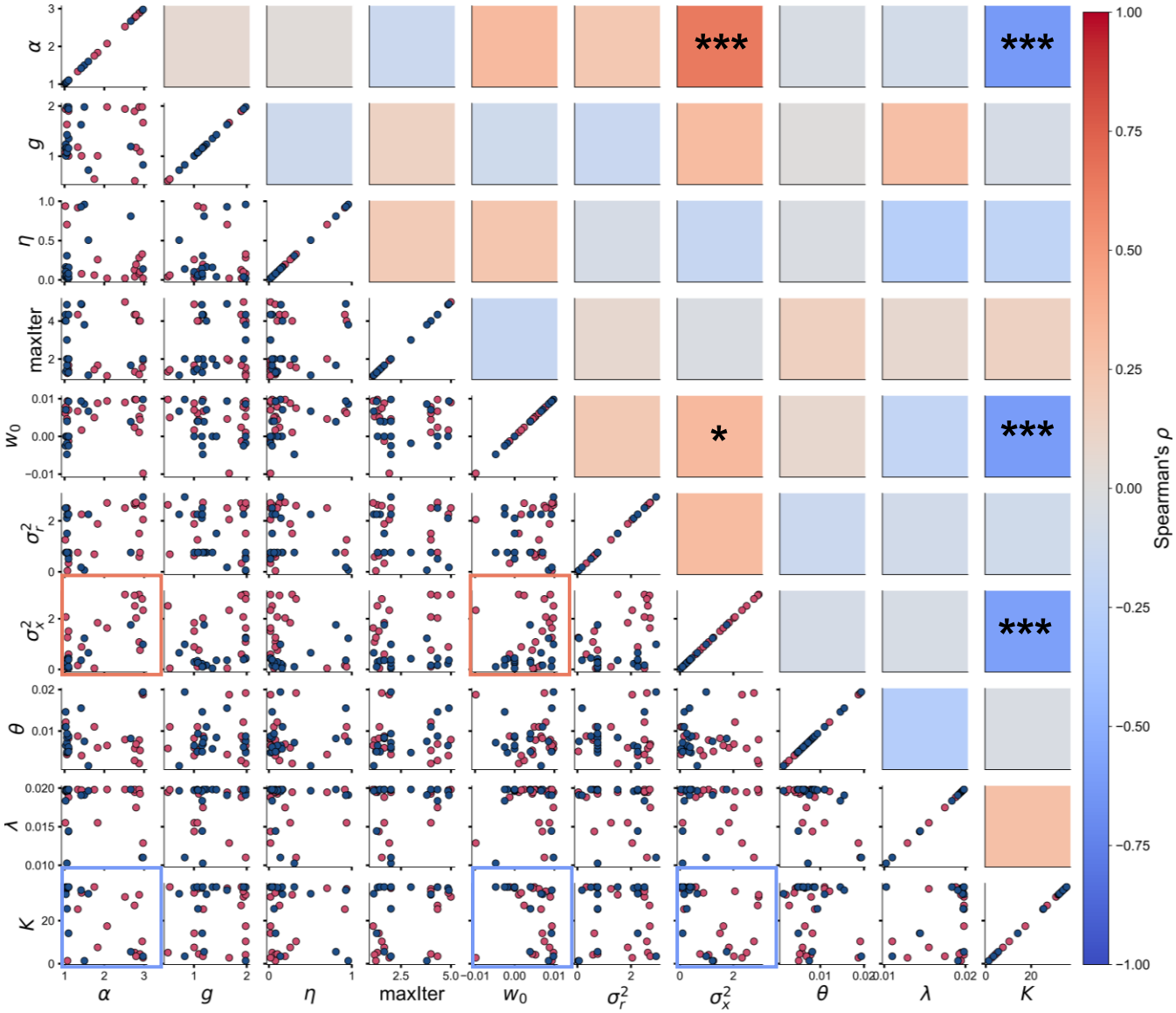


Figure 4 – figure supplement 2

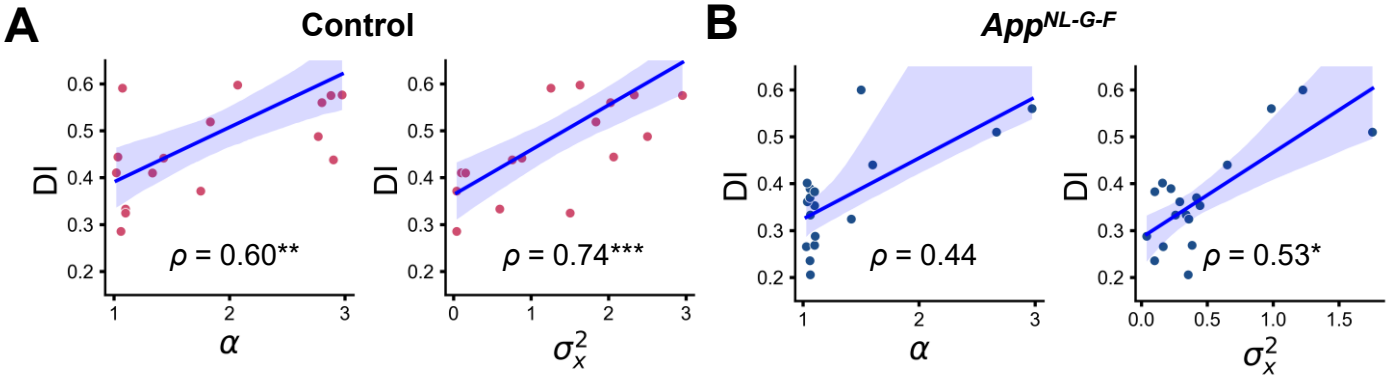


Figure 4 – figure supplement 3

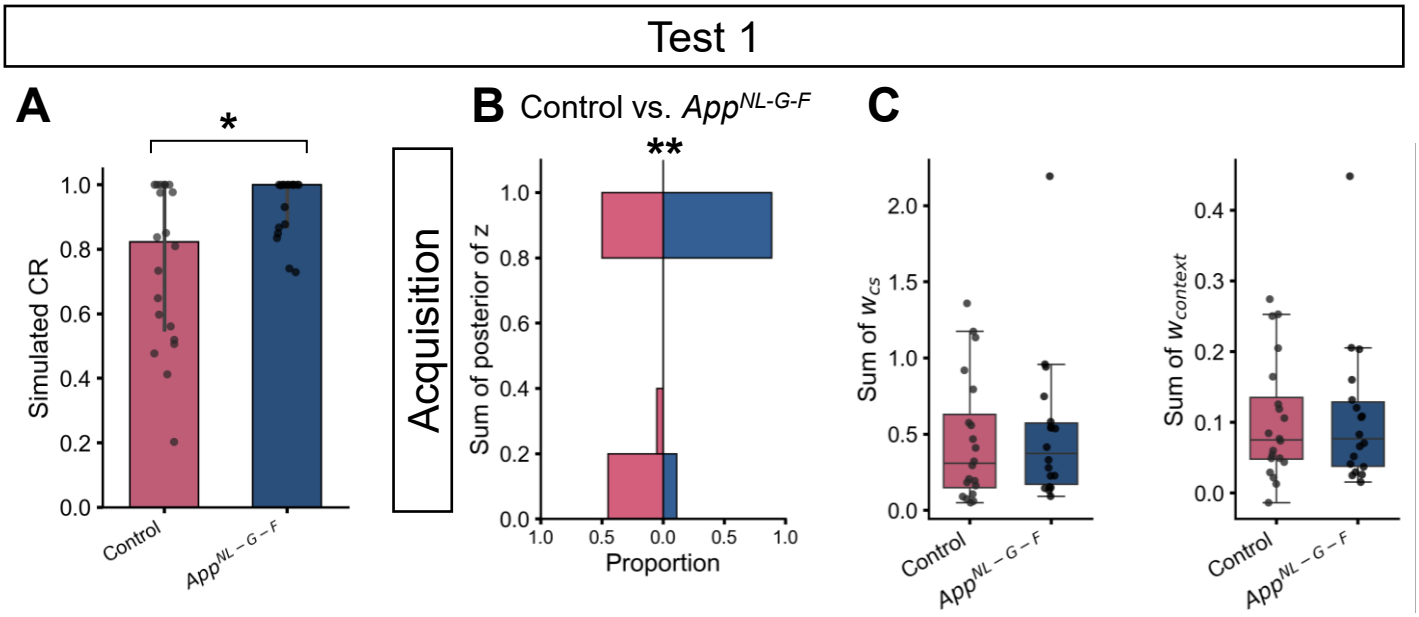


Figure 5 – figure supplement 1

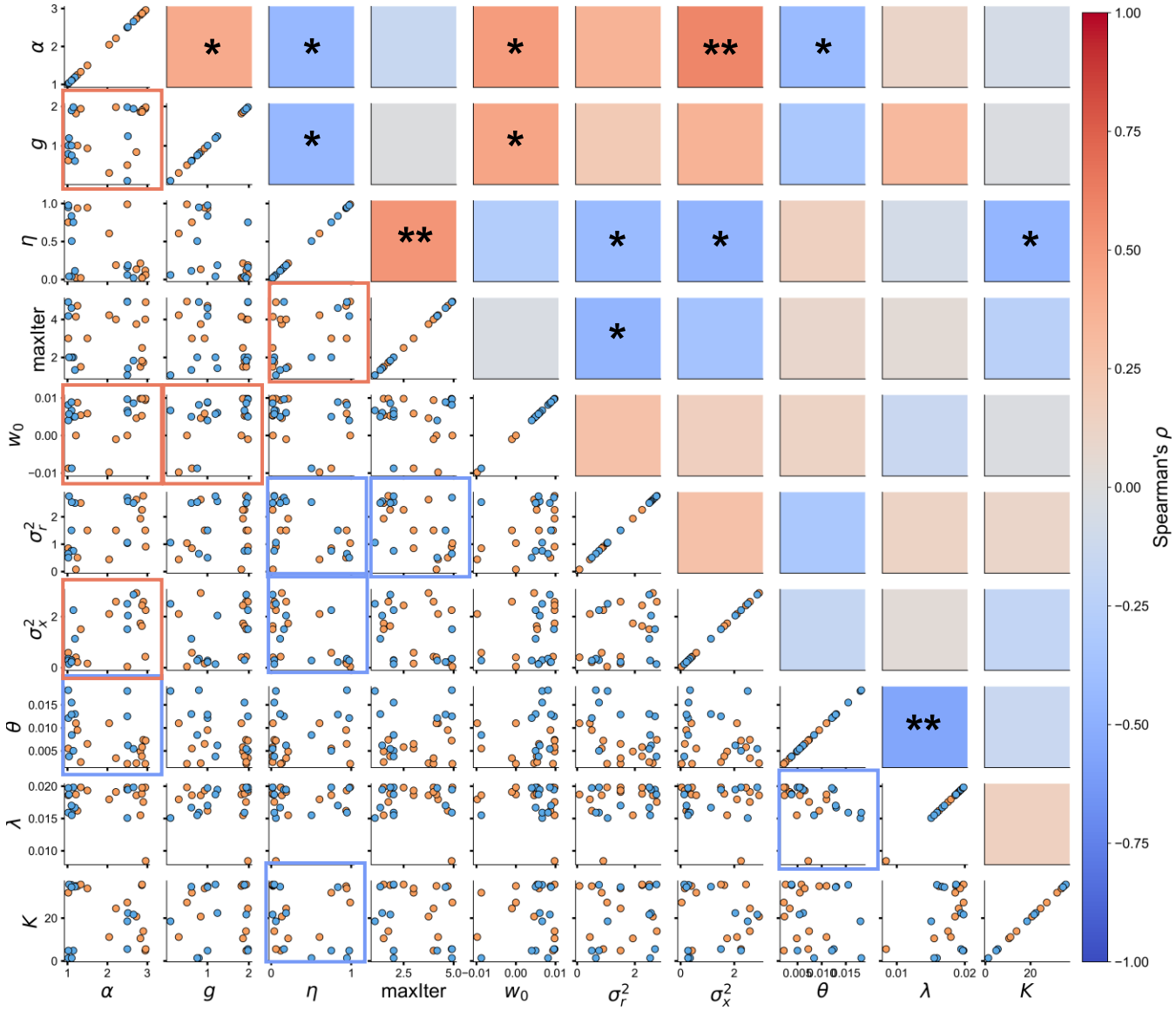


Figure 5 – figure supplement 2

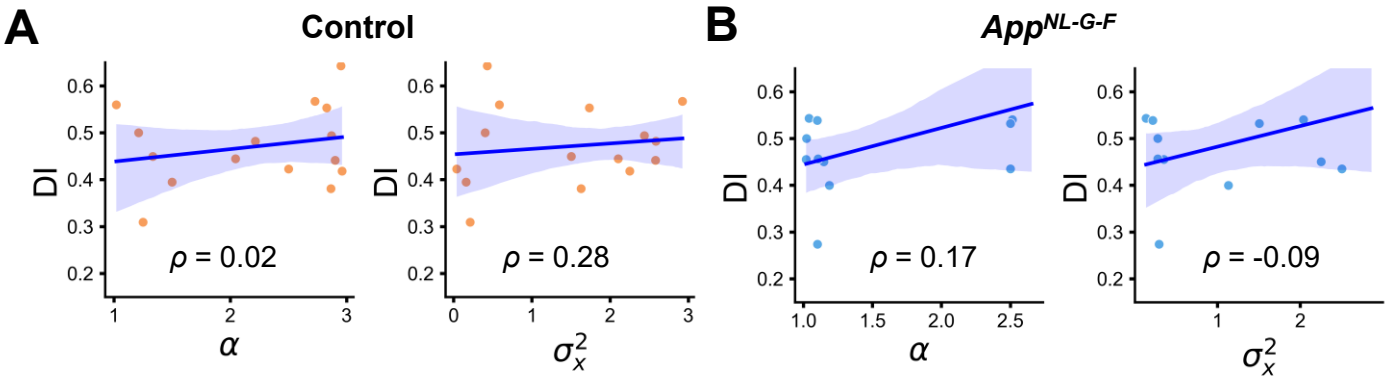


Figure 5 – figure supplement 3

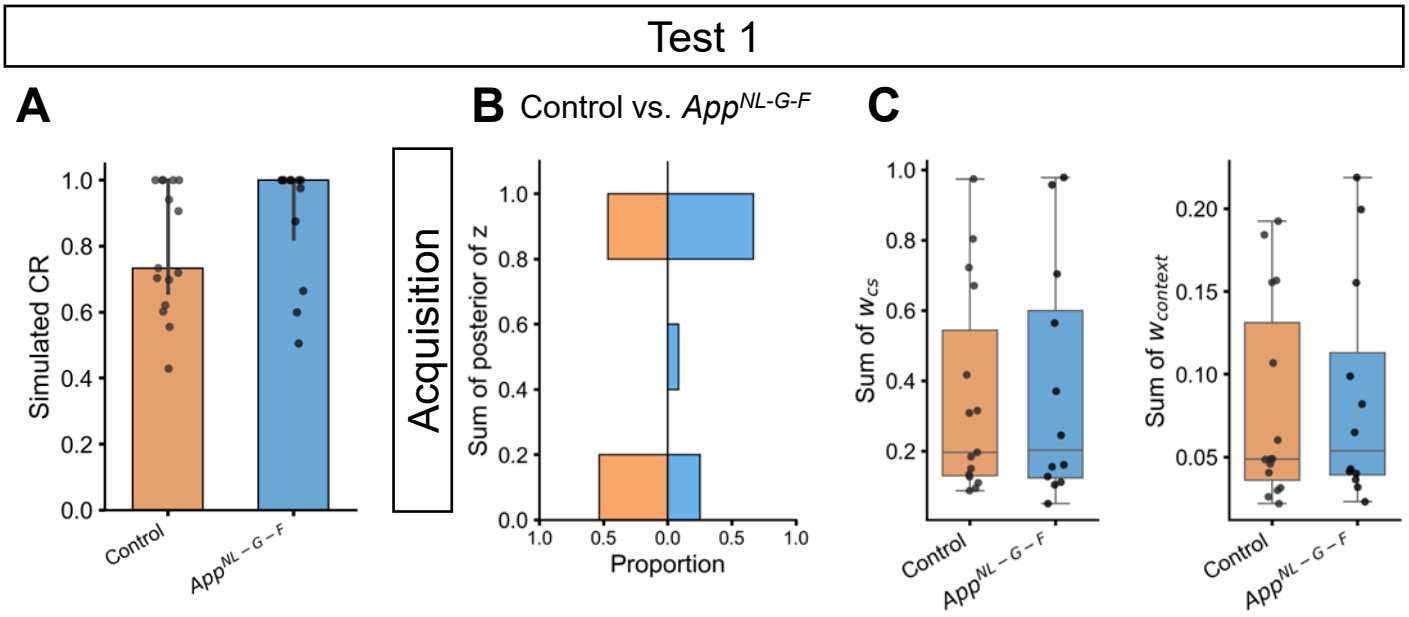
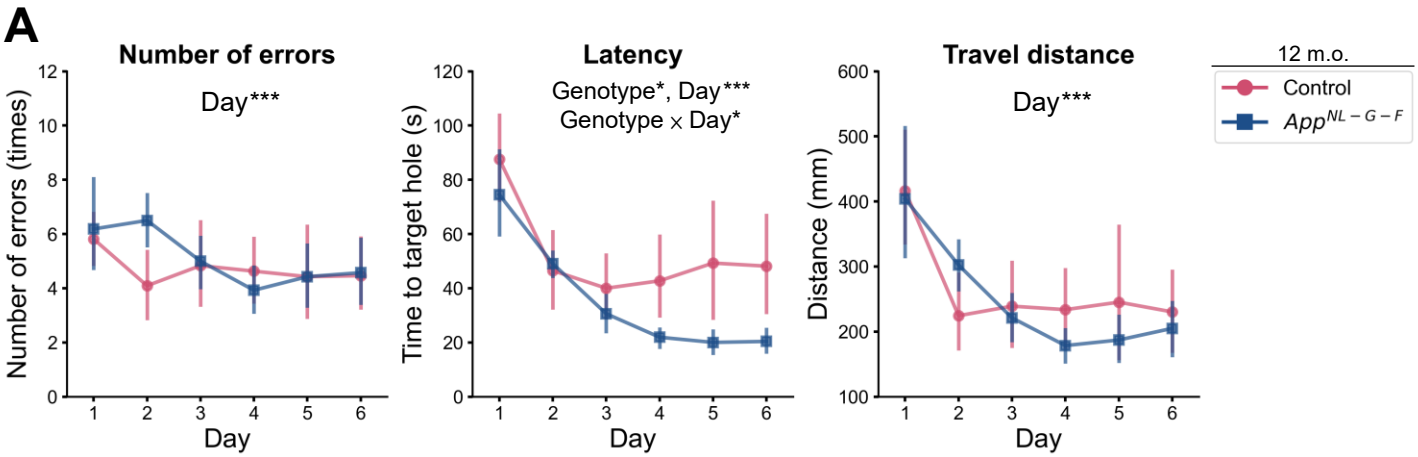
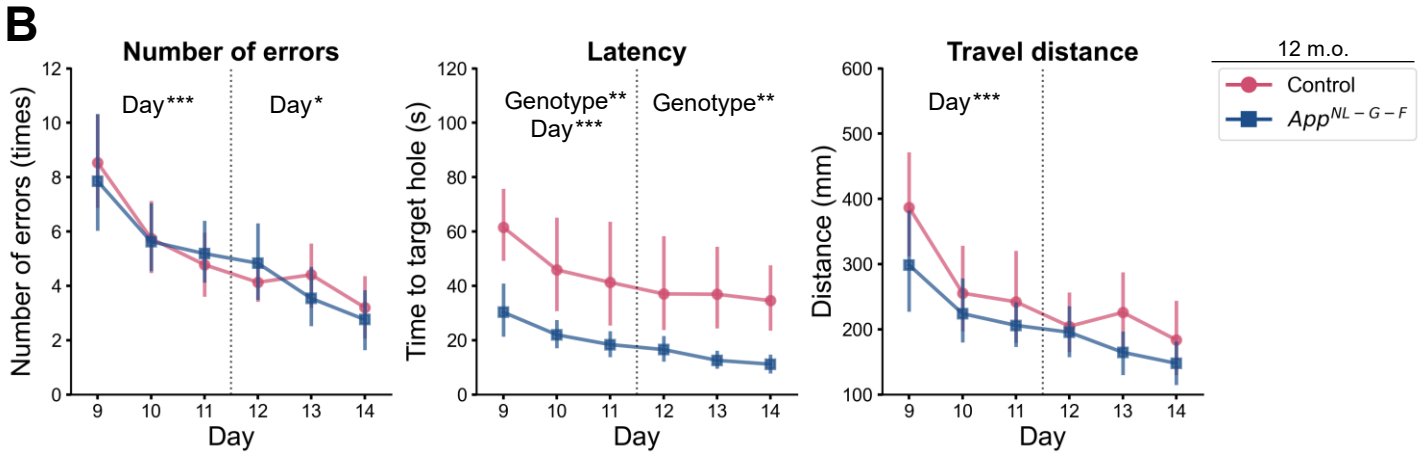


Figure 6 – figure supplement 1





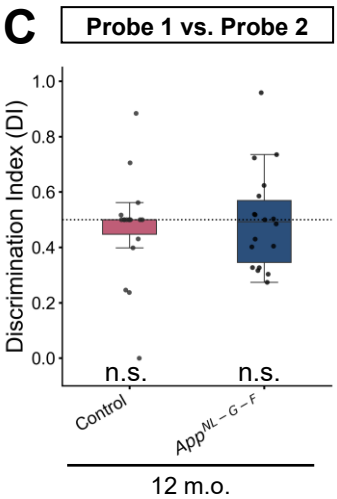
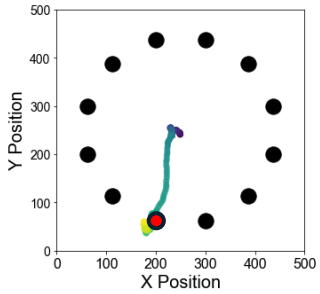


Figure 6 – figure supplement 2

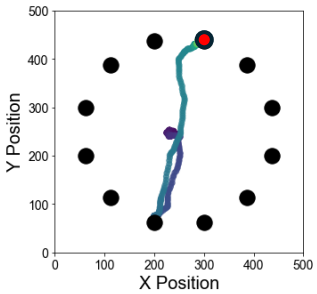
A

Spatial



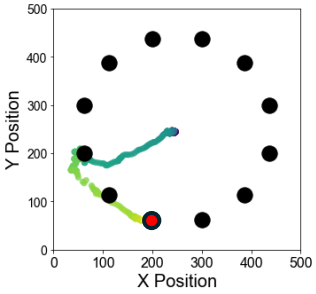
- Quadrant pass = 0
- The number of errors < 3
- The mouse only visited the target hole and holes beside the target

Confirmatory



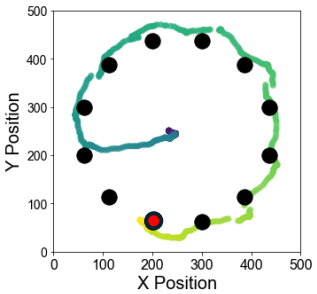
- The mouse visited the quadrant of the target hole and the opposite quadrant
- The time spent in the opposite quadrant was greater than 25% of the exploration time.

Serial



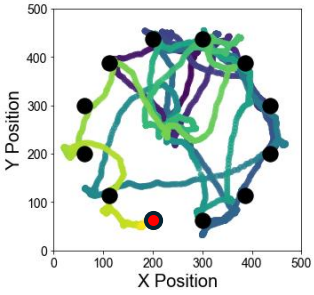
- The mouse visited the quadrant of the target hole and the two neighboring quadrants
- the mouse visited the hole sequentially

Perimeter



- Quadrant pass > 2
- The time spent in peripheral zone was greater than 75% of the exploration time

Random



- Otherwise

B

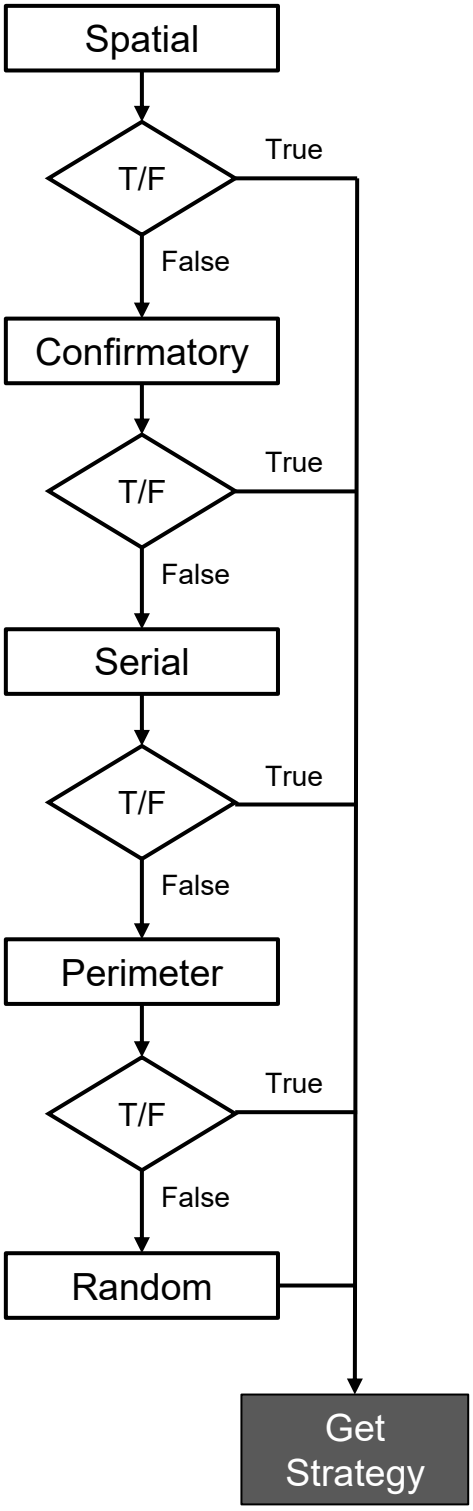
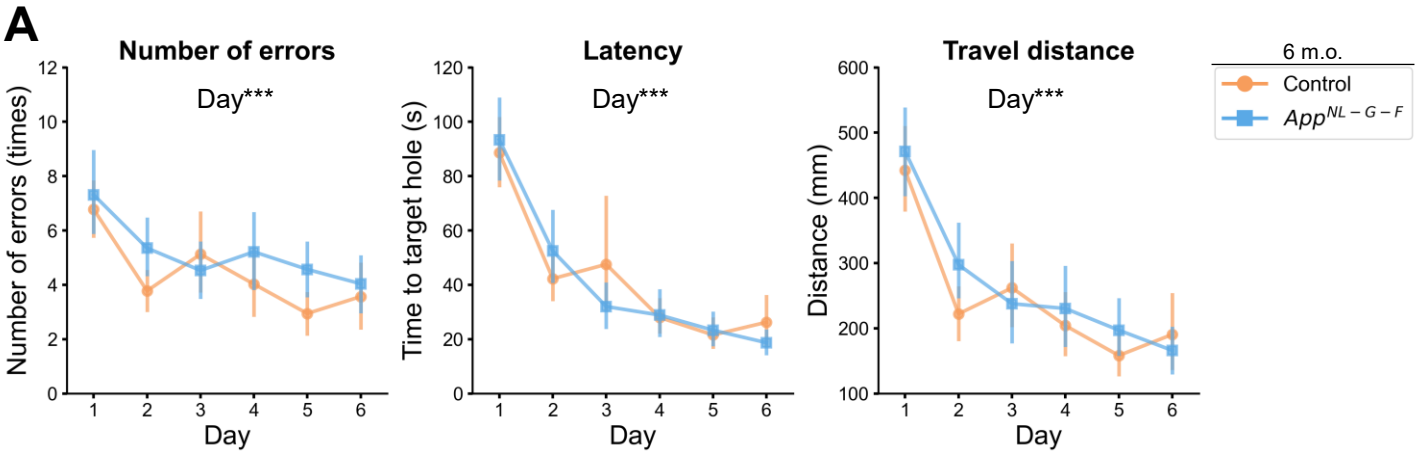
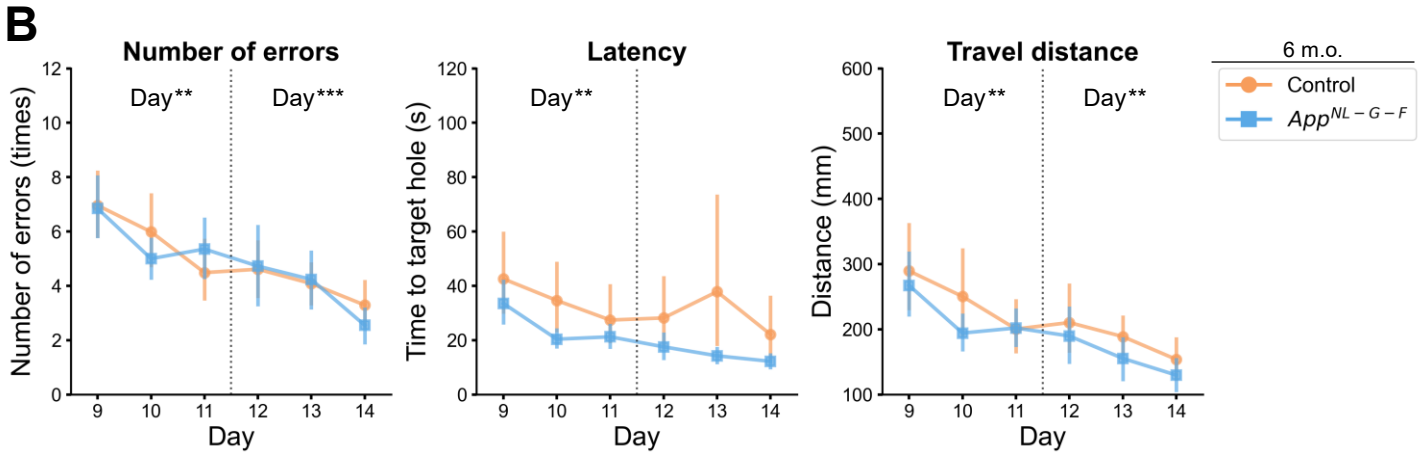
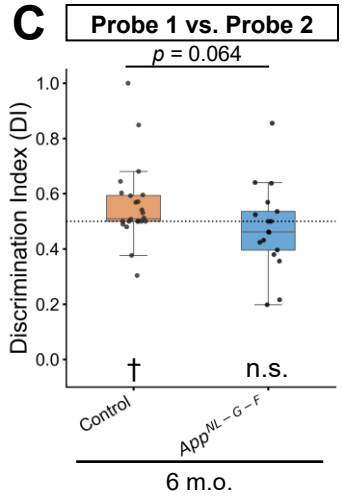



Figure 6 – figure supplement 3