

RESEARCH ARTICLE

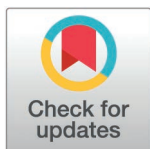
Failure to classically condition *G. dorotocephala* and *S. mediterranea* flatworms using historical and contemporary light-shock paradigms

Zachary S. Kelso¹*, Madeleine C. Snyder¹*, Samuel J. Gershman¹

Department of Psychology and Center for Brain Science, Harvard University, Cambridge, Massachusetts, United States of America

* These authors contributed equally to this work.

* zkelso@fas.harvard.edu (ZSK); maddiesnyder@fas.harvard.edu (MCS)



OPEN ACCESS

Citation: Kelso ZS, Snyder MC, Gershman SJ (2026) Failure to classically condition *G. dorotocephala* and *S. mediterranea* flatworms using historical and contemporary light-shock paradigms. PLoS One 21(9): e0358756. <https://doi.org/10.1371/journal.pone.0358756>

Editor: Maria Y Pakharukova, Institute of Cytology and Genetics SB RAS: FIC Institut citologii i genetiki Sibirskogo otdelenia Rossijskoj akademii nauk, RUSSIAN FEDERATION

Received: April 27, 2026

Accepted: September 4, 2026

Published: September 23, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0358756>

Copyright: © 2026 Kelso et al. This is an open access article distributed under the terms of

Abstract

Planarian flatworms represent one of the most evolutionarily informative nervous systems for an account of ancient bilaterian brains. Likewise, the unparalleled regenerative ability of planarians makes possible certain investigations of neural development, memory, and behavior that are simply impossible with other model organisms. Despite these facts, learning and memory are today underexplored in planarians, likely due in part to the controversial legacy of 20th-century experiments on the transfer of memories between individual flatworms. Here, we attempted to replicate and extend the classical conditioning experiments in planarians that were the basis of the later memory transfer work. We failed to find evidence of learning using both historical and contemporary conditioning protocols, and using computer vision methods to avoid subjectivity in manual video annotation. Our results cast doubt on the suitability of historic conditioning protocols for *Girardia dorotocephala* and *Schmidtea mediterranea* planarians. We encourage future researchers to explore more ethologically relevant conditioning paradigms to establish a behavioral foundation for investigating the molecular basis of learning.

1 Introduction

Evolutionary and comparative neuroscience suggest that planarian nervous systems represent the most basal, Urbilaterian-like brains of any contemporary organism [1–4]. Consequently, planarians offer a window into the evolutionary origins of cognitive brain functions, including learning, memory, and decision making. The twentieth-century discovery that planarians appear to be capable of learning and memory was a watershed in this research program. Even more remarkably, reports of memory retention after decapitation and regeneration, as well as memory transfer between individuals after RNA injections or cannibalism, suggest that planarian brains could shed light on novel mechanisms of memory storage. However, these

the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: Code to reproduce the main and supplementary figures in this paper is available here: <https://github.com/madeleinesnyder/PlanariaPipe/tree/main> and the mp4 files used for manual scoring and binary masks used for the feature extraction pipeline are available on the Harvard Dataverse (<https://doi.org/10.7910/DVN/NSTJGZ>).

Funding: This research was funded by a Polymath Award from Schmidt Sciences. <https://www.schmidtsciences.org/grantee/sam-gershman/>. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

claims have been mired in criticism and controversy, and contemporary research on planarian learning and memory is limited. The goal of this paper is to revisit basic questions about planarian learning in an effort to establish a behavioral foundation to further investigate these novel memory mechanisms.

Historically, planarian learning and memory has been tested with various forms of classical and operant conditioning [5,6]. The seminal experiment was conducted by James McConnell and Robert Thompson in 1955 [7], after which McConnell became the leading force in planarian learning and memory transfer experiments [8–11]. These classical conditioning procedures were typically conducted with *Girardia dorocephala* flatworms and employed light and shock as the conditioned (CS) and unconditioned (UCS) stimuli, respectively. All studies measured learning by quantifying turn and contraction responses to the onset of the CS after repeated stimulus pairings, and emphasized that the contraction response is more reliable than turns alone or pooled turns and contractions [5,6,12]. Allan Jacobson extended these experiments with additional controls and improved procedures [12–15], after which Jacobson remarked that “[i]t appears reasonable to assert that by the strictest criteria, classical conditioning has been conclusively demonstrated in planarians” [12]. Nonetheless, several groups that attempted replication failed to classically condition planarians or did so with only weak levels of learning. Though a review by Corning and Kelly [6] counted 36 inductions of classical conditioning using the original light-shock paradigm as measured by conditioned response acquisition, these were weighed against seven published failures, which included the mostly unsuccessful efforts of biochemist and Nobel laureate Melvin Calvin [16,17]. Other forays into planarian learning also yielded negative results [18]. By the end of the 1970s, most planarian psychologists, including McConnell, had shifted their area of research [19]. The procedural ambiguities of how to reliably train and elicit conditioned behavior from planarians were never clarified, perhaps due in part to the reluctance of subsequent planarian researchers to associate themselves with the controversial legacy of the memory transfer episode [20].

Modern attempts to classically condition or otherwise train planarians have been few in the years since the initial movement [21–27], even as planarians became a fixture of regenerative biology [20,28–32] and the subject of some neuroanatomical and pharmacological interest [33–35]. Research into planarians’ cognitive capacities has been renewed by recent investigations into non-synaptic forms of memory [36–41], which has returned attention to the McConnell-era transfer experiments and apparent ability of planarians to remember learned information even after decapitation and whole-brain regeneration [42,43]. Nonetheless, the foundational classical conditioning experiments have not been exactly replicated. This is surprising given the volume of the historical literature, the utility of planarians as a model of evolutionarily ancient brains, and the disruptive implications of regenerative memory retention to the neuroscience of memory.

We therefore sought to replicate the original planarian conditioning experiments in alignment with Allan Jacobson’s protocol [13,14]. We selected this particular protocol because it included the most rigorous set of controls (pseudoconditioning) to disambiguate true conditioning from sensitization or habituation (Methods). We refer to this

protocol and our reproduction of it as the “historic” paradigm. We then tested a contemporary conditioning paradigm inspired by aversive conditioning in rodents (referred to as the “contemporary” paradigm), as well as a selection of ethologically-motivated alternative stimulus pairings (“variant” paradigms). In all cases, we were unable to find evidence for classical conditioning in either *Girardia dorocephala* or *Schmidtea mediterranea* planarians. To address the possibility that these null results arose from the subjectivity of manual scoring (with high interscorer variability), we built a computer vision pipeline to analyze worm behavior across several days of training. The results of the computer vision analyses were likewise null.

2 Results and discussion

In what follows, we describe our findings from attempting to classically condition *G. dorocephala* flatworms using (1) a historic protocol based on the work of Allan Jacobson [12–15], (2) a contemporary protocol based on the work of James et al. (2026) [26], and (3) variant stimulus pairings informed by recent studies on planarian sensory capacities [13,26,44–46]. While our search space for a classical conditioning paradigm in planarians is not exhaustive, our assays show that salient stimuli across multiple sensory domains fail to elicit an increase in conditioned response rates comparable to those found in the original studies. We found no difference in conditioned response acquisition for any paradigm across two species: *Girardia dorocephala*, which were historically used for psychological experiments, and *S. mediterranea*, which are commonly used in contemporary regenerative biology research. Finally, we found no difference in conditioning across three separate *G. dorocephala* lineages from multiple locations across the United States (see Methods for more details).

2.1 Historic light-shock conditioning paradigm fails to produce conditioned response in planarians

We recreated the historic light-shock conditioning paradigm, in which *G. dorocephala* planarians are subjected to two sessions of light habituation followed by five sessions of paired exposure to light and shock, remaining as faithful to the original protocol as possible with modern equipment (Fig 1A–B; Methods). The classical conditioning group (CC; also called forward conditioning) received repeated pairs of light stimuli ending with an overlapping shock stimulus; the pseudoconditioning group (PC) received the same number of exposures to light and shock, but in a randomized and non-overlapping order. This protocol controls for non-associative sensitization: if pseudoconditioned worms acquire a response to light comparable to that of the CC group, the response increase can be seen as an indication of sensitization rather than a stimulus-specific conditioned response or learned behavior.

In keeping with the historic experiments’ methods, observed planarians were hand-scored for contraction responses from the onset of the CS to the onset of the UCS (two-way interscorer percent agreement: 91.92%, but see Results 2.6). We found no significant increase in the number of conditioned responses per block of light stimuli across days; rather, we observed a significant decrease in the conditioned response rate across days (two-sided per-worm slope test: $t(11) = -3.86, p = .003, n = 12$ worms; Page’s $L = 491.0, p = .998$; Fig 1C). There was no significant difference in conditioned responses between the CC and PC conditions (ANOVA main effect of condition: $F(1, 22) = 2.88, p = .104$). These results were markedly different from the stark increase in conditioned responses observed in the original study (Fig 1D; original figure reprinted in S3 Fig): the mean per-worm slope for our CC group was -0.02 ± 0.01 SEM responses/block, compared to Jacobson’s $+1.40 \pm 0.40$ SE responses/block (Welch’s $t(3.0) = -3.53, p = .039$, propagating the SE of Jacobson’s 5-block-mean regression; 95% CI on the slope difference $[-2.70, -0.14]$, excluding zero). The failure to replicate this result using the same species, stimulus parameters, apparatus, and pseudoconditioning controls suggests that either a meaningful phenotypic change in this species of worms has occurred since the original experiments, or that there is a critical but undocumented methodological detail that our recreation of the classic experimental setup fails to capture.

2.2 Individual differences in sensitivity to conditioned stimulus do not affect acquisition of the conditioned response

To investigate whether individual variability in response sensitivity to the conditioned stimulus may have obscured evidence of learning [47], we assayed six groups of six worms each with repeated presentations of the conditioned stimulus

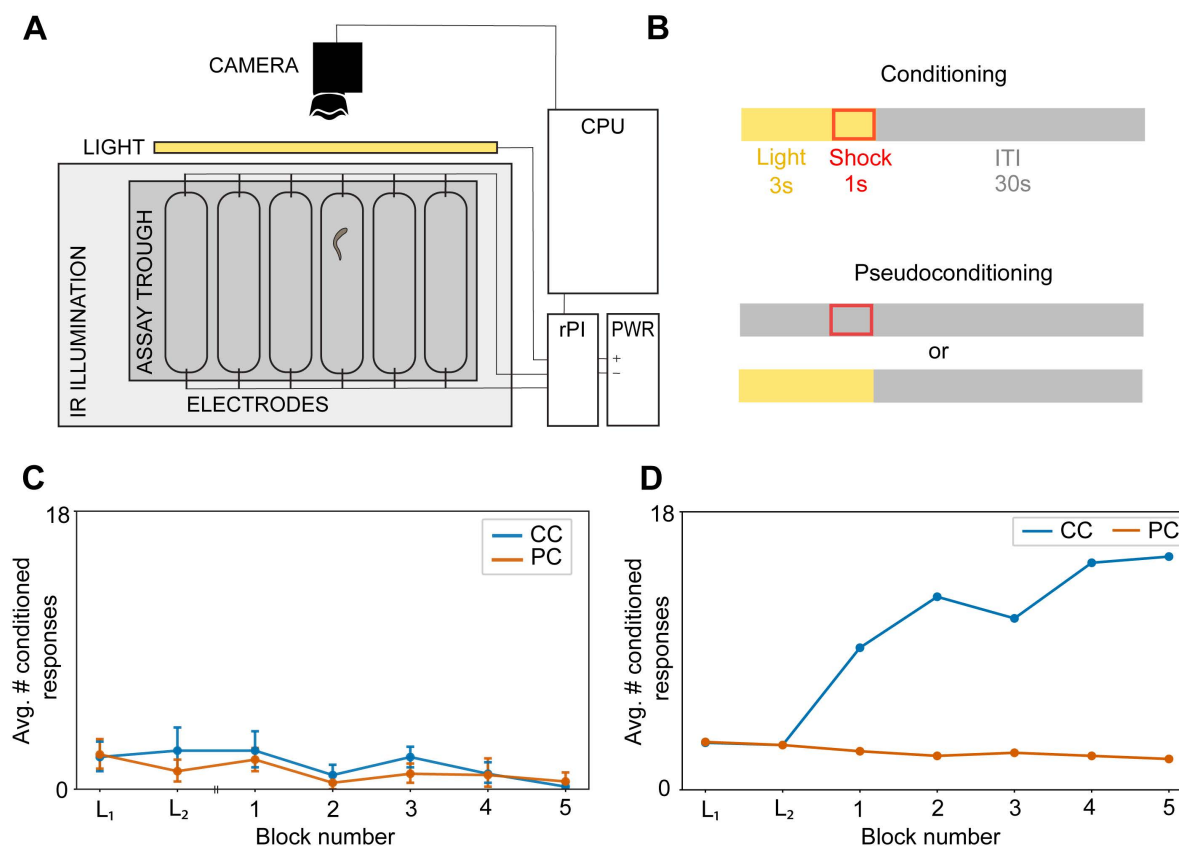


Fig 1. Recreation of historic conditioning paradigm. **A.** The experimental apparatus was a plastic trough of six parallel lanes with electrodes on the ends of each lane. A light source above the trough delivered the CS (white light), and an IR LED panel below the trough provided imaging illumination invisible to the planarians. A camera mounted above the trough recorded the experiments and a raspberry-Pi-based controller (rPi) controlled the delivery of CS and UCS (shock) stimuli in conjunction with a computer (CPU) and power supply (PWR). **B.** Each CC trial consisted of a three-second light CS that co-terminated with a one-second electric shock UCS (overlap shown by the red shock window within the yellow light window), followed by a 30-s intertrial ITI. In the PC group, the CS and UCS were delivered separately as unpaired shock-only trials or unpaired light-only trials, randomly interleaved (shown stacked). **C.** Mean number of conditioned responses across blocks, pooled across cohorts of *G. dorotocephala* worms. Blocks L₁ and L₂ are CS-only baseline blocks; blocks 1–5 are CC blocks. Blue line indicates CC cohorts (n = 12 worms); orange line indicates PC cohorts (n = 12 worms). Error bars are 95% CI. Both groups remained near baseline throughout conditioning, with no upward trajectory to suggest learning. The two groups did not differ from each other in overall response level. A stratified breakdown of the same data by baseline-responsiveness is presented in S2A Fig. **D.** Replotted from five published block means in Jacobson et al. (1967). The CC group shows a substantial, monotonically increasing response rate absent from our data. Original figure reprinted in Fig. S3.

<https://doi.org/10.1371/journal.pone.0358756.g001>

alone, then sorted them into low and high responders (S2A Fig). The groups had a significantly different probability of reacting to the conditioned stimulus (white light) presentation (independent samples *t*-test, $t(22) = 5.58$, $p < .001$). We found no significant upward trend in either of the historically standard conditioned responses (turns or contractions) [5, 12] for both the conditioning group (which showed a significant decrease, as reported above) or pseudoconditioning group ($t(11) = -1.82$, $p = .097$), and slopes did not significantly differ between high and low responders in either the conditioning (Welch's $t(10) = -0.27$, $p = .794$, $d = -0.16$) and the pseudoconditioning group (Welch's $t(10) = +0.51$, $p = .619$, $d = +0.30$) (S2B Fig). To minimize the risk of latent inhibition arising from stimulus pre-exposure during the responsiveness assay, we spaced the assay and the conditioning paradigm 2–3 days apart, during which the worms were only exposed to ambient light. Across the groups with high enough response numbers to discern from zero, we found no significant habituation to the conditioned stimulus (but see [6], Table VIII). The mean response rate increased significantly from Day 1 (11.8%

$\pm 2.0\%$ SEM) to Day 2 ($17.5\% \pm 2.5\%$ SEM; paired $t(35) = 2.39$, $p = .022$, two-sided; Cohen's $d_z = 0.40$). While this increase was consistent across five of six cohorts, no individual cohort reached significance alone. Slope analysis at the trial level did not reveal any significant gradual trend across trials (mean slope = 0, $t(35) = 0.16$, $p = .873$), suggesting the increase reflected a between-session shift rather than a continuous within-session change (S2C Fig). These findings rule out individual variation in baseline CS sensitivity as an explanation for our overall null result.

2.3 Feature-based quantification of the conditioned response in contemporary paradigm fails to provide evidence for learning

After the historic conditioning protocol yielded no appreciable evidence of learning, we attempted to replicate a contemporary conditioning protocol adapted from fear conditioning experiments in rodents [26] using *G. dorotocephala* and the distantly related planarian *S. mediterranea*, a common model organism in regenerative biology today [31,32]. To quantify the conditioned responses in this protocol, we used both manual annotation of turns and contractions and a novel computer vision pipeline (Fig 2A). From the hand-scored data we found that neither conditioned turns nor contractions showed a significant upward trend over training days in the conditioning group (turns: per-worm slope test, $t(41) = -1.95$, $p = .97$, one-sided; Page's $L = 1026.5$, $p = .90$; contractions: $t(41) = -0.96$, $p = .83$; Page's $L = 1035.5$, $p = .78$; $n = 42$ worms) or the pseudoconditioning group (turns: $t(11) = -1.11$, $p = .86$; contractions: $t(11) = 0.35$, $p = .37$; $n = 12$ worms). Linear mixed models (LMMs) confirmed no significant Day $\times$ Condition interaction for either turns ($\beta = -0.18$, $z = -0.21$, $p = .83$) or contractions ($\beta = 0.12$, $z = 0.36$, $p = .72$), indicating that the conditioning group did not produce responses at a different rate than controls. The two groups did differ in overall turn frequency (LMM: $\beta = -7.65$, $z = -3.12$, $p = .002$), with the conditioning group responding more than controls, but this difference was present from the outset and did not change with training (Fig 2B). Breaking the conditioning group down by species and strain (S1 Fig) confirmed this pattern in every measurable sub-population: per-worm conditioned response slopes were significantly lower than the historic learning rate of +1.40 responses/day (Welch's t -test, with denominator df adjusted to propagate the SE of Jacobson's regression). By subpopulation and conditioning paradigm: *S. mediterranea* CC (-1.75 responses/day, $n = 6$, $t(7.5) = -3.95$, $p = .005$), *G. dorotocephala* (Michigan, USA) CC (-0.33 , $n = 6$, $t(7.5) = -3.06$, $p = .017$), *G. dorotocephala* (stock) CC (-0.86 , $n = 30$, $t(21.3) = -3.08$, $p = .006$), and the matched *G. dorotocephala* pseudoconditioning reference (-0.43 , $n = 12$, $t(11.2) = -2.95$, $p = .013$); no sub-population produced a positive group-mean slope, and all were significantly lower than the Jacobson reference slope (Welch's t -tests, all $p \leq .017$).

Because variability in subjective manual scoring could obscure genuine group differences, we built a custom feature-extraction pipeline using Segment-Anything (Meta Inc.) and shape-analysis methods previously applied to planarian behavior [44,48]. Following a recently developed approach for comparing the acquisition of learned behavior across species and paradigms [49], we quantified worm behavior by calculating the Kullback–Leibler (KL) divergence between distributions of body-shape features sampled during conditioned-stimulus-on (CS-on) and intertrial (CS-off) intervals (Fig 2A). Intuitively, the KL divergence measures how much the presence of the CS changes the distribution of morphological features such as body area or perimeter. A worm that consistently responds to light will look systematically different during CS-on periods than during the ITI, producing high KL divergence, whereas a worm that ignores the CS will look similar in both windows, producing low KL divergence. Consequently, if animals acquire a conditioned response over training, the divergence between CS-on and CS-off feature distributions should increase across days, and should do so more steeply in the conditioning group than in the pseudoconditioning group.

We fit a linear mixed model predicting log-transformed KL divergence from Day (continuous), Condition (conditioning vs. pseudoconditioning), and their interaction, with random intercepts for individual subjects. The interaction term was not significant ($\beta = -0.08$, $z = -0.95$, $p = .342$), indicating no evidence that the conditioning group's trajectory differed from that of the pseudoconditioning group. This was corroborated by a direct comparison of per-subject slopes between groups (conditioning: $+0.33 \pm 0.13$; pseudoconditioning: -0.04 ± 0.31 ; Welch's $t = 1.11$, $p = .286$). The two groups did differ

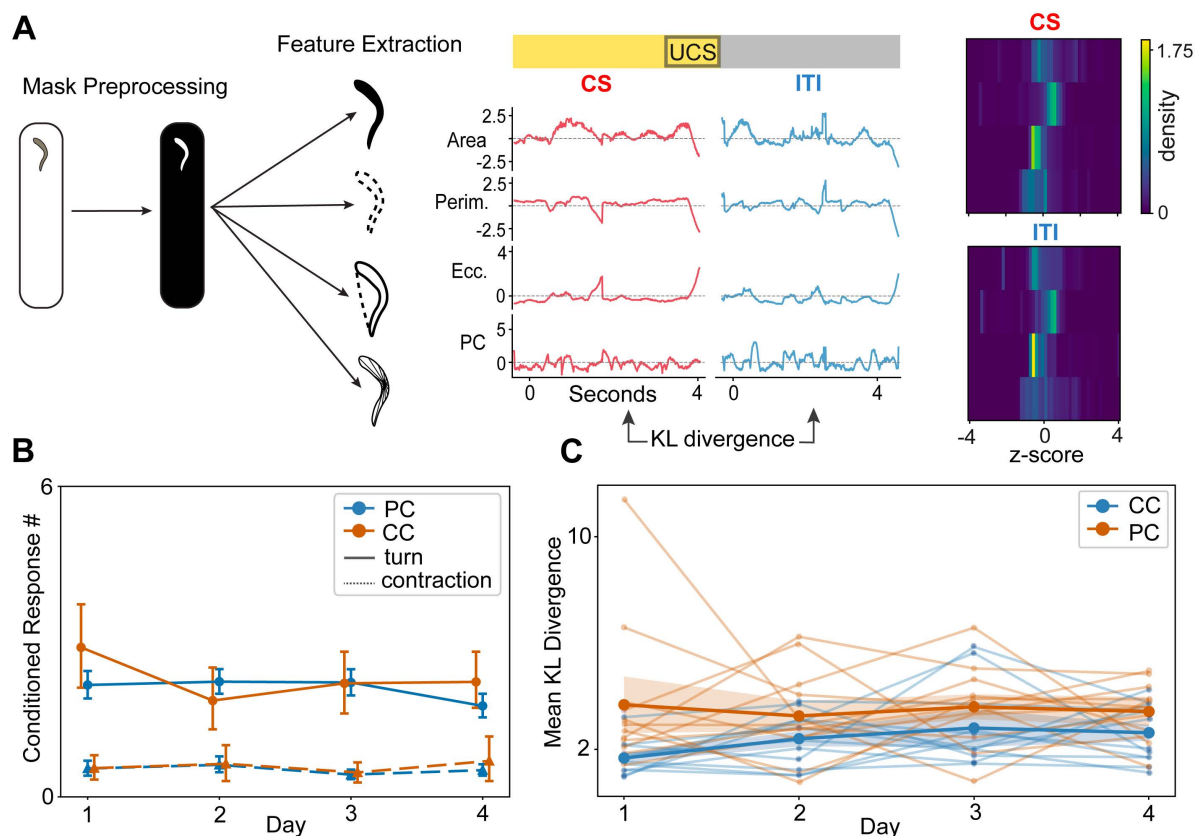


Fig 2. Application of machine vision pipeline to assess learning over the course of a contemporary conditioning protocol. **A.** Feature extraction pipeline. Four of ten morphological/spatial features shown (Area, Perimeter, Circularity, PC1). Traces show z-scored feature values for an example worm during CS-on (red) and ITI (blue) periods; heatmaps show corresponding feature distributions. Increasing area with stable perimeter indicates lateral body spread; spikes in circularity with decreasing perimeter indicate contraction. **B.** Hand-scored conditioned responses across training days (CC: $n=42$, blue; PC: $n=12$, orange), pooled across species and strains. Solid lines: turns; dashed lines: contractions. Error bars indicate 95% bootstrap CIs. Species-level breakdown in S1B Fig. **C.** Computer-vision-derived KL divergence (CS-on vs. ITI) across training days for stock *G. dorotocephala* (CC: $n=12$, blue; PC: $n=12$, orange). Thin lines: individual worms; bold lines: group means $\pm$ 95% bootstrap CI. Note that neither group shows a systematic upward drift and PC trajectories are if anything slightly higher than CC trajectories at every day, in the opposite direction to what learning would predict, and the two group-mean trajectories overlap within their CIs at all four time points.

<https://doi.org/10.1371/journal.pone.0358756.g002>

significantly in overall KL divergence ($\beta = 0.35$, $z = 3.61$, $p < .001$), though notably the pseudoconditioning group showed a higher mean KL divergence, which is opposite to what a conditioning effect would predict (Fig 2C). The higher mean KL divergence in the pseudoconditioning group may reflect greater behavioral variability induced by unpredictable stimulus timing, though we do not have a principled account of this difference. Critically, the convergence of two independent measurement methods (manual scoring and automated computer vision) on the same null result substantially reduces the likelihood that the negative outcome reflects a measurement artifact.

2.4 Ethologically relevant stimulus pairings fail to evoke a robust conditioned response

Modern literature on planarian sensory capacity motivated us to substitute light and shock for ultraviolet (UV) light, which is a more ethologically relevant stimulus [44–46]. Planarians have UV-sensitive extraocular opsins distributed across their bodies that mediate a rapid contraction in response to high-amplitude UV light exposure [45,50]. Low-amplitude and slowly modulating UV light evoke a more subtle and complex response [51], and are suitable as conditioned stimuli. We

selected two new stimulus pairings (CS and UCS): (1) white light and high-amplitude ultraviolet (UV) light, and (2) low-amplitude UV light and shock. We suspected that planarians may display implicit biases for detecting and associating these sets of stimuli, making it easier to elicit a robust conditioned response.

We first used low-amplitude UV light as the CS and shock as the UCS, reasoning that low-amplitude UV would evoke a salient yet suppressible response. In the historically used species (*G. dorotocephala*, $n=18$ worms), the per-worm acquisition slopes across five conditioning blocks did not increase (mean slope = -0.07 ± 0.07 SEM responses/block, 95% CI $[-0.22, +0.08]$) (Fig 3A). We then used white light as the CS and high-amplitude UV as the UCS, choosing the latter because planarians do not habituate to phasic, high-amplitude UV exposure [44]. Neither species we tested showed an upward trend (Fig 3B); the per-worm slopes were -0.67 ± 0.44 SEM responses/block in *G. dorotocephala* ($n=6$, 95% CI $[-1.80, +0.47]$) and -0.67 ± 0.49 SEM responses/block in *S. mediterranea* ($n=6$, 95% CI $[-1.94, +0.60]$). Across all three protocol $\times$ species combinations, the per-worm acquisition slope was significantly lower than the slope reported by Jacobson et al. (1967) (Fig 3C). The Jacobson CC reference slope, recomputed from his five published block means, is $+1.40 \pm 0.40$ SE responses/block. Two-sided Welch's t -tests comparing each of our populations to the Jacobson reference rejected equality

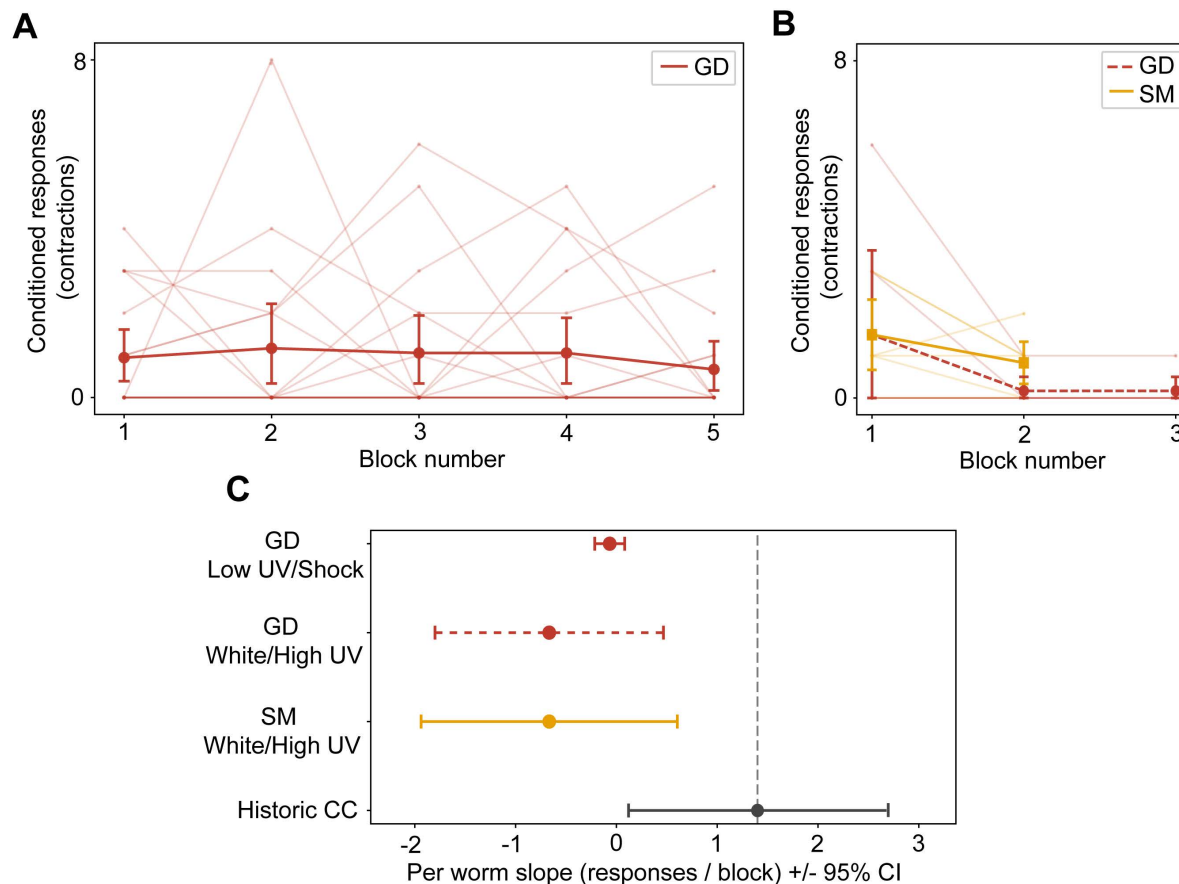


Fig 3. Failure to elicit learning using ethologically-motivated conditioning stimulus variants. **A.** Low-UV + Shock paradigm (*G. dorotocephala*, $n=18$). Thin red traces: individual worms; thick dark-red line: pooled mean $\pm$ 95% bootstrap CI. **B.** White + High-UV paradigm in stock *G. dorotocephala* (dashed red, $n=6$) and *S. mediterranea* (solid yellow, $n=6$), each completing the maximum number of conditioning blocks achievable for that species. Thin traces: individual worms; thick lines: group means $\pm$ 95% bootstrap CIs. **C.** Forest plot of per-worm OLS acquisition slopes (filled circles, 95% CI) versus the historic reference ($+1.4$ responses/block, dashed vertical line). Historic CC slope recomputed by OLS on five published block means ($df=3$). Note that all groups' slopes are significantly lower than the historic learning rate.

<https://doi.org/10.1371/journal.pone.0358756.g003>

at $\alpha = 0.05$ in every case: GD low-UV/shock $t(3.2) = -3.59$, $p = .034$ (95% CI on the difference $[-2.72, -0.21]$); GD white/high-UV $t(7.8) = -3.46$, $p = .009$ $[-3.45, -0.68]$; SM white/high-UV $t(8.0) = -3.24$, $p = .012$ $[-3.54, -0.60]$. In each of our populations the 95% CI excluded Jacobson's point estimate of +1.40, and the standardized effect against the fixed Jacobson reference was large (one-sample Cohen's $d = -4.91$, -1.91 , and -1.71 , respectively). These results indicate that neither variant using UV light replicated the conditioning rate originally reported by Jacobson et al.

2.5 Planarian species and origin have no observable effect on learning

Literature from the 1960s claimed that worm origin, species, and feeding interval (among many other factors) are important for the success of planarian learning [52]. To test whether worm provenance and species affected our worms' learning outcomes, we collected *G. dorotocephala* specimens from three locations across the United States, including the location (to the best of our knowledge) where McConnell procured his population of planarians for most (but not all) of his experiments. Commercially available *G. dorotocephala* from North Carolina (also called stock planarians), wild-caught Michigan *G. dorotocephala*, wild-caught California *G. dorotocephala* and laboratory *S. mediterranea* worms were used across our previously described experiments and showed indistinguishable levels of responding. While different species demonstrated notably different levels of general activity, we did not find that a particular species had an effect on the acquisition of conditioned responses (Fig 3A; S1 Fig). Laboratory *S. mediterranea* are from the Mediterranean [53], whereas *G. dorotocephala* are native to North America, and their evolutionary distance is possibly as vast as 100 million years [54]. Thus, the failure of American (Michigan, California, and Carolina strains) and Mediterranean species to demonstrate learning of conditioned responses is more consistent with a paradigm-level failure than with any species-specific deficit.

2.6 Behavioral scores are not consistent across trained observers

Having found no evidence of learning as measured through increases in conditioned response rates, we endeavored to determine the extent to which different scorers could agree on what types of behavior count as turns and contractions. For the historic paradigm, trials were scored in a binary fashion and 91.92% of trials were given the same value ("0" for no contraction or "1" for a contraction) by both scorers. This agreement value is dominated by trials where each scorer agreed that no response had occurred (89.67% for "0-0" and only 2.25% for "1-1"). When looking only at trials where at least one scorer reported a positive response, the interscorer agreement is 21.78%. For all trials, including those rated "0-0", Cohen's $\kappa = 0.31$, $P_0 = 0.92$, and $P_e = 0.88$. A reliable, robust assessment of contractions does not follow from these data.

In the contemporary paradigm, rather than a binary assessment of contraction, behavior was scored as counts of turns and contractions across the active duration of the conditioned stimulus, so a more complex metric of interscorer agreement was required. Three scorers scored a sample of 203 videos of our footage and timestamps of turns and contraction events were analyzed (see Methods). While the total sums of turns and contractions for each video are essentially the same for each scorer (Fig. S4), an analysis of the actual timestamps of reported responses reveals severe disagreement. The primary result is that unanimous agreement of contraction events (all scorers marked a given event in a given video as a contraction) is only 28.15% (67/238 of all contraction events). The agreement for turn events is even lower at 15.14% (106/700 of all turn events). These results echo those of Arlene Hartry [18], who similarly examined unanimous contraction agreement between three observers of planarian conditioning footage and found that so "few responses fell into this category" that they were statistically irrelevant. The low agreement of multiple scorers is in marked contrast to conventional classical conditioning procedures, where it is understood that conditioned responses should be relatively obvious to the trained eye.

To further examine interscorer agreement at the timestamp level, we computed F1 metrics for each pairwise combination of the three scorers' annotations for each video. The F1 metric is the harmonic mean of precision and recall measures (Methods). In the context of our data, precision is the number of the Test Scorer's events (turns or contractions) that the

Reference Scorer also marked. Recall is the proportion of Reference events that the Test Scorer correctly identified out of all of the Reference's events. These proportions describe how much the Test Scorer is able to capture a Reference Scorer's data while minimizing the detection of non-reference events. If a pair of scorers agreed perfectly, an F1 score of 1 would be expected, and severe disagreement would yield a 0. We also quantified event confusion based on the proportion of unpaired Reference events that have a temporally proximate Test event of the opposite type in the test events. These instances occur when the Test and Reference Scorer recognize the same event as having occurred, but categorize it differently.

Table 1 shows the recall, precision, F1, and confusion rates for each scorer pair and event type. Broadly, these measures indicate poor agreement about when events occur across scorers: on average, a given scorer detects about half of the turns (52.3%) or contractions (59.7%) that another scorer marks, and produces an equal amount that their reference fails to detect (55.6% for turns, 57.8% for contractions). To compare contraction agreement in response to the conditioned stimulus (light) versus the unconditioned stimulus (electric shock), Scorer 1 and Scorer 2 additionally examined 198 clips from the same set of videos for responses during the shock period. For shock-induced contractions, Scorer 1 and Scorer 2 agreed on 86.87% of trials (172/198). The discrepancy between high agreement about shock-induced contractions but low agreement about light-induced contractions fails to provide evidence that the CS-period contractions are specific, learned responses to light and not just ambient, noisy movement. This interpretation is supported by the observation that the total number of observed turns and contractions is essentially consistent across scorers despite starkly different time-stamp attributions for those behaviors (Fig. S4). The sum of these results suggests that manual assessment of planarian behavior is not robust for our experimental footage. That is, when multiple scorers observe and describe the same experimental footage, their reports are not reliably similar, even if the raw counts of turns and contractions are stable across scorers.

Our finding of low interscorer agreement is concerning for the interpretation of the historic literature on planarian conditioning, all of which relies on manual scoring. Though several of the classic studies do report high interscorer agreement, in many cases the complete data, the precise interscorer methods, and the experimental footage have been lost or are simply not available for review. The inconsistency in manual scoring of conditioned responses suggests, at least for our data, that we are attempting to quantify an absent learning signal, which is a poor foundation upon which to investigate potentially paradigm-shifting questions about memory transfer and the molecular basis of memory.

2.7 General discussion

Across three conditioning protocols, two species, four geographic populations, and two independent measurement methods, we found no evidence for the robust classical conditioning of freshwater planarians reported in historic literature. The

Table 1. Pairwise comparison of scorer agreement across behaviors as measured by recall, precision, and F1 scores of event timestamps of 203 sample training videos. Note that when a Reference and Test Scorer are flipped, the recall and precision values likewise swap places, but the F1 score remains the same.

Behavior	Reference	Test	Recall	Precision	F1	Confusion Rate
Turns	Scorer 1	Scorer 2	0.540	0.745	0.626	0.029
Turns	Scorer 2	Scorer 3	0.397	0.640	0.490	0.009
Turns	Scorer 3	Scorer 1	0.632	0.284	0.392	0.000
Turns	Average	—	0.523	0.556	0.503	0.013
Contractions	Scorer 1	Scorer 2	0.827	0.602	0.697	0.000
Contractions	Scorer 2	Scorer 3	0.492	0.556	0.522	0.031
Contractions	Scorer 3	Scorer 1	0.473	0.576	0.519	0.178
Contractions	Average	—	0.597	0.578	0.579	0.070

<https://doi.org/10.1371/journal.pone.0358756.t001>

consistency of this null result makes a simple procedural explanation unlikely. We suggest that the conditions for eliciting robust and reproducible conditioned behavior in the tested planarian species are highly idiosyncratic, and we encourage future research to explore more ethologically guided stimuli that are more likely to produce robust learning behaviors.

It is difficult to reconcile our null results with the 20th century planarian learning literature. Several laboratories (not only the laboratories of McConnell and his students) reported success in training planarians, and over fifty scientific papers were published reporting successful conditioning of planarians between 1959 and 1971 [6], including in high-impact journals like *Nature*. Additionally, editions of the *Worm Runner's Digest* contain letters from students and laypeople describing their efforts to condition planarians with an approximately 4:1 ratio of success to failure [17]. Publication bias toward positive results may partly explain the apparent preponderance of successes in the historic literature, though this cannot be the whole story given that different labs independently replicated a positive result. Other demonstrations of apparent learning include instrumental and maze paradigms (e.g., [55–57]), which we have not attempted to replicate here. Modern literature suggests that *C. elegans*, which have far fewer neurons than planarians (302 in hermaphrodites compared to 10,000–12,000 in the planarian brain) are capable of aversive learning [58–61]. We believe our efforts do not reflect on the widespread ability of planarians to learn in any situation, but conclude that we were unable to specifically elicit classical conditioning in the historically championed species *G. dorotocephala* or commonly used species *S. mediterranea* using historic and contemporary conditioning paradigms. Ultimately, we cannot explain why our results are in such contrast to those in the historic literature.

We close with a quote from a 1964 paper by Bennett and Calvin [16], who attempted to train planarians on a variety of tasks and were mostly unsuccessful:

“Until more adequate and more useful methods and descriptions of methods to train planarians are available, this animal appears to have little utility for proposed studies of the possible biochemical basis of memory.”

This assessment appears, at least in our hands, to hold true today. Going forward, we suggest that researchers methodically document their procedures, data, and footage demonstrating learned behavior in the context of the selected conditioning protocol. Perhaps more importantly, we recommend exploring alternative stimulus combinations informed by planarian ethology and ecology. Despite the failure to replicate historic results, we are optimistic that future work will discover a robust behavioral foundation for conditioning, providing an unprecedented model through which to study neural circuit development and the physical basis of learning and memory.

3 Materials and methods

There are three major paradigms in this study. The first is our attempt to reproduce the classic 20th century planarian light-shock conditioning experiments using a similar apparatus and set of experimental parameters as those used by James McConnell and Allan Jacobson (stimulus timing, planarian species, etc.). We refer to this as the “historic” classical conditioning paradigm. The second paradigm we call “variant” because they represent an exploration of novel stimulus configurations and timing parameters that are unprecedented in the planarian memory literature. Third comes the “contemporary” paradigm, which seeks to replicate a recent paper that used a spaced conditioning regime adapted from rodent experiments to classically condition planarians. The particular details of these three paradigms are summarized in [Table 2](#) and given in detail below.

3.1 Planarian acquisition and care

Planarians were housed in 6- and 11-cup rectangular Pyrex dishes with air holes carved into the plastic lids. The dishes were filled to a level of 2–3 inches with Montjuic water (1.6 mM NaCl, 1.0 mM CaCl₂, 1.0 mM MgSO₄, 0.1 mM MgCl₂, 0.1 mM KCl, 1.2 mM NaHCO₃ in deionized water). Housing dishes were cleaned biweekly on Tuesdays and Fridays. Cleaning

Table 2. Tabulation of each classical conditioning paradigm's experimental parameters for our attempted replications of the historic protocol, our variations of the historic protocol with alternative stimuli, and the contemporary protocol.

Different paradigms used in this study.			
	Historic replication	Variant paradigms	Contemporary replication
Role in study	Attempted replication of historic Jacobson [13,14] studies	Explorations of alternative stimuli that might more readily elicit learning	Attempted replication of recent planarian conditioning work based on rodent conditioning [26]
Species	<i>G. dorotocephala</i> , Carolina strain	<i>G. dorotocephala</i> of multiple strains and <i>S. mediterranea</i>	<i>G. dorotocephala</i> of multiple strains and <i>S. mediterranea</i>
Apparatus type	Parallel lanes with electrodes (PLA material)	Parallel lanes with electrodes (PLA material)	Parallel lanes with electrodes (milled plastic)
CS	3 sec white light	White or ultraviolet light	30 sec blue light
UCS	1 sec AC shock, coterminating with CS	Coterminous 1 sec UV light or AC shock	3 sec AC shock, coterminating with CS
Control	Pseudoconditioning by alternating light and shock	None	Pseudoconditioning by alternating light and shock
Trial timing	125 trials in 5 blocks of 25 trials across 2 days	25 trials per block in 3–5 blocks over 2–3 days	24 trials in 8 blocks of 3 trials across 4 days (2 blocks/day)
Intertrial interval	30 seconds	2 minutes	30 seconds
Interblock interval	10 minutes	10 minutes	4 hours
Scoring method	Hand-scored	Hand-scored	Hand-scored and computer vision pipeline
Other	Before training, worms received 2 blocks of 25 light-only trials.	Control conditions not run due to low response in noncontrol worms	See Methods for procedural divergence from original study

<https://doi.org/10.1371/journal.pone.0358756.t002>

consisted of corralling planarians into one corner using a water jet created by a transfer pipette or spray bottle containing fresh Montjuic water, draining the water from the dish, and wiping down the floor and walls with unbleached paper towels. The corner where the planarians had been concentrated was not wiped down, leaving an untouched triangular region with an area of 2–3 square inches. The unwiped corner was alternated between opposing sides with each cleaning so that a continuous zone of slime was always present in the trough. For experiments, planarians were individually isolated into six-well plates (Falcon, item no. 353046) containing approximately 10 mL of Montjuic water four days prior to the start of training.

Planarians were fed with freshly hatched *Artemia* brine shrimp nauplii (Brine Shrimp Direct, item no. BSEA8Z), hatched in salt water according to vendor instructions. For feeding, the hatchlings were rinsed with Montjuic water and transferred to the planarian dishes with a transfer pipette. Confirmation that the planarians ate the shrimp came in the form of pursuit of the shrimp, visible pharyngeal contractions, and the tendency for planarians to take on the color of their food (orange-red with brine shrimp). Planarians were fed once or twice per week, and in the case of experiments, starved for seven days prior to testing.

The planarians used in these studies varied in source. The historic experiments used worms purchased from Carolina Biological Supply (item no. 132954), which are wild-caught from various water sources in North Carolina. Miscellaneous experiments using alternative stimuli used the Carolina *G. dorotocephala* as well as *S. mediterranea* worms originating from the laboratory of Peter Reddien. Experiments modeled after the contemporary James et al. study used the Carolina *G. dorotocephala* worms and *S. mediterranea*, as well as a strain of wild-caught *G. dorotocephala* descended from an individual acquired from the Huron River in Ann Arbor, Michigan, where James McConnell is suspected to have sourced worms [62]. This founder planarian was found June 12, 2025, on the bottom of a rock in the shallow pool at the base of the Argo Dam on the outflow side (coordinates: 42°17'24.7"N, 83°44'42.6"W). Its species was determined visually by its

prominent auricles and brown, unspotted coloration, and distinguished from the similar species *G. tigrina* by its comparatively weaker adhesion to glass. Once in the laboratory, this individual worm was amplified into a colony by repeated cycles of bisection and regeneration. In a similar fashion, two worms of the same species were acquired from a stream on the University of California, Berkeley campus, which were used in a variant (White-UV) experiment.

3.2 Assay chamber, stimuli, and data acquisition

We designed an assay chamber resembling side-by-side duplicates of McConnell's original chamber [63]. Six separate roundbottomed, ovoid lanes were milled from a transparent block of polycarbonate and header pins were inserted at the ends of each lane. Sets of ipsilateral pins were connected by a copper wire to transmit the electrical shock stimulus to all lanes simultaneously (Fig 1A). Two alligator clips attached to the pin-spanning copper wires connected each side of the trough to a DC power supply via an H-bridge module that together produced a 10–12 V, 60 Hz square-wave AC shock stimulus. While there is debate regarding the extent to which the orientation of a DC current affects conditioning in planarians [64], the primary historical study we are replicating used alternating current [14]. Experiments were recorded at 10 Hz with an overhead camera (Teledyne FLIR GS3-U3-23S6C-C (2048px x 2048 px) or GS3-U3-41C6M-C (4096px x 3000px)) suspended approximately one foot above the trough. A long-pass filter (750 nm) was placed over the camera lens to eliminate saturation effects from the light stimulus. The trough was placed on a white acrylic panel which diffused another panel of infrared (IR) LED lights (850nm, Waveform Lighting flexstrips) below, so the worms were clearly visible from the overhead camera. For the white light in the historical reproduction, we employed a 100-Watt incandescent bulb (Fluker's Incandescent Basking Spotlight Bulb). Where ultraviolet light was a stimulus, a 100-Watt LED UV panel (395nm Glostars 100W LED, IP66) was used. In the case of the low ultraviolet condition, a diffuser was placed over the panel to reduce the light intensity. For the contemporary paradigm, we used an acrylic panel covered in rows of LED strips (6500K) with the panel itself covered in blue cellophane per the James et al. study. Small IR LEDs were directly coupled to the shock and light stimuli and visible in the camera frame to serve as indicators of stimulus onset for the viewer.

Custom software written in Python controlled the camera, conditioned stimulus, and unconditioned stimulus via a Raspberry Pi Pico (RP2040) microcontroller. Data were collected onto a Linux computer (Ubuntu 20.04.6 LTS, 64 bit) using a custom Python (version 3.8.10) recording script. The assay rig was contained in a rectangular box made of acrylic panels affixed to metal frame to isolate the experimental equipment and reduce ambient illumination.

3.3 Historic classical conditioning paradigm

In Allan Jacobson and colleagues' experiments [13–15], planarians were rapidly and repeatedly presented with overlapping light and shock that resulted in the worms contracting to light even in the absence of shock. We attempted first to reproduce these experiments in their timing stimulus parameters. Planarians of the species *G. dorotocephala* (known as *Dugesia dorotocephala* in Jacobson's time) were purchased from Carolina Biological Supply (item no. 132954) and subjected to 125 trials of 3 seconds of light, the final second of which included a 10–12 V, 60 Hz square-wave AC shock. Trials occurred in five blocks of 25 trials, each trial separated by 30 seconds, with ten-minute rest periods between blocks that were themselves spread over two days. Day 1 contained two consecutive light-only blocks in accordance with the historic paradigm. These were identical to the light-shock blocks, save that the shock did not occur. Presentation of the CS alone without any paired UCS had been reported to diminish the effectiveness of conditioning as far back as 1959 [65], but because the Jacobson group gave these light-only trials, we did likewise. The two Day 1 light-only blocks were followed by two light-shock blocks. On Day 2, worms received three light-shock blocks. In the case of the pseudoconditioning control, light-shock blocks were replaced with light-or-shock blocks. These consisted of randomly alternating light or shock presentation, with the constraint that no more than four consecutive light stimuli and no more than two consecutive shock stimuli were permitted.

Pseudoconditioning is nowadays often implemented by independent Poisson processes for different stimuli, but random alternation is the method for Jacobson's own pseudoconditioning control, so as with the pretraining light-only blocks, we did likewise. The pseudoconditioning was implemented by generating random lists of 25 alternating L and S symbols, then filtering for those that violated the consecutivity restraints. A random candidate LS sequence was selected at the start of each pseudoconditioning block and each trial in that block corresponded to a symbol in the LS sequence. For example, a pseudoconditioning block that selected the sequence "LLS..." would present two light trials and then a shock trial. The training footage was scored on a trial-by-trial basis and conditioned responses were marked when a worm performed a "discernible longitudinal contraction of the body" [13] in the two-second period between light onset and shock onset. This paradigm, along with the others described below, is summarized in Table 2.

We diverged from Jacobson's experiments in select ways consistent with established guidelines for training worms. First, before experiments began on any given day, a separate cohort of approximately 20 non-experimental worms were allowed to "slime" the trough. This was done in accordance with previous research that found that the presence of slime improved memory performance [52,66]. After twenty minutes, the non-experimental slime worms were removed and experimental worms were allowed to acclimate in the trough for ten minutes before experiments began. After a cohort of worms was done for the day, the trough was rinsed with deionized water and scrubbed with unbleached paper towels. Second, whereas the historic Jacobson et al. experiments trained large groups of worms in a single rectangular box as a convenience to acquire hundreds of worms for RNA purification, we elected to train our worms in parallel lanes. We did this because isolated worms are more easy to score behaviorally and because James McConnell's experiments on which Jacobson's are based primarily used a single-worm narrow lane apparatus with success [63]. Our parallel lane trough allowed us to train multiple worms simultaneously without hampering the visibility of individual worms.

Because our preliminary attempts to replicate the historic protocols showed low and unchanging conditioned response rates to light, we attempted to train a subset of worms that were already reactive to light prior to any training. To do this, we presented a subset of worms with two blocks of 20 light-only trials following the same timing characteristics as described above. Responses to light (turns or contractions) were scored for each trial. The twelve highest-responding worms were split into forward conditioning and pseudoconditioning groups and called "high responders"; the next highest-responding twelve worms were split into the "low responders" groups. S2B Fig shows a box-and-whisker plot of these two groups' pre-training responses to light.

3.4 Contemporary classical conditioning paradigm

Because a group at the University of Tasmania had recently reported success in conditioning planarians using different temporal parameters than our experiments, we attempted to reproduce their procedures [26]. In these experiments, adapted from fear conditioning in rodents, planarians were starved for seven days and subjected to six trials per day for four days, with trials split evenly across one morning and one afternoon block starting at approximately 10:00 and 14:00 respectively. Each trial consisted of 30 seconds of blue light exposure that coterminated with 3 seconds of shock, followed by 30 seconds of rest. The control condition consisted of an approximation of pseudoconditioning, with light-only and shock-only trials randomly dispersed throughout blocks, with at least one of each trial type occurring in each block. The day before training trials began, worms were permitted 10 minutes to explore the trough before being returned to their home dishes. On any training day, worms were given ten minutes to acclimate to the trough before trials began. Other experimental details remained the same as the historic experiments. Data from these experiments were assessed by blinded manual scoring and a computer vision pipeline, described below.

We note two methodological departures from the procedures described by James and colleagues. First, their worms appear to have been trained and tested in circular dishes, whereas ours were trained in linear lanes (as in McConnell's work [63]) because this enabled us to train multiple worms in one session. Second, while the James group notes that their planarians, described as *G. dorotocephala*, were ordered from a biological supply house, the

company that sells them informs us that these worms are collected from ponds in Australia (Southern Biological, personal communication, July 11, 2025). Whether these *G. dorotocephala* worms are invaders from North America, a truly endemic Australian variant, or a different species altogether is unknown. The circular training/test environment and different worm strain could also explain the discrepancy between our results and those of the James group, though further work is needed to resolve this.

3.5 Variant conditioning paradigms

Because repeated attempts to condition worms with the historic white light CS and an electric shock UCS (White-Shock) failed, we sought to use other sets of stimuli that might be more salient to planarians. There is some debate as to whether classical conditioning is agnostic to the stimuli used. However, it generally seems that stimuli with special relevance to the natural history of the animal being conditioned are more likely to facilitate learning [67,68]. For this reason, we added ultraviolet (UV) light to the assembly of stimuli in our training paradigms. We selected ultraviolet light as an additional stimulus because although electric shock is certainly aversive to planarians [69], we are unaware of any ethologically relevant situation where North American planarians would need to respond to electrical shocks. Furthermore, the repeated full-body shock of a submerged organism may cause side effects that could affect worms' behavior in a way that reflects acute injury rather than learned behavior. In contrast, planarians have dedicated sensory machinery to detect and respond to ultraviolet light, which includes a full-body contraction to sudden onset of intense UV [45,46]. These morphological and behavioral characteristics are easily reconciled with planarians' benthic, photophobic lifestyle. Thus, the miscellaneous variant paradigms consisted of combinations of UV light, white light, and shock.

We also varied the species and strain of the worms. In total, we intentionally did not cover the exhaustive space of these stimulus and strain combinations; rather, we sought to rapidly test a subsample for any indication of planarian learning and converge on promising candidate paradigms. Because no promising candidates emerged, we did not pursue these paradigms further.

1. **White-UV:** The same white light from the White-Shock paradigm was paired with an ultraviolet pulse instead of shock. Two *G. dorotocephala* worms collected from Berkeley, California were subjected to 3 seconds of white light coterminating with 1 second of UV light. Worms were trained in three blocks of 25 trials (2-minute ITI), with the first block occurring on Day 1 and the next two blocks on Day 2. A ten minute rest period separated neighboring blocks. Contractions to the onset of white light were scored for each trial. An additional six *S. mediterranea* worms were trained identically, save that they only received two blocks. These experiments took place between 9:30 and 14:30.
2. **UV-Shock:** The white light from White-Shock was swapped with a UV light instead and reduced in strength so as not to trigger baseline contractions. Eighteen *G. dorotocephala* worms were subjected to five blocks of 25 trials of three seconds of UV with onesecond of coterminating shock using an ITI of 2 minutes. Blocks one and two occurred on Day 1, three and four on Day 2, and the fifth on Day 3. Worms were allowed 10–20 minutes to acclimate to their trough lanes before experiments began and given a ten minute rest period between blocks. The shock characteristics were the same 10–12 V, 60 Hz square-wave AC described in our historic protocol. Worms were scored for contractions to the onset of the UV light. These experiments occurred between 10:00 and 16:00.

Because the unconditioned response to a sufficiently high UV stimulus is in itself a contraction, ultraviolet light cannot be considered neutral stimulus in the conventional sense of the term. However, as argued by Jacobson [12], even if conditioned and unconditioned stimuli elicit the same response (contraction), learning can still be assessed by the increasing difference (or lack thereof) in response rates between pseudoconditioning controls and forward/classical conditioning worms. In any case, because we observed no increase in responding among worms trained on any of these paradigms, pseudoconditioning controls were not necessary.

3.6 Manual scoring and interscorer agreement

Learning in the classical planarian conditioning assay has historically been measured as the number of full-body contractions that occur in response to the onset of the conditioned light stimulus over the course of training, scored by human observers. In line with this quantification method, we had between one and three scorers manually score each video for conditioned responses occurring after the onset of the conditioned stimulus and before the onset of the unconditioned stimulus. For historic experiments, scorers simply watched complete overhead footage of the experiment as it took place. Because the CS scoring period was only two seconds per trial, each scorer's response was binary (one contraction or no contraction). By comparing Scorer 1 and Scorer 2's binary data on a trial-by-trial and worm-by-worm basis, we were able to count the proportion of trials in which each scorer agreed positively (contraction and contraction), agreed negatively (no contraction and no contraction), and disagreed (contraction and no contraction, or vice versa). The proportion of agreement trials divided by the total number of trials is the interscorer agreement for this experiment. We additionally computed Cohen's κ as a measure of interscorer agreement corrected for chance agreement, using the standard two-rater formula $\kappa = (P_0 - P_e)/(1 - P_e)$, where P_0 is the observed proportion of agreement and P_e is the proportion of agreement expected by chance given the marginal response rates of each scorer.

Footage from the contemporary protocol experiments was scored differently using a custom-made computer interface called the Blind Scoring Interface (BSI, Fig. S5). This interface enabled minimally trained scorers to count turns and contractions for each trial of the experimental footage blindly. Three scorers counted conditioned responses across the same subset of 203 trials split between forward conditioning (144) and pseudoconditioning groups (59). Scorer 2 scored an additional thirteen videos (216 total), and Scorer 1 scored all footage of the contemporary paradigm with the interface. Video files representing one worm and one trial were given codenames composed of two random adjectives and one random noun; when a user scored a video, only this codename was seen, so no information about experimental group (forward or pseudoconditioning) was available to them, nor was there any indication of how much previous training the worm had undergone. The videos were created such that each trial containing light presentation was segmented into its own 27-second video spanning the onset of the light up to the frame before the onset of the shock stimulus. Timestamps of turns and contractions were recorded from each video using the blind interface by pressing a "Turn" or "Contraction" button on the frame which the behavior occurred. The video title, as well as the sum and list of event timestamps for turns and contractions, were saved to a CSV from the interface. Data for interscorer reliability and agreement measures were obtained by comparing the event timestamps of turn and contraction events between the three scorers.

Because each scorer scored the same set of videos and noted timestamps of turns and contractions, we were able to compare exactly when each turn and contraction took place. This was accomplished by comparing two scorers' timestamp lists of turns and contractions per video (one video is equivalent to one trial). For each trial, every possible pair of same-type timestamps was tabulated and the difference between those timestamps calculated. Pairs were sorted by increasing difference and pairs with the least difference were considered matches up to a threshold of ten frames (1 second) for turns and fifteen frames (1.5 seconds) for contractions. Any timestamps remaining when all pair combinations had been exhausted were considered orphans. This was done for same-type events, turns with turns and contractions with contractions, until each type had a list of orphan timestamps. Next, each orphan timestamp was compared to orphan timestamps of the opposite type and, if within an event type's difference threshold, paired with the opposite-type orphan. Thus, for two scorers' assessment of a given video, we were able to calculate the number of matching timestamps, the number of unmatched timestamps, and which of those unmatched timestamps could be matched with an opposite-type event. We performed this procedure for each video and for each scorer pair and used those measures to calculate precision, recall and F1 scores.

The F1 metric is the harmonic mean of precision and recall. When treating one scorer's timestamps as reality (Reference) and testing another scorer against it (Test), precision is the proportion of true positives recovered by the Test dataset out of all Test positives. Recall is the proportion of true Test positives relative to the entirety of the Reference

positives. In this context, precision is the proportion of a Test Scorer's events that can be matched to the Reference Scorer. Likewise, recall is the proportion of Reference events that the Test Scorer can match out of all of the Reference's events. These proportions describe how much the Test Scorer marks the same events as the Reference Scorer while also not marking events the Reference Scorer does not have. The formulae for precision, recall, and F1 scores are given. (TP=true positive, FP=false positives, FN=false negative.)

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

There are two possible variants of F1 score to calculate for scorer pairs. One option is to calculate an F1 score for each video and then average across all video-specific F1 scores to produce a scorer pair's final F1 score. However, this method risks inflating F1 scores because any video where scorers agree nothing happened (no timestamps given and therefore no possible disagreement) would raise the average F1 score. Instead, we summed the TP, FP, and FN values across all videos, then calculated precision/recall only once to produce a final, more conservative F1 score. The F1 metrics calculated this way are more reflective of agreement about the events of interest.

The measures of unanimous turns and contractions across all three scorers were calculated similarly. For each video, triplets of same-type events were produced and sorted by their range (maximum frame value minus minimum frame value). The range threshold for turns was 10 and 15 for contractions. For each video, some number of unanimous contractions was counted. Because of partial agreement (e.g., Scorer 1 and 2 agreed on a contraction but Scorer 3 disagreed), we used venn diagram notation to calculate the total number of distinct turn and contraction events across all scorers and videos. The sum of all unanimous turns and contractions per video was divided by this number to yield the unanimity rate.

The above methods described agreement about events occurring within the window of the conditioned stimulus, i.e., conditioned responses. To score agreement about contractions induced by shock (the unconditioned stimulus) between Scorers 1 and 2, 198 individual worm trials were sampled from the classical conditioning group of the footage from our attempted contemporary protocol replication. The trials were observed non-blindly for contractions at the onset of the shock stimulus. Blind scoring was deemed unnecessary for shock stimuli because training day or classical/pseudoconditioning status should have no bearing on unconditioned responses barring health issues or habituation, neither of which was observed. The proportion of the number of trials where Scorer 1 and Scorer 2 agreed divided by the number of all scored videos formed the shock agreement metric.

3.7 Computer vision pipeline

Because manual scoring was shown to have low interscorer agreement, we sought to develop an alternative method to assess planarian behavior that would be resistant to human bias. The pipeline we developed began by quantifying worm morphology across continuous experimental footage. We used 10 morphological/spatial features (area, perimeter, area-perimeter ratio, circularity, hull area, centroid coordinates, relative concavity area, and two principal components from the 100-dimensional vector of the perimeter shape) to characterize the pose of individual planarians at each frame. To extract these features we tailored the Segment-Anything computer vision pipeline (Meta, Inc) to produce a timeseries of binary masks of the worms from the same overhead experimental footage used to manually score the worms' behavior. First, each video was spatially cropped such that each video contained a single worm. Then, we labeled two points for the first frame of each video to indicate the worm, and two points to indicate the electrodes (static occlusions in the environment). We used the Segment-Anything pipeline to propagate these points through each video and obtain a timeseries of binary masks. We manually inspected each video, and if the mask extraction quality was suboptimal, split the video before the quality drop and labeled new starting points to propagate through the next video segment. This process was

repeated until the entire timeseries of masks was of sufficient quality. Next, we calculated the mask area (sum of all black pixels in the mask), perimeter (using the findContours function from the OpenCV python library), area-perimeter ratio, circularity ($((4 * \pi * A)/(P * P))$ where A is the area and P is the perimeter), hull area (using convexHull and contourArea from OpenCV python library), centroid coordinates (using moments from OpenCV), and relative concavity area ($((H - A)/H)$ where H is hull area). To complement these features, for each frame we extracted the first two principal components of a 100-dimensional vector of distances from the centroid of the mask to evenly-spaced points on the perimeter of the mask. This analysis was published in [44]. The areas timeseries was used to identify outlier timepoints (> 3.5 standard deviations), which were excluded from all feature vectors.

While it is theoretically possible to use the morphological feature information to create a criterion for what counts as a turn or contraction (e.g., a derivative above a certain value in the “Area” timeseries), this approach still relies on human judgment to determine what signature(s) would be suggestive of a conditioned response or merely behavioral noise. The problem inherent to manual scoring is therefore still present. Instead, we took an alternative approach where we quantified how worm behavior differed between conditioned-stimulus (CS-on) and intertrial (CS-off) periods by computing the Kullback–Leibler (KL) divergence between the morphological feature distributions [70]. We first extracted the session-specific windows of time where the conditioned stimulus was on (CS-on) and off (CS-off) and approximated each condition as a multivariate Gaussian with diagonal covariance. We estimated means and unbiased variances of each feature in each condition (ignoring NaNs and adding a small constant for numerical stability), and computed the closed-form 1D Gaussian KL divergence between the CS-on and CS-off Gaussians. We obtained the total divergence $KL(\text{CS-on} \parallel \text{CS-off})$ for a given worm during a training block by summing these per-feature KL values. Thus, individual worms’ divergence scores could be plotted to observe any potential change in divergence over the course of the experiment’s four days (eight training blocks).

This computer vision method that assesses learning by the divergence of morphological distributions relies on (1) continuous uninterrupted footage of behaving worms for mask creation, and (2) a sufficient number of frames for each condition (CS-on and CS-off/ITI) to reliably estimate the KL divergence from Gaussian parameters. As described above, the contemporary paradigm contained 30 seconds of CS exposure per trial and an ITI of 30 seconds, which provides ample and continuous footage for analysis. The historic reproduction, conversely, had three-second CS exposure periods with a terminal one-second shock. That provides mere two-second CS-only periods for analysis, which is too little data to rigorously compare to the 30-second ITI. The same issue exists for the variant paradigms. The computer vision method was therefore by necessity restricted to the contemporary paradigm’s footage.

3.8 Statistical tests

Historic replication: To test for acquisition of the conditioned response over training days, we fit a linear slope to each worm’s daily response rate and tested whether the mean slope differed from zero (one-sample t -test, two-sided). We confirmed this with Page’s L test for monotone increasing trends across ordered time points. Group differences (conditioning vs. pseudoconditioning) were assessed with a one-way ANOVA. All analyses used Python (SciPy, statsmodels).

Contemporary replication: To test for acquisition of the conditioned response from the hand-scored data, we computed a linear slope of each worm’s daily response count across training days and tested whether the mean slope was significantly greater than zero using a one-sample t -test (one-sided, $\alpha = 0.05$). We supplemented this with Page’s L test, and fit LMMs with Day (continuous), Condition (conditioning vs. pseudoconditioning), and their interaction as fixed effects and individual subject as a random intercept, estimated via maximum likelihood. The Day $\times$ Condition interaction served as the primary test of the conditioning hypothesis, assessing whether the rate of behavioral change over training days differed between groups. LMMs were fit separately for turns and contractions. To compare each subpopulation’s mean per-worm slope to the historic reference, we used Welch’s t -tests with denominator degrees of freedom adjusted to propagate the standard error of Jacobson’s published regression slope.

Contemporary protocol scoring with the computer vision pipeline: To test for changes in KL divergence values between conditioning and pseudoconditioning groups over training, we log-transformed the KL divergence values to improve normality and analyzed with a LMM predicting log KL from Day, Condition, and their interaction, with random intercepts for individual subjects. The Day $\times$ Condition interaction term served as the primary test of differential learning trajectories; the main effect of Condition was additionally examined to assess overall group differences in KL divergence level. Per-subject slopes were also compared between conditions using Welch's *t*-test.

Miscellaneous conditioning paradigms: For each paradigm (UV-shock and white light–UV), we computed a linear slope of each worm's contraction count across training blocks and compared each population's mean per-worm slope to the historic Jacobson reference of +1.40 responses/block using two-sided Welch's *t*-tests with denominator degrees of freedom adjusted to propagate the standard error of Jacobson's published regression slope, and computed one-sample Cohen's *d* against the fixed Jacobson reference value. Worms were pooled across cohorts within each paradigm; slopes from cohorts contributing different numbers of blocks were weighted equally at the worm level.

Individual sensitivity: To test if different sensitivity to the conditioned stimulus affected acquisition of the conditioned response, we computed each worm's linear slope across conditioning blocks and tested whether the mean slope differed from zero (one-sample *t*-test, two-sided), separately for the conditioning and pseudoconditioning groups. To confirm that the high and low responder groups were meaningfully separated at baseline, we compared their pre-training CS response rates using an independent-samples *t*-test and computed Cohen's *d* as a measure of effect size. Slopes were compared between high and low responders within each condition using Welch's *t*-test with Cohen's *d*. To assess whether response rates changed across the light-only assessment sessions, we compared per-worm Day 1 and Day 2 means using a paired *t*-test (two-sided) with Cohen's *d_z*. To distinguish a gradual within-session trend from a between-session shift, we additionally fit per-worm linear slopes across individual trials using a one-sample *t*-test and computed Spearman's rank correlation between trial number and response rate.

Supporting information

S1 Fig. Conditioned-response acquisition from contemporary conditioning protocol, by species and group. A.

Mean per-trial conditioned response score per worm per day, for turns (solid) and contractions (dashed), broken down by species and strain (data from Fig. 2B). Error bars indicate 95% bootstrap CIs on the daily mean. One PC reference group (stock *G. dorotocephala*) shown for context. B. Forest plot of population-mean per-worm OLS slopes ($\pm$ 95% CI) for three CC populations and two references. Slopes pool turns and contractions. Dark-gray diamond: Jacobson's historic CC reference slope of $+1.40 \pm 0.40$ SE responses/day, derived by OLS on five published block means.

(TIF)

S2 Fig. Pre-conditioning baseline characterization and per-block conditioned response acquisition for historical protocol cohort. A.

Mean conditioned responses per block, by baseline-responsiveness stratum and condition ($n=6$ per line; error bars: 95% bootstrap CI). Note that the CC and PC groups did not differ in overall response level, and slopes did not differ between strata within either condition. B. Per-worm mean CS response rate across two baseline days of exposure to the CS only ($n=12$ per stratum). Box edges: 25th–75th percentiles; line: median; whiskers: $1.5 \times$ IQR. High and low strata were significantly separated at baseline ($t(22) = 5.58$, $p < .001$, Cohen's $d \approx 2.28$). C. Per-worm mean response rate across the two-day light habituation test ($N=36$; grey lines: individuals; red line: grand mean $\pm$ SEM). No within-session trend was detected; response rate increased between sessions, contrary to habituation ($+1.5 \times 10^{-4} \pm 9.3 \times 10^{-4}$ resp./trial, $t(35) = +0.16$, $p = .873$; Spearman $\rho = +0.002$, $p = .95$).

(TIF)

S3 Fig. Reprinted original plot from Jacobson et al. 1967.

(TIF)

S4 Fig. Three independent blind scorers all fail to detect an increase in conditioned responses during the contemporary conditioning protocol. A. Across-worm mean turn count on Day 1 and Day 4 for each scorer (CC: solid lines, 72 worm-sessions per scorer $\times$ day, with Scorer 3 at $n=65-68$ due to missing sessions); PC: dashed lines, $n=36$ worm-sessions per scorer $\times$ day, with Scorer 3 at $n=35$). Error bars indicate 95% bootstrap CIs. No scorer shows an increase from Day 1 to Day 4 in either group. B. Same as A, for contraction responses. (TIF)

S5 Fig. Blind Scoring Interface. Graphical user interface used to score worm behavior from the contemporary protocol. Each trial was cropped to a 27 s video spanning CS onset to just before UCS onset and assigned a random codename. Scorers marked turn and contraction events without knowledge of the worm's experimental group or training history. (TIF)

Acknowledgments

We extend our gratitude to the laboratories of Peter Reddien and Michael Levin for supplying worms and providing technical guidance, and to the Harvard NeuroTechnology Core for aid in building apparatuses. We additionally thank Reeva and Daniel Kimble for discussions of their experiences training planarians in the laboratory of James McConnell in the 1950s and 1960s, and for generously providing us with access to the complete *Worm Runner's Digest*.

Author contributions

Conceptualization: Zachary S. Kelso, Madeleine C. Snyder, Samuel J. Gershman.

Data curation: Zachary S. Kelso, Madeleine C. Snyder.

Formal analysis: Zachary S. Kelso, Madeleine C. Snyder.

Funding acquisition: Samuel J. Gershman.

Investigation: Zachary S. Kelso, Madeleine C. Snyder.

Methodology: Zachary S. Kelso, Madeleine C. Snyder.

Project administration: Samuel J. Gershman.

Resources: Samuel J. Gershman.

Software: Madeleine C. Snyder.

Supervision: Samuel J. Gershman.

Validation: Zachary S. Kelso, Madeleine C. Snyder.

Visualization: Zachary S. Kelso, Madeleine C. Snyder.

Writing – original draft: Zachary S. Kelso, Madeleine C. Snyder.

Writing – review & editing: Zachary S. Kelso, Madeleine C. Snyder, Samuel J. Gershman.

References

1. Bailly X, Reichert H, Hartenstein V. The urbilaterian brain revisited: novel insights into old questions from new flatworm clades. *Dev Genes Evol.* 2013;223(3):149–57. <https://doi.org/10.1007/s00427-012-0423-7> PMID: [23143292](#)
2. Holland LZ, Carvalho JE, Escriva H, Laudet V, Schubert M, Shimeld SM, et al. Evolution of bilaterian central nervous systems: a single origin? *EvoDevo.* 2013;4(1):27. <https://doi.org/10.1186/2041-9139-4-27> PMID: [24098981](#)
3. Holland L. Invertebrate origins of vertebrate nervous systems. In: *Evolutionary Neuroscience*. Elsevier; 2020. p. 51–73.
4. Sarnat HB, Netsky MG. When does a ganglion become a brain? Evolutionary origin of the central nervous system. *Semin Pediatr Neurol.* 2002;9(4):240–53. <https://doi.org/10.1053/spen.2002.32502> PMID: [12523550](#)

5. Corning W, Riccio D. The planarian controversy. *Molecular approaches to learning and memory*. 1970. p. 107–50.
6. Corning W, Kelly S. Platyhelminthes: the turbellarians. In: *Invertebrate Learning: Volume 1 Protozoans Through Annelids*. Springer; 1973. p. 171–224.
7. Thompson R, McConnell J. Classical conditioning in the planarian, *Dugesia dorotocephala*. *J Comp Physiol Psychol*. 1955;48(1):65–8. <https://doi.org/10.1037/h0041147> PMID: 14354075
8. McConnell JV, Jacobson AL, Kimble DP. The effects of regeneration upon retention of a conditioned response in the planarian. *J Comp Physiol Psychol*. 1959;52(1):1–5. <https://doi.org/10.1037/h0048028> PMID: 13641455
9. McConnell JV, Jacobson R, Humphries B. The effects of ingestion of conditioned planaria on the response level of naive planaria: a pilot study (or: “You are what you eat?”). *J Neuropsychiat*. 1961;3.
10. McConnell JV. Memory transfer through cannibalism in planarians. *J Neuropsychiat*. 1962;3:542–8.
11. Zelman A, Kabat L, Jacobson R, McConnell JV. Transfer of training through injection of “conditioned” RNA into untrained planarians. *Worm Runner's Digest*. 1963;5:14–21.
12. Jacobson AL. Classical conditioning and the planarian. In: *Chemistry of Learning: Invertebrate Research*. Springer; 1967. p. 195–216.
13. Jacobson AL, Fried C, Horowitz SD. Planarians and Memory: I. Transfer of Learning by Injection of Ribonucleic Acid. *Nature*. 1966;209(5023):599–601. <https://doi.org/10.1038/209599a0>
14. Jacobson AL, Fried C, Horowitz SD. Planarians and memory: II. The Influence of Prior Extinction on the Ribonucleic Acid Transfer Effect. *Nature*. 1966;209(5023):599–601.
15. Jacobson AL, Horowitz SD, Fried C. Classical conditioning, pseudoconditioning, or sensitization in the planarian. *J Comp Physiol Psychol*. 1967;64(1):73–9. <https://doi.org/10.1037/h0024808> PMID: 6064436
16. Bennett E, Calvin M. Failure to train planarians reliably. *Neurosci Res Program Bullet*. 1964;2(4):3–24.
17. McConnell JV. Failure to interpret planarian data correctly: a reply to Bennet and Calvin. What goes here? 1965.
18. Hartry AL. *Learning and Memory Transfer in Planarians*. The Claremont Graduate University; 1970.
19. Gaito J. Molecular psychobiology of memory: Its appearance, contributions, and decline. *Psychobiology*. 1976;4(4):476–84. <https://doi.org/10.3758/bf03326600>
20. Newmark PA, Sánchez Alvarado A. Schmidtea happens: Re-establishing the planarian as a model for studying the mechanisms of regeneration. *Curr Top Dev Biol*. 2022;147:307–44. <https://doi.org/10.1016/bs.ctdb.2022.01.002> PMID: 35337453
21. Peretti P, Zinthefer M, Fernandez H. Influence of Learning Trials on Memory Retention of Planaria: *Dugesia tigrina*. *Pakistan Journal of Psychology*. 1999;30(1):35.
22. Abbott SM, Wong GK. The conditioning and memory retention of planaria (*Dugesia tigrina*) for directional preferences. *BIOS*. 2008;79(4):160–70. <https://doi.org/10.1893/0005-3155-79.4.160>
23. Prados J, Alvarez B, Howarth J, Stewart K, Gibson CL, Hutchinson CV, et al. Cue competition effects in the planarian. *Anim Cogn*. 2013;16(2):177–86. <https://doi.org/10.1007/s10071-012-0561-3> PMID: 22976576
24. Samuel K, Hakes AK, Suviseshamuthu ES, Fichera ME. Retention and Distribution of Dopamine-Dependent Reward Memory in Regenerating Planaria. *Biomolecules*. 2026;16(5):649. <https://doi.org/10.3390/biom16050649> PMID: 42194000
25. Shimojo K, Shimojo E, Shimojo S. New evidence for preservation of conditioned behavior based on UV light sensitivity in dissected tail halves of planarians. *J Vis*. 2021;21(9):2326. <https://doi.org/10.1167/jov.21.9.2326>
26. James FLJ, Ahamed SK, Yang C, Thinley J, Vicario CM, Canales JJ, et al. Adapting rodent cued threat conditioning to planarians: Memory acquisition, consolidation, and reconsolidation. *Behav Processes*. 2026;235:105321. <https://doi.org/10.1016/j.beproc.2025.105321> PMID: 41390058
27. Forde F. *Probing the Limits of Memory: Can a Learned Response Persist Through Decapitation and Regeneration in Planaria?* [Ph.D. Thesis]. Wellington, New Zealand: Open Access Te Herenga Waka-Victoria University of Wellington; 2025.
28. Elliott SA, Alvarado AS. Planarians and the history of animal regeneration: paradigm shifts and key concepts in biology. *Planarian regeneration: Methods and protocols*. 2018. p. 207–39.
29. Beane WS, Morokuma J, Lemire JM, Levin M. Bioelectric signaling regulates head and organ size during planarian regeneration. *Development*. 2013;140(2):313–22. <https://doi.org/10.1242/dev.086900> PMID: 23250205
30. Lobo D, Beane WS, Levin M. Modeling planarian regeneration: a primer for reverse-engineering the worm. *PLoS Comput Biol*. 2012;8(4):e1002481. <https://doi.org/10.1371/journal.pcbi.1002481> PMID: 22570595
31. Reddien PW, Sánchez Alvarado A. Fundamentals of planarian regeneration. *Annu Rev Cell Dev Biol*. 2004;20:725–57. <https://doi.org/10.1146/annurev.cellbio.20.010403.095114> PMID: 15473858
32. Reddien PW. The Cellular and Molecular Basis for Planarian Regeneration. *Cell*. 2018;175(2):327–45. <https://doi.org/10.1016/j.cell.2018.09.021> PMID: 30290140
33. Kusayama T, Watanabe S. Reinforcing effects of methamphetamine in planarians. *Neuroreport*. 2000;11(11):2511–3. <https://doi.org/10.1097/00001756-200008030-00033> PMID: 10943713

34. Buttarelli FR, Pellicano C, Pontieri FE. Neuropharmacology and behavior in planarians: translations to mammals. *Comp Biochem Physiol Part C: Toxicol Pharmacol*. 2008;147(4):399–408.
35. Cho M, Nayak SU, Jennings T, Tallarida CS, Rawls SM. Predator odor produces anxiety-like behavioral phenotype in planarians that is counteracted by fluoxetine. *Physiol Behav*. 2019;206:181–4. <https://doi.org/10.1016/j.physbeh.2019.04.003> PMID: 30951749
36. Gershman SJ. The molecular memory code and synaptic plasticity: A synthesis. *Biosystems*. 2023;224:104825. <https://doi.org/10.1016/j.biosystems.2022.104825> PMID: 36610586
37. Abraham WC, Jones OD, Glanzman DL. Is plasticity of synapses the mechanism of long-term memory storage? *NPJ Sci Learn*. 2019;4(1):9. Available from: <https://www.nature.com/articles/s41539-019-0048-y>
38. Langille JJ, Gallistel CR. Locating the engram: Should we look for plastic synapses or information-storing molecules? *Neurobiol Learn Memory*. 2020;169:107164. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1074742720300083>
39. Gallistel CR. The Coding Question. *Trends Cogn Sci*. 2017;21(7):498–508. <https://doi.org/10.1016/j.tics.2017.04.012> PMID: 28522379
40. Gallistel CR. The physical basis of memory. *Cognition*. 2021;213:104533. <https://doi.org/10.1016/j.cognition.2020.104533> PMID: 33375954
41. Akhlaghpour H. An RNA-based theory of natural universal computation. *J Theor Biol*. 2022;537:110984. <https://doi.org/10.1016/j.jtbi.2021.110984> PMID: 34979104
42. Shomrat T, Levin M. An automated training paradigm reveals long-term memory in planarians and its persistence through head regeneration. *J Exp Biol*. 2013;216(Pt 20):3799–810. <https://doi.org/10.1242/jeb.087809> PMID: 23821717
43. Deochand N, Costello MS, Deochand NE. Behavioral Research with Planaria. *Perspect Behav Sci*. 2018;41(2):447–64. <https://doi.org/10.1007/s40614-018-00176-w> PMID: 31976405
44. Bray SR, Wyss LS, Chai C, Lozada ME, Wang B. Adaptive robustness through incoherent signaling mechanisms in a regenerative brain. *Cell Rep*. 2024;43(8):114580. <https://doi.org/10.1016/j.celrep.2024.114580> PMID: 39133614
45. Shettigar N, Chakravarthy A, Umashankar S, Lakshmanan V, Palakodeti D, Gulyani A. Discovery of a body-wide photosensory array that matures in an adult-like animal and mediates eye-brain-independent movement and arousal. *Proc Natl Acad Sci U S A*. 2021;118(20):e2021426118. <https://doi.org/10.1073/pnas.2021426118> PMID: 33941643
46. Shettigar N, Joshi A, Dalmeida R, Gopalkrishna R, Chakravarthy A, Patnaik S, et al. Hierarchies in light sensing and dynamic interactions between ocular and extraocular sensory networks in a flatworm. *Sci Adv*. 2017;3(7):e1603025. <https://doi.org/10.1126/sciadv.1603025> PMID: 28782018
47. McConnell JV. On the turning of worms: A reply to James and Halas. *Psychol Rec*. 1964;14(1):13–20. <https://doi.org/10.1007/bf03393556>
48. Bray S, Wyss L, Wang B. Planameterization: Software to segment and parameterize videos of planarian behavior. GitHub repository. 2023. <https://github.com/samuelbray32/planameterization>
49. Harris JA, Gallistel CR. Information, certainty, and learning. *Elife*. 2026;13:RP102155. <https://doi.org/10.7554/eLife.102155> PMID: 41626953
50. Le D, Sabry Z, Chandra A, Kristan WB III, Collins EMS, Kristan WB Jr. Planarian fragments behave as whole animals. *Curr Biol*. 2021;31(22):5111–7. <https://doi.org/10.1016/j.cub.2021.09.056>
51. Wyss LS, Bray SR, Wang B. Cellular diversity and developmental hierarchy in the planarian nervous system. *Curr Opin Genet Dev*. 2022;76:101960. <https://doi.org/10.1016/j.gde.2022.101960> PMID: 35878572
52. McConnell JV. Specific Factors Influencing Planarian Behavior. In: Corning WC, Ratner SC, editors. *Chemistry of Learning*. Boston, MA: Springer US; 1967. p. 217–33.
53. Lázaro EM, Harrath AH, Stocchino GA, Pala M, Baguña J, Riutort M. *Schmidtea mediterranea* phylogeography: an old species surviving on a few Mediterranean islands? *BMC evolutionary biology*. 2011;11(1):274.
54. Sluys R, Riutort M. Planarian Diversity and Phylogeny. *Methods Mol Biol*. 2018;1774:1–56. https://doi.org/10.1007/978-1-4939-7802-1_1 PMID: 29916154
55. Best JB, Rubinstein I. Maze learning and associated behavior in planaria. *J Comp Physiol Psychol*. 1962;55:560–6. <https://doi.org/10.1037/h0044394> PMID: 13868584
56. Best JB. Protopsycho. *Sci Am*. 1963;208(2):54–62. Available from: <https://www.scientificamerican.com/article/protopsycho>
57. Best JB. Behavior of Planaria in Instrumental Learning Paradigms. NASA-CR-80380. Accession No. 67N13007. National Aeronautics and Space Administration; 1963. Available from: <https://ntrs.nasa.gov/citations/19670003678>
58. Ardiel EL, Rankin CH. An elegant mind: learning and memory in *Caenorhabditis elegans*. *Learn Mem*. 2010;17(4):191–201. <https://doi.org/10.1101/Im.960510> PMID: 20335372
59. Moore RS, Kaletsky R, Lesnik C, Cota V, Blackman E, Parsons LR, et al. The role of the Cer1 transposon in horizontal transfer of transgenerational memory. *Cell*. 2021;184(18):4697–712.e18. <https://doi.org/10.1016/j.cell.2021.07.022> PMID: 34363756
60. Nishijima S, Maruyama IN. Appetitive Olfactory Learning and Long-Term Associative Memory in *Caenorhabditis elegans*. *Front Behav Neurosci*. 2017;11:80. <https://doi.org/10.3389/fnbeh.2017.00080> PMID: 28507513
61. Brown DDR, Pearson BJ. A Brain Unfixed: Unlimited Neurogenesis and Regeneration of the Adult Planarian Nervous System. *Front Neurosci*. 2017;11:289. <https://doi.org/10.3389/fnins.2017.00289> PMID: 28588444

62. McConnell JV. On the procuring and the care of planarians. Manual of Psychological Experimentation on Planarians, second ed Planarian Press, Ann Arbor, Michigan. 1967. 128 p.
63. McConnell JV, Cornwell PR, Clay M. An apparatus for conditioning Planaria. Am J Psychol. 1960;73:618–22. <https://doi.org/10.2307/1419955> PMID: [13773818](https://pubmed.ncbi.nlm.nih.gov/13773818/)
64. Barnes CD, Katzung BG. Stimulus polarity and conditioning in planaria. Science. 1963;141(3582):728–30. <https://doi.org/10.1126/science.141.3582.728> PMID: [13969357](https://pubmed.ncbi.nlm.nih.gov/13969357/)
65. Lubow RE, Moore AU. Latent inhibition: the effect of nonreinforced pre-exposure to the conditional stimulus. J Comp Physiol Psychol. 1959;52:415–9. <https://doi.org/10.1037/h0046700> PMID: [14418647](https://pubmed.ncbi.nlm.nih.gov/14418647/)
66. McConnell JV, Mpitsos G. Effects of presence or absence of slime on classical conditioning in planarians. In: American Zoologist. vol. 5. Soc Integrative Comparative Biology; 1965. p. 653+.
67. Seligman ME. On the generality of the laws of learning. Psychol Rev. 1970;77(5):406–18. <https://doi.org/10.1037/h0029790>
68. Domjan M, Cusato B, Krause M. Learning with arbitrary versus ecological conditioned stimuli: evidence from sexual conditioning. Psychon Bull Rev. 2004;11(2):232–46. <https://doi.org/10.3758/bf03196565> PMID: [15260188](https://pubmed.ncbi.nlm.nih.gov/15260188/)
69. Reho G, Lelièvre V, Cadiou H. Planarian nociception: Lessons from a scrunching flatworm. Front Mol Neurosci. 2022;15:935918. <https://doi.org/10.3389/fnmol.2022.935918> PMID: [35959107](https://pubmed.ncbi.nlm.nih.gov/35959107/)
70. Gallistel CR, Latham PE. Bringing Bayes and Shannon to the Study of Behavioural and Neurobiological Timing and Associative Learning. Timing Time Percept. 2022;11(1–4):29–89. <https://doi.org/10.1163/22134468-bja10069>