PROBE HORIZON IS NOT PREDICTABLE MEMORY: MEASURING THE HISTORY A SURROGATE NEEDS ON A COMPLETE CONNECTOME MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

A surrogate model must be told how much of the past to see, and practice takes that number from a probe of the simulator, typically an impulse horizon, an autocorrelation time or a settling time read at a threshold such as $1/e$. Such a probe reports how long one chosen excitation leaves a visible trace in one observed statistic. What a surrogate needs is different: the horizon over which past input still changes the noise-averaged response it is trained to predict. On a leaky integrate-and-fire model of the first complete connectome of an adult male *Drosophila* central nervous system, the three probes return 30, under 10 and 150 ms, while a windowed MLP student keeps gaining fidelity out to 960 ms of history and the memoryless MLP trails a GRU by 0.15 in fidelity. We quantify what the student needs with a shared-suffix measurement whose estimator subtracts the teacher's own stochastic floor in squared distance. On this connectome the measurement returns a tail that is still above a tenth of its initial amplitude at 1 s, the last lag we measure, matching the range of history the student keeps using and well beyond every probe. On a second complete brain connectome under identical neuron dynamics the same measurement returns 83 ms, and its students need 80 ms. On a synthetic teacher with a planted 500 ms state it returns a fitted slow timescale of 521 ms [429, 693]. The student's required history survives a sixteenfold larger training bank, a change of input noise and a 9.6-fold change in capacity, and wherever the input contains transitions it moves with the predictable tail across teachers and interventions. A gain-matched intervention makes the separation causal. Read after subtracting its stochastic floor, the teacher's step response lengthens from 220 ms to past the end of the step-response window, while the conventional reading shortens. The history a surrogate needs is a measurement of predictable memory, and predictable memory is not the horizon a probe reports.

1 INTRODUCTION

Surrogate modelling and teacher–student distillation share a design question: how much of the past must the student see? Existing practice answers it with a probe of the system, applied before any model is trained. Delay embeddings set their lag from the $1/e$ time of the autocorrelation or the first minimum of the mutual information, which with the embedding dimension fixes how far back they reach (Fraser & Swinney, 1986; Kantz & Schreiber, 2004; Tan et al., 2023). Step-response models in predictive control are truncated at the settling time of the plant (García et al., 1989; Shridhar & Cooper, 1997). Memory closures for reduced-order models are truncated where the memory kernel has decayed, and when the closure is a recurrent network, that estimate brackets the input lengths tried (Tian et al., 2021; Wang et al., 2020). Intrinsic timescales of neural populations are read from autocorrelation (Murray et al., 2014). A learned surrogate inherits this practice because such probes are what is available before training.

We introduce two terms for surrogate modelling that separate what a probe reports from what a surrogate requires. The *probe horizon* is the duration for which a chosen excitation leaves a trace above threshold in a chosen statistic of the response. The *predictable memory* is the duration for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

which past input still changes the noise-averaged response under the input distribution the surrogate is run on. A probe horizon can fall short of the predictable memory of the same system for two reasons. First, a brief excitation may move a slow component so little that its trace stays below the detection threshold. For instance, A 10 ms pulse shifts a state with a 500 ms time constant by only 2%. Second, a statistic of a noisy response decays not to zero but to a stochastic floor set by the noise that survives averaging. A threshold on it therefore marks where the predictable signal sinks into the floor rather than where dependence on past input ends. Two things are missing before predictable memory can be established on a simulator. One is the magnitude of the shortfall on a scientific simulator of realistic scale. The other is a measurement of predictable memory that requires no prior specification of the timescale it is to detect.

We propose the shared-suffix measurement, a direct measurement of predictable memory by intervention. It constructs pairs of input episodes that share their last k ms and are independent before that. Each episode is run many times and the two noise-averaged outputs are compared. Any difference that remains is due to input older than k ms. The difference of two finite-sample averages carries stochastic power of its own. We therefore work in squared distance where that power is additive and subtract a floor estimated from two averages of the same episode. The result is a curve read at descriptive quantiles after the fact rather than stopped at a threshold during acquisition, so no timescale has to be specified in advance.

The main teacher of our experiments is MaleCNS, the first complete connectome of an adult male fly central nervous system (Berg et al., 2026), simulated as a leaky integrate-and-fire network with published parameters (Shiu et al., 2024). Its probe horizons are 30 ms by impulse, under 10 ms by autocorrelation and 150 ms by settling after a step. To find the history a surrogate actually requires, we train students on the teacher's responses with input windows of increasing length. If the probes were right, fidelity would stop improving at 150 ms. Yet a windowed MLP student keeps improving out to 960 ms of history, and the memoryless MLP reaches a fidelity of 0.757 against 0.903 for a GRU. Applied to the teacher alone, the shared-suffix measurement returns a closely matching result. More than a tenth of the predictable dependence survives to 1 s, the last lag we measure, and the fitted slow timescale is 824 ms. On FlyWire, a second complete brain connectome simulated under identical dynamics (Dorkenwald et al., 2024; Schlegel et al., 2024), predictable dependence falls to a tenth of its initial amplitude within 83 ms, and students need 80 ms.

Contributions. (1) **A surrogate's required history is an operational measurement of predictable memory.** It is not explained by memorisation of the training bank, by averaging of input noise or by student capacity: a sixteenfold larger bank, a change of input noise and a 9.6-fold change in parameter count all leave it at the top of the measured grid (§3, §6.2). (2) **A measurement of predictable memory that does not have to be told which timescale to look for.** The shared-suffix measurement intervenes on the surrogate's own stimulus family and is read after subtracting the teacher's stochastic floor in power rather than in distance. On a teacher with a planted 500 ms state it returns $\tau_2 = 521$ ms $[429, 693]$, while the impulse probe returns 10 ms (§4, §5.1). (3) **Probe horizon and predictable memory are causally separable.** Two complete connectomes under identical dynamics differ by more than an order of magnitude in required history, with predictable tails whose intervals do not overlap, and a gain-matched intervention on one raises the predictable power left at 1 s ninefold while all three conventional probes report the same or a shorter horizon (§5.2, §6.1).

2 Problem setup

A teacher is a stochastic simulator that turns an input history into a response, and a surrogate is a student trained on episodes from a fixed stimulus family to predict that response. Because the teacher is stochastic, what the student can learn is the noise-averaged response, and the student must be told how much of the past to see: a windowed student receives only the last c bins of input. We ask two questions. How large must c be, and can that number be read from a probe of the teacher before any student is trained? This section introduces the setup of the teachers, the students and the score. §3 answers the first question and §4 answers the second.

Teachers. MaleCNS v1.0 (Berg et al., 2026) contains 166,700 annotated neurons, 166,483 of which participate in a 25.6M-edge graph covering the brain and ventral nerve cord (VNC) of an adult male. FlyWire v783 (Dorkenwald et al., 2024; Schlegel et al., 2024) is an adult female brain

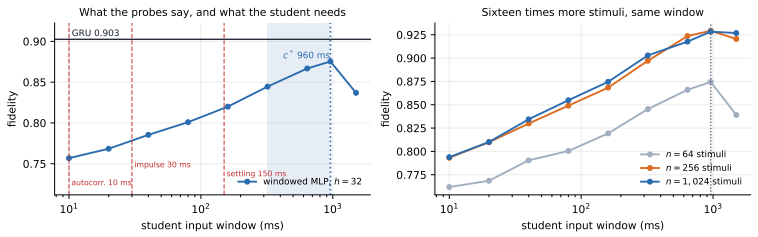

Figure 1: Left: windowed-MLP fidelity against the length of its input window, with the three conventional probes marked and the GRU's best fidelity as a horizontal line. The student is still improving at 96 bins. (Last point: whole-episode window, excluded from c^*.) Right: the same sweep at three training-bank sizes. A sixteenfold larger bank raises the curve without reducing the long-window benefit.

ending at the neck connective; of its 139,255 proofread neurons we simulate the 138,639 on the list of Shiu et al. (2024). The two are independent reconstructions of different individuals. Both are simulated as leaky integrate-and-fire networks with the parameters of Shiu et al. (2024), identical across connectomes ($\tau_m = 20$ ms, $\tau_s = 5$ ms, $w_{syn} = 0.275$ mV; Appendix A). Weights scale with synapse count and signs follow neurotransmitter annotation. Sensory drive is a deterministic rate profile per input population, delivered as Poisson spike trains inside the model. Given the spike trains, the model is bit-deterministic. Both teachers reproduce the sugar-feeding result of Shiu et al. (2024). Sugar drive reaches the feeding motor neuron MN9 at 112.7 Hz on FlyWire and 39.0 Hz on MaleCNS, and size-matched control sets, degree-preserving rewiring and, on FlyWire, bitter co-activation leave at most a few Hz (MaleCNS has no bitter annotation; Appendix B).

Students. The student target is the firing rate of every readout cell type in 10 ms bins: 486 descending and 186 motor types on MaleCNS, 672 channels in all, and 472 descending types on FlyWire, which has no motor readout. Stimuli drive the wind-sensing afferents of Johnston's organ, split by hemisphere into two groups, as rate profiles with temporal structure (steps, ramps, pulse trains) whose constant segments have a median length of 390 ms. The training set is 64 stimulus parameterisations $\times$ 3 Poisson seeds (192 episodes of 1.5 s), with 36 validation and 60 test episodes split by stimulus and seed; we call this stimulus set the main family. Students receive the deterministic rate profile that drove the teacher. A 3 s family from the same generator is used where stated. Students map the rate time series to the channel rates. The recurrent student is a single-layer GRU (Cho et al., 2014) of hidden dimension h. The windowed MLP has the same width and takes the last c bins as input; at $c = 1$ it is the memoryless MLP. Unless stated otherwise, each configuration is trained to convergence with three seeds, on 150-bin chunks that are whole episodes on the 1.5 s families. Optimiser, schedule and depth are in Appendix C.

Scoring. Write a trial as $Y = f(s) + \varepsilon$, with independent noise of variance σ^2. Two teacher trials are then expected $2\sigma^2$ apart in squared distance, while a perfect student incurs σ^2, so the attainable floor is half the teacher–teacher distance, $D_{TT}/2$. Normalised fidelity is 1 at this floor and 0 at the loss of the training-mean predictor, and is reported unclipped, so a fidelity gain is a gain in predicting $f(s)$. The same anchoring is standard for scoring models of noisy neural responses (Hsu et al., 2004; Schoppe et al., 2016). c^* is the smallest window c at which the windowed MLP's mean fidelity is within one pooled seed standard deviation of its maximum over $c \in \{1, 2, 4, 8, 16, 32, 64, 96\}$ bins and stays there, with a bootstrap 95% interval over seeds and test stimuli. Where the curve is still rising at the top of the grid, c^* is written $\geq$ 960 ms rather than extrapolated. On the 3 s family the grid extends to 200 bins, where the curve peaks and turns rather than running into the top of the grid (§6.2).

3 Are the conventional probes good enough?

Figure 1 exposes the mismatch. A 10 ms rate perturbation leaves a Holm-significant trace on the descending response for 30 ms, the autocorrelation of that response falls to $1/e$ within one 10 ms bin, and the response settles to $1/e$ of a step within 150 ms, so the three probes put the teacher's

memory under 150 ms (protocols in Appendix E). Yet the windowed MLP's fidelity rises from 0.757 without memory and 0.801 with an 80 ms window to 0.845 at 320 ms, more than twice the longest probe, and 0.875 at 960 ms, against 0.903 for the best GRU, at $h = 64$. Given 96 bins of input, the MLP closes most of its gap to the GRU, so the shortfall is one of history rather than architecture. The conventional probes stop well before additional history stops helping the student. The probes themselves, and the practices that set a model's history from them, are surveyed in §7.

Either the probes miss memory in the teacher, or the student uses a long window for something other than that memory. We test the second explanation first. Memorisation, input denoising and limited capacity would each give the student such a use, and each predicts that the required history should shrink under a corresponding intervention. A fourth possibility, that single trajectories carry long-range structure a window could exploit, is tested by splicing.

Memorisation of the training bank. Memorisation would turn a long window into a fingerprint reader, letting the student identify which of the 64 training stimuli it is in and look up the answer. A bank too large to memorise should then shorten the required history. We regenerated the family with 256 and 1,024 unique stimuli and repeated the sweep at $h = 32$ (nearest-neighbour cosine distance falls from 0.025 to 0.022, so the bank grows in coverage and not in duplicates). For every bank, c^* reaches the top of the grid, and the gain from a 10 ms to a 960 ms window goes from $+0.113$ at $n = 64$ to $+0.135$ at $n = 1{,}024$ (Table 5). The long window is therefore not explained by memorisation of a small training bank.

Averaging of input noise. A student receiving a noisy input would integrate over time to estimate the underlying rate, and a cleaner input should then shorten its required history. Our students receive the deterministic rate profile, and the Poisson sampling happens inside the teacher, where each drive group pools 324–348 afferents and the realised drive deviates from its rate by 2.7%. Giving the student a single Poisson realisation instead, or the mean of eight, leaves c^* at the top of the grid in all three conditions, and the fidelity curves differ by at most 0.005 at every window.

Capacity of the student. A 192-dimensional input compressed into $h = 32$ units could make a long window a way of buying representational room rather than a statement about the teacher, in which case a wider student should need less history. We repeated the MaleCNS sweep at $h \in \{16, 32, 64, 128\}$, a 9.6-fold range in parameter count. Every width keeps $c^* \geq 960$ ms, memoryless fidelity varies by 0.006 across the four, and the gain from a 960 ms window over a memoryless one lies between 0.105 and 0.127 at every width (Table 6).

Structure in single trajectories. The student receives the deterministic rate profile, so by construction it cannot exploit structure in the teacher's noise. Splicing single runs confirms it: beyond 30 ms, two runs sharing their recent input are as far apart in output as two runs of the same episode under different Poisson seeds.

The long window survives all three interventions, and single trajectories give it nothing else to fit. It is spent on the noise-averaged response $\mathbb{E}[Y \mid \text{input}]$, the object the $D_{TT}/2$ floor scores against. A measurement meant to explain the student therefore has to be made on that average. A raw-response probe reads a statistic that still contains this noise.

4 THE SHARED-SUFFIX MEASUREMENT

The shared-suffix measurement. We operationalise dependence on the distant past by intervention rather than by fitting a model of it. For $N = 32$ pairs of episodes whose rate profiles share the last k ms and are independent before that, we run each episode with $M = 16$ Poisson seeds, average, and compare the two averages at the final bin. If the response there depended only on the last k ms of input, the two would coincide. The measurement is matched to the student in two ways. It draws its episodes from the stimulus family the surrogate is trained and evaluated on, rather than from an excitation chosen by the experimenter, and it compares responses in the same unnormalised squared distance on raw rates that the fidelity uses, so each channel counts for as much as the variance it carries.

Predictable power, not distance. A seed average does not remove the teacher's noise, it only shrinks it. Write the two averages as $\bar{Y}_a = \mu_a + \epsilon_a$ and $\bar{Y}_b = \mu_b + \epsilon_b$, where μ is the noise-averaged response we want to compare and ϵ is what remains of the noise after averaging M seeds. Their distance is therefore never zero, even when $\mu_a = \mu_b$, and the question is how to remove the part that noise contributes. In distance the two parts do not separate, but in squared distance they do. Because ϵ_a and ϵ_b are independent and zero-mean, the cross term vanishes in expectation and

$$\mathbb{E}\big\|\bar{Y}_a - \bar{Y}_b\big\|_2^2 \;=\; \big\|\mu_a - \mu_b\big\|_2^2 \;+\; \mathbb{E}\big\|\epsilon_a - \epsilon_b\big\|_2^2. \tag{1}$$

The second term can be measured on its own. Two independent M-seed averages of the same episode share μ, so their squared distance, F^2, is noise alone. Writing $D^2(k) = \mathbb{E}_{\text{pairs}}\|\bar{Y}_a - \bar{Y}_b\|_2^2$ for pairs sharing their last k ms, the predictable power

$$P(k) = D^2(k) - F^2 \tag{2}$$

estimates how much input older than k still changes the noise-averaged response. Its normalisation $\rho(k) = P(k)/P(0)$, which starts at one, is the dependence curve. Subtracting the floor in distance instead would treat noise as additive where it is not, and shortens the tail (Appendix H).

Because the conventional probes are read on amplitude, we read the dependence curve on amplitude too. τ_{50}, τ_{90} and τ_{95} are the lags at which the amplitude of predictable dependence has fallen to 0.5, 0.1 and 0.05 of its initial value, where ρ crosses 0.25, 0.01 and 0.0025. We report τ_{90} as the predictable tail. We fit one and two timescales to the same amplitude, $\rho(k) = [ae^{-k/\tau_1} + (1 - a)e^{-k/\tau_2}]^2$ with $\tau_1 < \tau_2$, and call the fitted τ_2 the slow timescale. Intervals are bootstrapped over pairs. The estimator also carries its own check. At the largest k the two episodes are identical, so $P(k)$ must be zero, and it is in every condition, with values from -0.007 to $+0.008$. Because the identity point is zero by construction, it serves only as this check and enters neither the quantiles nor the fits. A quantile the curve has not reached by the last measured lag is reported as lying beyond it.

The estimator recovers a known curve. We test the estimator on a benchmark with a planted amplitude curve and a stochastic floor of the experiment's shape (Appendix H). Subtracting the floor in distance biases the tail downward, and the bias grows as the signal falls toward the floor. The distance-domain estimator returns $\tau_{90} = 325$ ms against a true 561 and a slow timescale of 329 ms against a true 479, whereas the power-domain estimator returns 572 and 500 (Table 9).

5 Experiments

5.1 A teacher with a planted timescale

Before applying the measurement to a connectome, we test it, together with the impulse probe and the student sweep, on a teacher whose memory we chose. Teacher A is a 200-unit tanh recurrent network (spectral radius 0.9, $\tau = 20$ ms), driven by the same rate profiles through a fixed random projection, with input and observation noise matched to the fly's. It carries one additional scalar state with a planted time constant of 500 ms, driven by a projection of the input and added to the output with a gain set so that it explains 30% of the predictable output variance. Teacher B is identical with that gain set to zero (Appendix D). The question is whether each of these three recovers a timescale known by design.

The impulse probe cannot tell the two teachers apart (Table 1). Their impulse curves coincide to two decimals and both horizons are 10 ms, although A contains a state that carries 30% of what the student must predict. The dependence curve separates them, and on A its slow timescale is 521 ms, with a bootstrap interval of $[429, 693]$ that contains the planted 500 ms. Nothing in the measurement was set from that value. The shared-suffix grid, the seed budget and the quantiles are the same on every teacher in this paper. The student's required history separates the teachers as well, 640 ms [640, 960] on A against 160 ms [80, 160] on B.

What the impulse probe would need. A 10 ms pulse steps a 500 ms integrator by 2%, and detecting the slow state of teacher A with it would take about 2×10^5 seeds per condition, against 4,608 episodes for the whole dependence curve. A 200 ms pulse finds the state at 384 seeds, which means choosing the pulse from the quantity being measured (Appendix G).

Table 1: Synthetic teachers with (A) and without (B) a planted 500 ms state, measured in the same way as the connectome teachers. The impulse probe does not distinguish them. The dependence curve and the required history do, and the fitted slow timescale matches the planted value without the measurement having been given it.

	A (500 ms state, 30% of variance)	B (no slow state)
Impulse horizon (Holm, 24 seeds)	10 ms	10 ms
Settling after a step, $1/e$	240 ms	30 ms
Predictable tail τ_{90}	922 ms [814, 999]	98 ms [96, 99]
Fitted slow timescale τ_2	521 ms [429, 693]	unidentified
Required history c^* ($h = 32$, 10 seeds)	640 ms [640, 960]	160 ms [80, 160]

Table 2: Predictable dependence of the noise-averaged response on both connectomes, under identical neuron dynamics. Bootstrap 95% intervals over 32 pairs; an open upper end (—) marks an interval that extends beyond the last measured lag. Middle column: MaleCNS scored on its 486 descending cell types alone. The other conditions are in Appendix Tables 10 and 11.

	MaleCNS		FlyWire
	672 channels	486 descending	472 descending
Predictable tail τ_{90} (ms)	>1,000 [583, —]	>1,000 [571, —]	83 [63, 99]
Fitted slow timescale τ_2	824 ms	788 ms	unidentified
Impulse horizon	30 ms	30 ms	20 ms
Memoryless MLP vs GRU	0.757 vs 0.903	—	0.952 vs 0.976
Required history c^*	$\geq$ 960 ms	$\geq$ 960 ms	80 ms [40, 80]

5.2 Two connectomes, one with a long predictable tail

MaleCNS. At 300 ms, twice the settling time and ten times the impulse horizon, 14% of the initial predictable power remains [6%, 25%] (Table 2, Figure 2). The two-timescale fit places 54% of the amplitude on a slow component with a timescale of 824 ms. The point estimate is still above a tenth of the initial amplitude at the last measured lag, 1,000 ms, so τ_{90} lies beyond that lag, with an interval bounded below at 583 ms. The same limit is why the required history is written $c^* \geq 960$ ms. When dependence outlives the measured range, both report that range, not a number they cannot see.

FlyWire. FlyWire provides the contrasting case. Under the same neuron dynamics and class of drive, its predictable dependence falls to a tenth of its initial amplitude by $\tau_{90} = 83$ ms [63, 99], no detectable power remains at 300 ms, and the two-timescale fit collapses to a single component ($a = 1.00$). Its students need correspondingly little history, $c^* = 80$ ms [40, 80], and the memoryless MLP lies 0.024 below the GRU, against a gap of 0.15 on MaleCNS. Where the teacher has one timescale, probes, dependence curve and student agree to within a factor of four (Figure 2, right). The two τ_{90} intervals do not overlap, the lower end for MaleCNS being nearly six times FlyWire's upper end.

MaleCNS also carries 186 motor channels with no counterpart in FlyWire. Scored on its 486 descending types alone, it keeps both the slow component and the long required history (Table 2, middle column), so the difference between the connectomes is not an artefact of the extra channels.

6 Understanding the gap between probe horizon and predictable memory

6.1 Why a probe horizon can stay short while predictable memory grows

The two connectomes also differ in operating point. Under the same drive, MaleCNS fires at five times FlyWire's rate, 14.4 Hz across the graph against 2.8. As a control on the comparison, we reduced its synaptic weight from the published 0.275 mV to 0.1063 mV, bringing it to 2.87 Hz.

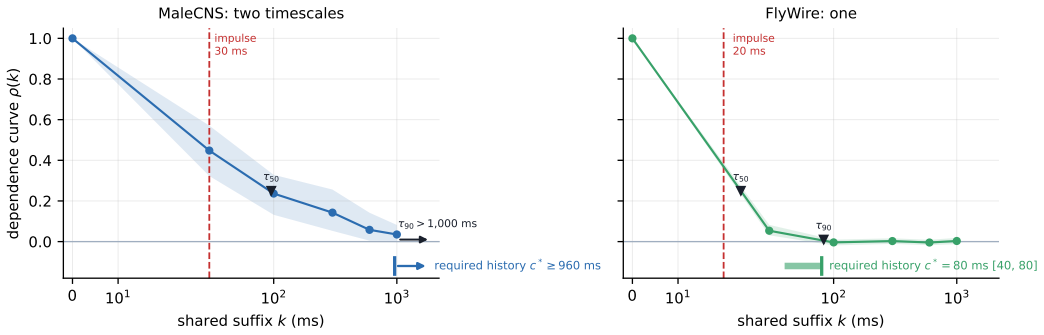

Figure 2: Dependence curve $\rho(k)$ for MaleCNS (left) and FlyWire (right), with bootstrap band. Dashed: the impulse horizon, 30 ms and 20 ms. Triangles: τ_{50} and τ_{90}; on MaleCNS an arrow marks a τ_{90} beyond the last measured lag. Bar: the c^* interval, with c^* at the tick; an arrow marks c^* at the top of its grid. On MaleCNS the impulse horizon sits in the first tenth of the curve and the tail runs past the last measured lag; on FlyWire the curve, the impulse horizon and the student agree to within a factor of four.

Table 3: A gain-matched intervention on MaleCNS. The predictable power left at 1 s grows ninefold. None of the three conventional probes follows it, and two of them shorten; only the floor-corrected settling reading in the last row lengthens. Settling is read as the last lag still above the level, the rule of Appendix F, so the baseline reads 140 ms rather than the 150 ms of §3.

	Published gain	Gain-matched
Whole-graph firing rate	14.4 Hz	2.87 Hz
Predictable power left at 1 s, $\rho(1000\,\text{ms})$	0.035 [0.00, 0.08]	0.320 [0.15, 0.49]
Impulse horizon (Holm)	30 ms	20 ms
Autocorrelation, $1/e$	<10 ms	<10 ms
Settling, $1/e$	140 ms	120 ms
Settling, 10% of floor-corrected excess	220 ms	>1,490 ms (end of window)

The gain-matched intervention separates the two quantities directly (Table 3). Predictable power left at 1 s rises ninefold, from 0.035 to 0.320, with intervals that do not overlap, while none of the three conventional probes reports a longer horizon. The impulse horizon falls from 30 to 20 ms, the autocorrelation time stays inside one bin, and settling at $1/e$ shortens from 140 to 120 ms. Predictable memory can therefore grow while all three conventional probes report the same or a shorter horizon. At FlyWire's rate MaleCNS retains more predictable power, not less, so the gap between the connectomes is not one of operating point.

Read at 10% of its *floor-corrected* excess, the same step response moves the other way, from 220 ms to past the end of the 1,500 ms step-response window. The two readings disagree in direction, and what separates them is the floor. A raw settling curve approaches a non-zero stochastic level, so a fractional threshold on it times the moment the predictable signal falls into that level, which strengthening a slow component can move earlier. Subtracting the floor first turns the same curve into a test of whether predictable influence has vanished, and the corrected reading tracks the intervention.

With the impulse result of §5.1, this gives both reasons a conventional probe can underestimate predictable memory. Its excitation may engage a slow component too weakly to pass a test, as a 10 ms pulse does a 500 ms state, or its statistic may reach a stochastic floor while predictable dependence remains, as the raw settling curve does here. The shared-suffix measurement avoids the first by using the surrogate's own stimulus family, and the second by subtracting the floor.

6.2 Required history co-varies with predictable dependence

If the required history measures predictable memory, it should move when predictable memory moves. Beyond the two connectomes and two synthetic teachers, four further conditions test this.

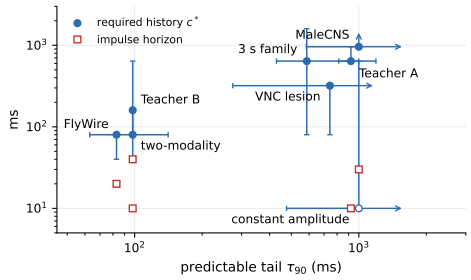

Figure 3: Required history c^* against the predictable tail τ_{90}, log–log, with 95% bootstrap intervals on both axes. Unprefixed labels are MaleCNS conditions. Up arrow: c^* at the top of its grid. Right arrow: an interval extending beyond the last measured lag, 1,000 ms, where MaleCNS and the constant-amplitude family are placed. Open marker: the constant-amplitude family, whose drive contains no transitions. Open squares: the impulse horizon of the teachers on which it was measured, which spans a factor of four while τ_{90} and c^* each span more than an order of magnitude.

Episode length. The 3 s family comes from the same stochastic generator as the main family, with 2 to 6 switches per episode instead of 1 to 3. On it, τ_{90} shortens from more than 1 s to 586 ms [429, 1,192] and c^* from at least 960 ms to 640 ms [80, 1600]. We treat this family as an intervention on episode length, not as a change of the stationary input process.

A structural perturbation. FlyWire ends at the neck and MaleCNS includes the VNC, which makes it the one structure that can be removed from one teacher and not the other. Removing every VNC neuron and every edge touching one (22,822 neurons and 17% of edges, sparing all wind afferents and descending neurons) shortens τ_{90} from more than 1 s to 743 ms, the slow timescale from 824 to 186 ms, and c^* from at least 960 ms to 320 ms [80, 320] (control in Appendix I).

A drive without transitions. The constant-amplitude family holds the wind drive at one amplitude for each whole 1.5 s episode, so its natural input never changes. Splicing two such episodes at $t - k$ creates a step, and the shared-suffix measurement returns a long tail, with τ_{90} beyond 1,000 ms, the last measured lag. The student needs no history, $c^* = 10$ ms [10, 640], with memoryless fidelity 0.980 against a best of 0.981, since on the natural distribution the response is a function of the current amplitude. This marks the boundary of the correspondence. The measurement captures memory under a drive with one transition at lag k, and here that transition exists only in the measurement, so it times settling after an artificial step rather than history useful for prediction. It matches what a surrogate needs when the natural drive itself contains transitions on that timescale, as the main family does (median constant segment 390 ms). That condition is read off the stimulus family, before anything about the teacher's timescale is known.

Across conditions. Figure 3 collects every condition in which both τ_{90} and c^* were measured, including a two-modality family on MaleCNS, which drives wind and gustatory afferents through four drive groups that switch independently on their own schedules (median segment 400 ms). On that family $\tau_{90} = 98$ ms [96, 141] and $c^* = 80$ ms [40, 640]. Among the conditions whose drive contains transitions, the three whose tails are under 100 ms need at most 160 ms of history and the four whose tails are above 500 ms need 320 ms or more, on point estimates, while the impulse horizon, wherever it was measured, stays between 10 and 40 ms. Because c^* is set by a performance tolerance while τ_{90} summarises a curve, we do not claim that c^* equals any quantile of the curve, and the figure carries no identity line. The figure shows that a quantity read off a trained surrogate and one measured on the teacher by intervention agree about which conditions have long memory, and that the impulse horizon does not.

7 Related work

Probes that set a model's past. The practices listed in §1 all read a probe horizon and use it in place of predictable memory. Our results apply wherever the excitation engages a slow component

weakly or the statistic being read has a stochastic floor, and on a complete connectome model both hold. Neural operators and physics emulators (Li et al., 2021) condition on a short fixed history, and the procedure of §8 applies to them directly.

What is known about these probes. That a linear or short-kick probe does not exhaust a system's memory is an old result. Jaeger (2002) bounded a recurrent network's memory capacity by its state dimension, and Ganguli et al. (2008) showed that memory in non-normal networks is not set by the eigenvalue spectrum, so a summary read off one response can mischaracterise it. Boyd & Chua (1985) give the fading-memory condition under which a finite past window suffices at all, and nonlinear system identification separates the impulse response of the linearisation from the memory of the system (Schoukens & Ljung, 2019). Zeraati et al. (2022) show that the standard autocorrelation estimator of an intrinsic timescale is biased on finite data. We add the size of the gap on a complete connectome model, an intervention that separates the two quantities causally, and a measurement of predictable memory. Correcting for trial-to-trial variability is standard when scoring models of noisy neural data, where an explainable-variance ceiling comes from repeated presentations (Hsu et al., 2004; Schoppe et al., 2016). Our fidelity anchor is that correction, and the estimator of §4 applies the same idea to a difference between two conditions rather than to one response, which makes it additive in squared distance.

Reading a teacher from its student. Connectome-constrained models (Shiu et al., 2024; Lappalainen et al., 2024) and complete connectomes (Dorkenwald et al., 2024; Schlegel et al., 2024; Berg et al., 2026) supply our teachers. Closest in spirit, Beniaguev et al. (2021) read the computational complexity of a biophysical neuron from the depth and temporal extent of a network that emulates it. We make the same reading for memory and supply what it needs to be trusted: interventions that exclude other uses of the window, a teacher with a planted timescale, and an independent measurement it agrees with. Behavioural cloning studies what a clone loses under distribution shift (Ross et al., 2011); we hold the distribution fixed and read the teacher from what the clone needs.

8 Discussion

What the surrogate measures. A student scored against the $D_{TT}/2$ floor carries every predictable component of the response, weighted by the variance it explains rather than by how easily a probe excites it, so its required history reads out the predictable memory available under its input distribution. Predictable memory is not a property of the teacher alone: it depends jointly on teacher, drive and task, as the 3 s and constant-amplitude families show. This is by design, since the history that matters is the one a surrogate needs under the distribution it will be run on.

Setting a surrogate's history. Four inexpensive steps carried the argument here. Score the student against $D_{TT}/2$, so that trajectory-level noise is not charged to it. Sweep the window, then repeat with a larger training bank, a different input noise level and a different width; a window that shrinks under any of these was spent on memorisation, denoising or capacity. Where the input can be spliced and the natural drive contains transitions, measure predictable dependence on the noise-averaged response, subtracting the floor in power. And report where each curve was cut off.

Scope. The teachers are two connectome models under one set of published dynamics and two synthetic recurrent networks, driven on the wind and, in one family, the gustatory pathway. The students are one windowed MLP and one GRU, varied in parameter count but not in depth or inductive bias. Whether required history tracks predictable dependence for non-stationary simulators, adaptive dynamics and other surrogate families is open.

9 Conclusion

A simulator can retain predictable memory long after its conventional transient has gone, and this matters whenever a surrogate must reproduce the simulator rather than its response to one probe. Here the history a surrogate needs tracks floor-corrected predictable dependence wherever the drive contains transitions, and survives large changes in training data and capacity. A timescale used to design a surrogate should be measured on what it must predict, under the distribution it runs in.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code for the teachers, the three conventional probes, the shared-suffix measurement and the student training and evaluation, with the YAML configurations and seeds used for the reported runs. The neuron model and its parameters are given in Appendix A, training hyperparameters in Appendix C, the synthetic teachers in Appendix D and the probe protocols in Appendix E. Connectome data are fetched from their official sources by a script in the supplementary material and are not redistributed; FlyWire is CC-BY-NC and is used for research only.

AI USE STATEMENT

Large language models were used to polish the language of the manuscript, to improve the readability of code that the authors had already completed, and to give feedback on the paper during writing. All edits and suggestions from these tools were reviewed by the authors, who take full responsibility for the content of this work.

REFERENCES

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv:1607.06450*, 2016.

D. Beniaguev, I. Segev, and M. London. Single cortical neurons as deep artificial neural networks. *Neuron*, 109(17):2727–2739, 2021.

S. Berg, I. R. Beckett, M. Costa, P. Schlegel, M. Januszewski, E. C. Marin, et al. Sexual dimorphism in the complete Drosophila male central nervous system connectome. *Cell*, 189(18):5504–5526.e15, 2026. doi: 10.1016/j.cell.2026.08.015.

S. Boyd and L. O. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32:1150–1161, 1985.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.

S. Dorkenwald et al. Neuronal wiring diagram of an adult brain. *Nature*, 634:124–138, 2024.

A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33:1134–1140, 1986.

S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105:18970–18975, 2008.

C. E. García, D. M. Prett, and M. Morari. Model predictive control: theory and practice—a survey. *Automatica*, 25(3):335–348, 1989.

D. Hendrycks and K. Gimpel. Gaussian error linear units (GELUs). *arXiv:1606.08415*, 2016.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

A. Hsu, A. Borst, and F. E. Theunissen. Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems*, 15(2): 91–109, 2004.

H. Jaeger. Short term memory in echo state networks. GMD Report 152, German National Research Center for Information Technology, 2002.

A.-M. Jürgensen, A. Khalili, E. Chicca, G. Indiveri, and M. P. Nawrot. A neuromorphic model of olfactory processing and sparse coding in the Drosophila larva brain. *Neuromorphic Computing and Engineering*, 1(2):024008, 2021.

K. S. Kakaria and B. L. de Bivort. Ring attractor dynamics emerge from a spiking model of the entire protocerebral bridge. *Frontiers in Behavioral Neuroscience*, 11:8, 2017.

H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, second edition, 2004.

J. K. Lappalainen et al. Connectome-constrained networks predict neural activity across the fly visual system. *Nature*, 634:1132–1140, 2024.

A. A. Lazar, T. Liu, M. K. Turkcan, and Y. Zhou. Accelerating with FlyBrainLab the discovery of the functional logic of the Drosophila brain in the connectomic and synaptomic era. *eLife*, 10: e62362, 2021.

Z. Li et al. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

J. D. Murray et al. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, 17:1661–1663, 2014.

M. M. Paul, M. Pauli, N. Ehmann, S. Hallermann, M. Sauer, R. J. Kittel, and M. Heckmann. Bruchpilot and Synaptotagmin collaborate to drive rapid glutamate release and active zone differentiation. *Frontiers in Cellular Neuroscience*, 9:29, 2015.

S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Artificial Intelligence and Statistics*, pp. 627–635, 2011.

P. Schlegel et al. Whole-brain annotation and multi-connectome cell typing of Drosophila. *Nature*, 634:139–152, 2024.

O. Schoppe, N. S. Harper, B. D. B. Willmore, A. J. King, and J. W. H. Schnupp. Measuring the performance of neural models. *Frontiers in Computational Neuroscience*, 10:10, 2016.

J. Schoukens and L. Ljung. Nonlinear system identification: a user-oriented road map. *IEEE Control Systems Magazine*, 39:28–99, 2019.

P. K. Shiu et al. A Drosophila computational brain model reveals sensorimotor processing. *Nature*, 634:210–219, 2024.

R. Shridhar and D. J. Cooper. A tuning strategy for unconstrained SISO model predictive control. *Industrial & Engineering Chemistry Research*, 36(3):729–746, 1997.

M. Stimberg, R. Brette, and D. F. M. Goodman. Brian 2, an intuitive and efficient neural simulator. *eLife*, 8:e47314, 2019.

E. Tan, S. Algar, D. Corrêa, M. Small, T. Stemler, and D. Walker. Selecting embedding delays: an overview of embedding techniques and a new method using persistent homology. *Chaos*, 33(3): 032101, 2023.

Y. Tian, Y. T. Lin, M. Anghel, and D. Livescu. Data-driven learning of Mori–Zwanzig operators for isotropic turbulence. *Physics of Fluids*, 33(12):125118, 2021.

Q. Wang, N. Ripamonti, and J. S. Hesthaven. Recurrent neural network closure of parametric POD-Galerkin reduced-order models based on the Mori–Zwanzig formalism. *Journal of Computational Physics*, 410:109402, 2020.

Z. Yao and K. Scott. Serotonergic neurons translate taste detection into internal nutrient regulation. *Neuron*, 110(6):1036–1050.e7, 2022. doi: 10.1016/j.neuron.2021.12.028.

R. Zeraati, T. A. Engel, and A. Levina. A flexible Bayesian framework for unbiased estimation of timescales. *Nature Computational Science*, 2:193–204, 2022.

A NEURAL DYNAMICS AND NEURON ACCOUNTING

Membrane equation. Outside the refractory period, for each neuron i,

$$\frac{dv_i}{dt} = \frac{v_{\text{rest}} - v_i + g_i}{\tau_m}, \qquad \frac{dg_i}{dt} = -\frac{g_i}{\tau_s}, \tag{3}$$

with g a conductance term carried in units of voltage, as in the reference implementation of Shiu et al. (2024). A neuron spikes when $v_i > v_{th}$, after which $v_i \leftarrow v_{\text{reset}}$, $g_i \leftarrow 0$, and both equations are frozen for t_{ref}. A spike in neuron j reaches each postsynaptic partner i after a fixed delay and adds

$$g_i \mathrel{+}= s_j \, n_{j \to i} \, w_{\text{syn}}, \tag{4}$$

where $n_{j \to i}$ is the number of synapses from j to i in the connectome and $s_j \in \{+1, -1\}$ is the sign of j's neurotransmitter.

Parameters. Every value is copied from the reference implementation, together with the citation recorded there; none is re-tuned here.

Symbol	Value	Meaning	Source
v_{rest}, v_{reset}	-52 mV	resting and reset potential	Kakaria & de Bivort (2017)
v_{th}	-45 mV	spike threshold	Kakaria & de Bivort (2017)
τ_m	20 ms	membrane time constant	Kakaria & de Bivort (2017)
τ_s	5 ms	synaptic time constant	Jürgensen et al. (2021)
t_{ref}	2.2 ms	refractory period	Lazar et al. (2021)
t_{delay}	1.8 ms	synaptic delay	Paul et al. (2015)
w_{syn}	0.275 mV/syn.	synaptic weight scale	free parameter of Shiu et al. (2024)
dt	0.1 ms	integration timestep	Brian2 default, as in the reference
r_{poi}, f_{poi}	150 Hz, 250	stimulation rate, scaling	Shiu et al. (2024)
w_{poi}	68.75 mV	$= w_{\text{syn}} f_{\text{poi}}$, jump per event	derived

Neurons receiving Poisson stimulation have their refractory period set to zero, as in the reference implementation, so that an afferent rate is transduced faithfully instead of being clipped at $1/2.2\,\text{ms} = 455\,\text{Hz}$.

Integration. Both equations are linear between spikes, so each timestep is advanced with the exact propagator rather than by Euler steps: $g(t+dt) = g(t)e^{-dt/\tau_s}$ and $v(t+dt) = v_{\text{rest}} + (v(t) - v_{\text{rest}})E_m + (g(t)/\tau_m)(E_s - E_m)/(\tau_m^{-1} - \tau_s^{-1})$ with $E_m = e^{-dt/\tau_m}$, $E_s = e^{-dt/\tau_s}$. This matches the `method='linear'` integration of Brian2 (Stimberg et al., 2019). Only the outgoing edges of neurons that spiked are touched each step.

Neuron accounting, MaleCNS v1.0. We simulate every body carrying a `superclass` annotation: **166,700** neurons, of which **166,483** have at least one synaptic connection and **217** have none and are simulated as isolated. Glia, orphan fragments and unannotated segments are excluded. The graph retains **25,582,938** neuron-to-neuron edges carrying **124,177,617** synapses (39.8% of the 311.8M in the release; the remainder touch fragments and glia), across **11,751** distinct type labels, 43 of which are `_unclear` placeholders, against the 11,710 neuron types of the release. Signs come from the official `consensus_nt`: 107,438 excitatory and 59,262 inhibitory, with **3,177 neurons (1.9%)** carrying a guessed sign. Histamine is treated as inhibitory, which is this paper's one extension to the published assignment.

Neuron accounting, FlyWire v783. **138,639** of the release's 139,255 proofread neurons, carrying **15,091,983** unthresholded edges and **54,492,922** synapses, brain only: the volume ends at the neck connective, so there is no ventral nerve cord and no leg or wing motor pool.

Readout populations. MaleCNS: 1,332 descending neurons in 485 named cell types and 815 motor neurons in 185 named types, each pool carrying one further channel for the bodies with no type, 672 output channels in all. FlyWire: 1,299 descending neurons in 472 cell types and no motor readout. Rates are pooled per cell type in 10 ms bins, which is the level at which the connectome is annotated and at which a rate is stable across noise seeds.

Silent channels. A cell type whose rate is constant across a split is excluded from per-channel R^2 and counted rather than clamped, since R^2 is undefined for it. Of the 672 channels, 263 are silent under wind drive. Fidelity is a mean squared error over all channels and is unaffected by the exclusion.

B TEACHER VALIDATION

We drive the sugar pathway and read the feeding motor neuron MN9, following Shiu et al. (2024). Rates are means over three Poisson seeds of a 500 ms presentation, and the rewiring control is a degree-preserving shuffle of the same graph.

Table 4: Sugar/bitter validation. FlyWire uses the published sugar and bitter GRN identifiers; MaleCNS annotates neither, so its sugar pathway is defined from connectivity as a gustatory receptor neuron synapsing onto a sugar-SEL neuron (Yao & Scott, 2022).

Connectome	Condition	Driven	MN9 (Hz)
FlyWire	sugar	20	112.7
	bitter	20	0.0
	sugar + bitter	40	1.7
	size-matched control	20	0.0
	sugar, rewired	20	0.0
MaleCNS	sugar	44	39.0
	size-matched control	44	2.3
	sugar, rewired	44	0.0

The bitter arm. MaleCNS v1.0 carries no bitter receptor annotation and a connectivity-defined proxy was not specific, so the MaleCNS teacher is validated on the sugar arm.

C TRAINING AND STUDENT ARCHITECTURE

Windowed MLP. The input is the drive vector of the last c bins, stacked, so its width is $2c$ on the wind families. A linear projection to width h is followed by LayerNorm (Ba et al., 2016), two blocks of $\mathrm{Linear}(h, h)$ and GELU (Hendrycks & Gimpel, 2016), and a linear head to the 672 output channels. There is no dropout. At the start of an episode the window is filled by repeating the first bin; the first 20 bins are excluded from scoring, so for $c > 21$ the earliest scored bins still contain padding.

GRU. A single layer of hidden dimension h with a linear head to the same 672 channels.

Optimisation. AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-3}, not scaled with width, weight decay 0, gradient-norm clipping at 1.0, 200 warmup steps and then a cosine schedule to 5% of the peak, batch size 32, at most 20,000 steps. Validation runs every 200 steps, and training stops when the validation loss has not improved for 12 evaluations or its relative change over 5 evaluations is below 0.005. Every fit in the width sweep converged under this rule, with a median best step of 2,200.

Sequences and targets. Chunks are 150 bins long, which is the whole episode on the 1.5 s families and half of it on the 3 s family, with 20 warmup bins excluded from the loss and a stride of 10 bins. Targets are $\log(1 + r)$ of the rate r in Hz; fidelity is computed on the untransformed rate.

Cost. CPU only, one core per fit. The median fit took 50 s (range 37–78 s) across the 360 fits of the width sweep, which uses ten seeds per cell rather than three.

The GRU is used as the recurrent reference because it is the strongest of the sequence students we trained under the same protocol on MaleCNS: at width 64 it reaches 0.903, against 0.880 for a state-space model and 0.874 for a Transformer. Full fidelity-versus-width and fidelity-versus-window tables for every architecture and connectome, unclipped, are in the supplementary material.

Table 5: Windowed-MLP fidelity on MaleCNS against training-bank size, $h = 32$. The required history does not move. The $n = 64$ column is an independent run of the main configuration, which is why its memoryless cell differs from Table 6 by 0.005.

Window (ms)	$n = 64$	$n = 256$	$n = 1024$
10	0.762	0.793	0.794
80	0.801	0.849	0.855
320	0.845	0.897	0.903
960	0.875	0.929	0.928
Gain from long window	+0.113	+0.136	+0.135
c^*	$\geq$960	$\geq$960	$\geq$960

Table 6: Windowed-MLP fidelity on MaleCNS against hidden width, spanning a 9.6-fold range in parameter count (eightfold in hidden units), at ten seeds per cell. Memoryless fidelity varies by 0.006 across the four widths and every width keeps $c^* \geq 960$ ms.

Window (ms)	$h = 16$	$h = 32$	$h = 64$	$h = 128$
Parameters at $c = 96$	15,088	30,528	64,480	144,672
10	0.758	0.757	0.756	0.752
80	0.804	0.801	0.803	0.796
320	0.844	0.845	0.845	0.839
960	0.863	0.875	0.883	0.873
Gain, 960 over 10 ms	$+0.105$	$+0.118$	$+0.127$	$+0.121$
c^*	$\geq$960	$\geq$960	$\geq$960	$\geq$960

D SYNTHETIC TEACHERS

Both teachers are a 200-unit tanh recurrent network with $\tau = 20$ ms, integrated at $dt = 1$ ms and read out in 10 ms bins after a 50 ms warmup, with 100 output channels. Every matrix is drawn from one seed and saved, so the two teachers differ in exactly one number. The recurrent matrix is Gaussian, rescaled to spectral radius 0.9 from 0.9974. The input consists of the rate profiles the fly teachers see, passed through a fixed random projection, with multiplicative noise applied per step at 0.0854 so that the binned drive deviates from its rate by 0.027, the fly's figure. The slow variable is one extra scalar state with $\tau = 500$ ms, driven by a one-dimensional random projection of the input and added to the output with gain 3.0177 in teacher A, the value that gives it a 0.300 share of the predictable output variance, and 0 in teacher B. Readout noise is Gaussian per bin with $\sigma = 0.2785$, set so that the single-trial signal-to-noise ratio matches the fly's. The predictable variance is 0.0956 for A and 0.0672 for B.

E PROBE PROTOCOLS

Impulse horizon. The stimulus is left intact apart from a rate perturbation of fixed magnitude on a single 10 ms bin at lag k before a probe time, with 32 Poisson seeds per condition (24 for the synthetic teachers of §5.1). The effect is $\|\bar{y}_{\mathrm{pert}} - \bar{y}_{\mathrm{ctrl}}\|_2$ at the probe time, tested by permutation with 2,000 relabelings, $p = (b + 1)/(n + 1)$, and Holm correction (Holm, 1979) over 100 lags. The horizon is the last Holm-significant lag, 30 ms on MaleCNS and 20 ms on FlyWire. On the two-modality drive of §6.2 it is 40 ms, and on the gain-matched teacher 20 ms.

Settling time. The drive is held at one amplitude for 1,000 ms, switched to a second, and held for 1,500 ms, for 32 amplitude pairs at 32 Poisson seeds each. The settling time is the time for $\|\bar{y}(t) - \bar{y}_\infty\|_2$ to fall to $1/e$, with $\bar{y}_\infty$ the mean over the last 500 ms. Averaging over seeds is what makes this measurable: on a single trial the 672-channel distance is dominated by Poisson noise and never reaches $1/e$, and a per-channel reading is set by rate quantisation rather than by the response, since a two-neuron cell type moves in steps of 50 Hz. A control holds the second amplitude throughout. The floor-corrected reading of Table 3 subtracts, in squared distance, the distance between two independent 16-seed means of that control, and is read as the last bin still above 10% of the corrected excess. The analysed window is the 1,500 ms after the switch, so its last

Table 7: Every probe at a strict and a lenient threshold, on the four teachers. All values in ms. The strict readings are the Holm-corrected impulse horizon and $1/e$; the lenient ones are raw $\alpha = 0.05$ and 5%. Autocorrelation and settling are read as the last lag still above the level, the same rule the impulse horizon uses. c^* at the top of its grid is written $\geq$, and τ_{90} beyond the last measured lag is written $>$.

Probe	MaleCNS	FlyWire	Teacher A	Teacher B
Impulse horizon, Holm	30	20	10	10
Impulse horizon, raw α	50	30	20	20
Autocorrelation, $1/e$	0	0	0	0
Autocorrelation, 5%	190	210	280	240
Settling, $1/e$	140	20	240	30
Settling, 5%	1490	1490	1490	1490
Settling, 10% of excess	220	30	440	50
Shared-suffix measurement, τ_{50}	96	24	90	28
Shared-suffix measurement, τ_{90}	>1,000	83	922	98
Required history c^*	$\geq$ 960	80	640	160

bin is at 1,490 ms and a reading of 1,490 is the end of the window rather than a resolved crossing. At 1,000 ms the gain-matched excess is still 0.321 of its initial value, against 0.014 at the published gain.

Autocorrelation time. We compute the autocorrelation of each descending channel within an episode, after removing that channel's mean in that episode, and average over channels and 60 held-out episodes. The lag-1 correlation is already 0.094, below $1/e$, so the horizon is under one 10 ms bin. The lag is read as the last one still above the level, because the curves are not monotone: FlyWire's lag-1 correlation is negative.

F Every probe at a common threshold

Each conventional probe carries its own stopping rule, so Table 7 reads every probe at a strict and at a lenient level. Autocorrelation and settling are read as the last lag still above the level, the rule the impulse horizon uses, which is why settling at $1/e$ appears as 140 ms there and as 150 ms in §3.

At the strict readings the impulse probe puts all four teachers between 10 and 30 ms and does not separate teacher A from teacher B, which differ only in a 500 ms state. Autocorrelation separates nothing at either level: at $1/e$ it crosses within one bin on all four, and at 5% it returns 190–280 ms on all four. Settling at $1/e$ is the one conventional probe that separates the teachers into a long and a short group, though not in the order the required history gives, and it falls short of c^* on all four. Settling at 5%, close to the 2–5% band used to size step-response models in control practice (Shridhar & Cooper, 1997), runs to the end of the step-response window on all four and separates nothing. A seed-averaged step response read at 10% of its floor-corrected excess is the closest any probe reading comes, and §6.1 shows what the floor correction changes.

This paper reports τ_{90} and defines c^* by a tolerance, so it does not escape analysis choices. What differs is where the choice sits. A probe threshold is part of acquisition, deciding when measurement stops, and a pulse length must be chosen before the timescale it should find is known, whereas a quantile of ρ summarises a curve already estimated. The $1/e$ convention is not wrong, but it is applied to a curve still containing the teacher's noise, and §6.1 shows the cost.

G How much budget the impulse probe would need

Teacher A is the only teacher for which we know the answer the probe is asked for, so it is where the probe's power can be measured rather than argued. A pulse of length T steps a component with time constant τ by $1 - e^{-T/\tau}$, so a 10 ms pulse steps a 500 ms integrator by 2.0%, a 50 ms pulse by 9.5% and a 200 ms pulse by 33.0%.

Table 8: Impulse probe on teacher A, whose slow variable is a known 500 ms integrator, over pulse length and seed budget at fixed magnitude. Each cell gives the Holm-corrected impulse horizon and, in brackets, the p-value of the effect at lag 500 ms. The published protocol is the top-left cell. The effect at that lag depends on the pulse and not on the seeds: 0.066, 0.340 and 1.590 for the three pulse lengths, against the 2.0%, 9.5% and 33.0% steps a 500 ms integrator takes.

	seeds per condition		
Pulse	24	96	384
10 ms	10 ms [1.00]	30 ms [1.00]	60 ms [1.00]
50 ms	80 ms [1.00]	110 ms [1.00]	150 ms [1.00]
200 ms	240 ms [1.00]	400 ms [0.08]	800 ms [<0.001]

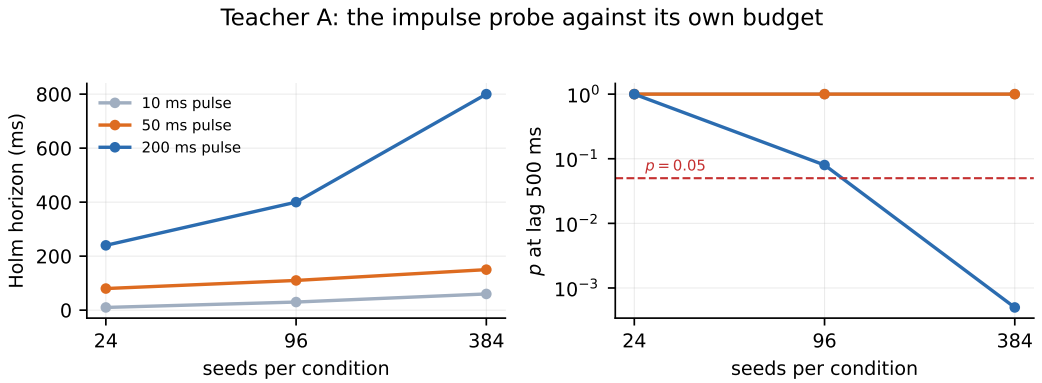

Figure 4: Left: the impulse horizon grows with both pulse length and seed budget, as a significance threshold must. Right: the p-value of the effect at lag 500 ms, where the slow state lives. Only the 200 ms pulse reaches it, and only at 384 seeds.

The horizon grows with both pulse length and seeds, as a significance threshold must (Figure 4). What matters is the effect at lag 500 ms, where the slow state lives. That effect is set by the pulse and not by the seeds, and it becomes significant only in the bottom-right entry. The standard protocol, a 10 ms pulse at 24 seeds, leaves an effect there 24 times smaller than the 200 ms pulse does, so matching the sensitivity that resolved it needs about 24^2 times 384 seeds, roughly 2×10^5 per condition. The shared-suffix measurement that finds the same state costs 4,608 episodes. The row that succeeds uses a pulse of 200 ms, which is the right choice for a 500 ms state and one that an experimenter can make only after learning that the state is there.

H ESTIMATOR VALIDATION IN DETAIL

Table 9: Floor correction in distance space against power space, on a benchmark with a known dependence curve. All values in ms. The distance-domain estimator shortens the tail by 42% at τ_{90} and the slow timescale by 31%; the power-domain estimator is within 5% of the truth row everywhere.

Estimator	τ_{50}	τ_{90}	τ_{95}	τ_2
Truth	30	561	924	479
Distance-domain estimator	27	325	558	329
Power-domain estimator (this paper)	30	572	953	500

The benchmark has 672 channels, $M = 16$ seeds per mean, per-channel noise $\sigma = 1$ and 32 pairs, drawn once at a fixed seed. The planted amplitude is $100\,(0.70\,e^{-k/25} + 0.30\,e^{-k/500})$ on the seven-point k grid of the experiment, forced to 0 at $k = 1500$ where the two episodes coincide. Each distance is the norm of a real 672-vector plus two independent noise vectors, so the estimator sees

data of the same shape the experiment produces. If a seed mean over n_{ch} channels is written as $f+\epsilon$ with per-channel noise variance σ^2/M, the expected squared distance is $\|f_a - f_b\|^2 + 2n_{\mathrm{ch}}\sigma^2/M$ and the floor has expectation $2n_{\mathrm{ch}}\sigma^2/M$ alone, so $D^2 - F^2$ is unbiased everywhere, including where the signal is zero. All three rows of Table 9 read the same quantiles: the truth and distance rows at amplitude 0.50, 0.10 and 0.05, the power row at their squares.

Identity check. At the largest k the two spliced episodes are identical, so the true predictable dependence is zero and an unbiased estimator should fluctuate symmetrically about it. Across conditions the power-domain estimator returned values between -0.007 and $+0.008$.

I EVERY CONDITION

Random-lesion control. The size-matched random lesion of Table 10 acts differently from the VNC lesion of §6.2. It lowers $\rho(30\,\mathrm{ms})$ from 0.448 to 0.105 [0.075, 0.144] and leaves nothing at 100 ms, where the VNC lesion still retains 0.207. A lesion moves the operating point as well as the pathway, so we read the VNC lesion as a perturbation of predictable dependence, not as the removal of one slow mode or as an anatomical localisation.

Table 10: Predictable dependence of the noise-averaged response at four lags, for every condition, power-domain estimator, with bootstrap 95% intervals over 32 stimulus pairs. Rows use the main family except the 3 s row, which is read against its own k grid and its own identity point at $k = 3000$, and the constant-amplitude and two-modality rows (§6.2). *VNC lesion* removes every VNC neuron and every edge touching one; *random lesion* removes as many neurons drawn uniformly from everything but the driven afferents and the readouts. Table 11 gives the tails and fits of the same rows.

Condition	$\rho(100)$	$\rho(300)$	$\rho(600)$	$\rho(1000)$
MaleCNS	0.236 [0.13, 0.33]	0.143 [0.06, 0.25]	0.058 [0.00, 0.13]	0.035 [0.00, 0.08]
MaleCNS, 3 s family	0.170 [0.09, 0.27]	0.045 [0.02, 0.09]	0.008 [0.00, 0.02]	−0.000 [−0.01, 0.01]
MaleCNS, VNC lesion	0.207 [0.08, 0.31]	0.015 [0.00, 0.04]	0.015 [0.00, 0.04]	0.001 [−0.01, 0.01]
MaleCNS, random lesion	−0.004 [−0.02, 0.02]	−0.013 [−0.03, 0.00]	−0.005 [−0.02, 0.01]	−0.008 [−0.03, 0.01]
MaleCNS, gain-matched	0.590 [0.45, 0.71]	0.467 [0.28, 0.63]	0.432 [0.23, 0.61]	0.320 [0.15, 0.49]
MaleCNS, constant-amplitude family	0.168 [0.05, 0.29]	0.122 [0.03, 0.23]	0.076 [0.00, 0.16]	0.040 [−0.01, 0.10]
MaleCNS, two-modality family	0.006 [0.00, 0.01]	0.001 [0.00, 0.00]	0.001 [0.00, 0.00]	0.002 [0.00, 0.01]
FlyWire	−0.004 [−0.02, 0.01]	0.003 [−0.01, 0.02]	−0.005 [−0.02, 0.01]	0.003 [−0.01, 0.02]
Teacher A	0.223 [0.12, 0.33]	0.106 [0.05, 0.17]	0.029 [0.02, 0.04]	0.005 [0.00, 0.01]
Teacher B	0.003 [0.00, 0.01]	0.003 [0.00, 0.01]	0.002 [0.00, 0.01]	0.001 [0.00, 0.00]

Table 11: Predictable tail and two-timescale fit for the rows of Table 10. τ_{90} is read from the measured lags only, where the amplitude falls to 0.1, that is $\rho = 0.01$. Where ρ is still above that level at the last measured lag, 1,000 ms, τ_{90} is written >1,000, and an interval that extends beyond that lag has an open upper end (—). The fit uses the same lags and is $[ae^{-k/\tau_1} + (1-a)e^{-k/\tau_2}]^2$; where a is at or near 1 the slow component carries no weight and τ_2 is unidentified (—). ‡: the fitted τ_2 lies beyond the last measured lag and is not reported. $D(k_{\max})/F$ is the identity check at each row's own identity point, where the two episodes are the same episode.

Condition	τ_{90} (ms)	a, τ_1, τ_2	$D(k_{\max})/F$
MaleCNS	>1,000 [583, —]	0.46, 27, 824	1.01
MaleCNS, 3 s family	586 [429, 1192]	0.43, 17, 307	0.99
MaleCNS, VNC lesion	743 [274, —]	0.23, —, 186	1.03
MaleCNS, random lesion	91 [80, 154]	1.00, 27, —	0.99
MaleCNS, gain-matched	>1,000	0.22, 23, ‡	0.98
MaleCNS, constant-amplitude family	>1,000 [476, —]	0.59, 34, ‡	0.98
MaleCNS, two-modality family	98 [96, 141]	0.96, 32, —	1.00
FlyWire	83 [63, 99]	1.00, 20, —	0.98
Teacher A	922 [814, 999]	0.43, 21, 521	1.00
Teacher B	98 [96, 99]	1.00, 36, —	1.00