Gradient descent does not preserve a trainable PING loop

exp049 · 9 June 2026

The trained networks this entry uses are produced once in the shared training hub, exp022 (Training), and reused here rather than retrained.

Abstract

exp025 freezes the recurrent conductances W^{EI}, W^{IE} at biophysical values. When they are trainable, Adam does *not* preserve or recover effective PING from any tested initialisation (canonical, zero, or 10% of canonical): E→I recruitment weakens or remains absent, inhibitory firing collapses, and the network moves toward dense E firing at ≈ 90% accuracy, close to the frozen PING control (≈ 91%). The matrices store non-negative conductance magnitudes; inhibition is produced by the GABA reversal potential, not by a negative W^{IE}. PING is therefore a structural prior imposed by the frozen loop in this setup, not one gradient descent recovers on its own.

Method

Architecture. $N_E = 1024$ excitatory, $N_I = 256$ inhibitory, mem-mean readout, Dale's law enforced. Hyperparameters match the exp025 PING baseline: Adam at lr 4×10^{-4}, batch 256, surrogate slope 1, $W_{\text{in}} \sim \mathcal{N}(1.2, 0.12)$ at 95% sparsity, gradient norm clipped to 1.0, $\Delta t = 0.1$ ms, $T = 200$ ms, no firing-rate regulariser.

Both recurrent matrices are non-negative conductance magnitudes. After each optimiser step they are projected onto the non-negative cone. Their physiological sign is supplied by the pathway reversal potential: I→E contributes $g_I(E_I - V)$ with $E_I = -80$ mV and is therefore inhibitory despite $W^{IE} \geq 0$.

Sweep. Four conditions × three seeds (42, 43, 44) on the 10% MNIST subset (5600 train / 1400 test), 50 epochs. Only the initial (W^{EI}, W^{IE}) and the trainable-or-not flag vary:

Condition	W^{EI}, W^{IE} init	Trainable?
frozen_ping (control)	canonical biophysical	no
trainable_ping_init	canonical biophysical	**yes**
trainable_zero_init	0 (COBA-equivalent start)	**yes**
trainable_small_init	0.1 × canonical	**yes**

Canonical biophysical means $W^{EI} \sim \mathcal{N}(1.0, 0.1)$ µS, $W^{IE} \sim \mathcal{N}(2.0, 0.2)$ µS at $N_I = 256$, fan-in-normalised, so the trainer reports per-edge means of ≈ 0.0010 and ≈ 0.0078. "PING is on" means the inhibitory loop is active: E recruits I, and I paces a gamma rhythm through GABA conductance. "The loop is lost" means E→I recruitment is too weak to sustain that regime, I activity is low or absent, and E fires densely. The firing rates and rhythmicity are the cleanest read of which regime a trained network reaches.

Results

The whole sweep collapses to one picture. In the (E firing rate, I firing rate) plane, a network "found PING" if it sits in the low-E / high-I corner, and "lost the loop" if it sits in the dense-E / silent-I corner.

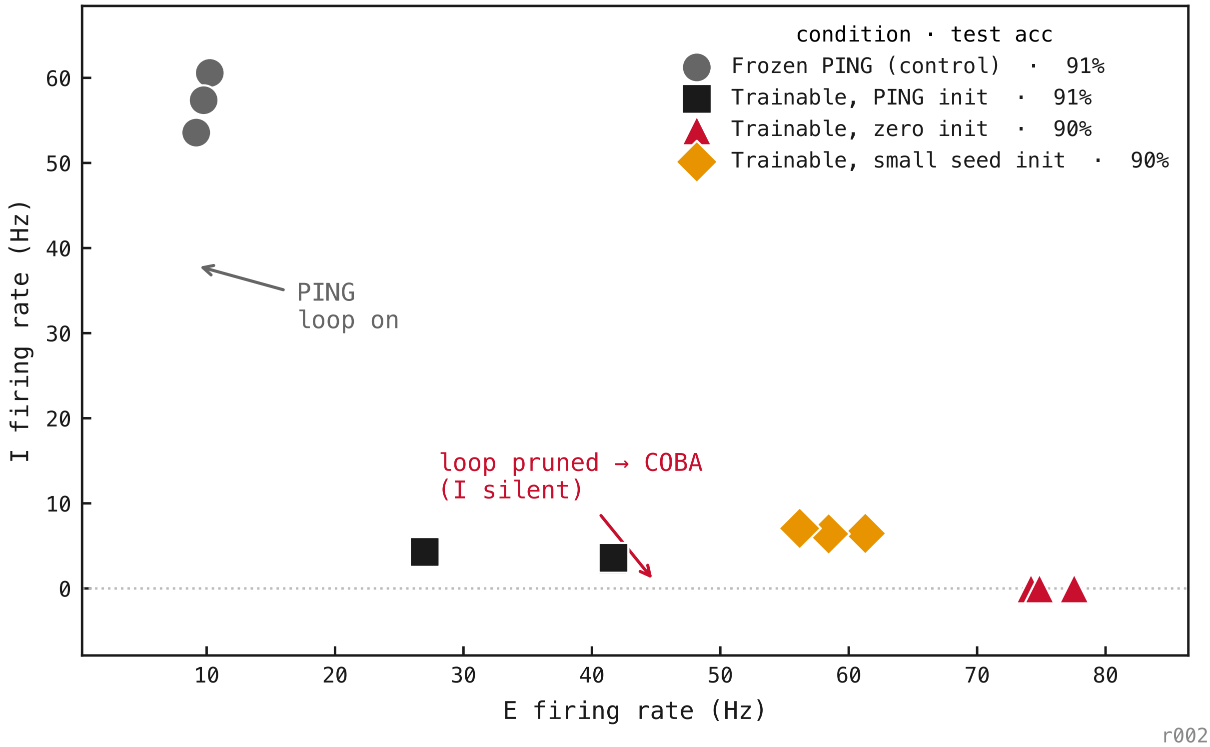

Figure 1: Each point is one trained network (4 conditions × 3 seeds). The **frozen control** (grey) sits in the PING corner, E ≈ 10 Hz and I ≈ 57 Hz, the inhibitory loop alive. **Every trainable condition** (started at canonical PING, at zero, or at 0.1× canonical) collapses to the dense-E / silent-I floor, E ≈ 40–75 Hz and I ≈ 0–7 Hz, at ≈ 90% accuracy, essentially matching the ≈ 91% control. Where the loop *starts* makes no difference; only whether it is allowed to train. Gradient descent never reaches PING, and dropping it costs no accuracy.

Training curves — functional loop collapse, epoch by epoch

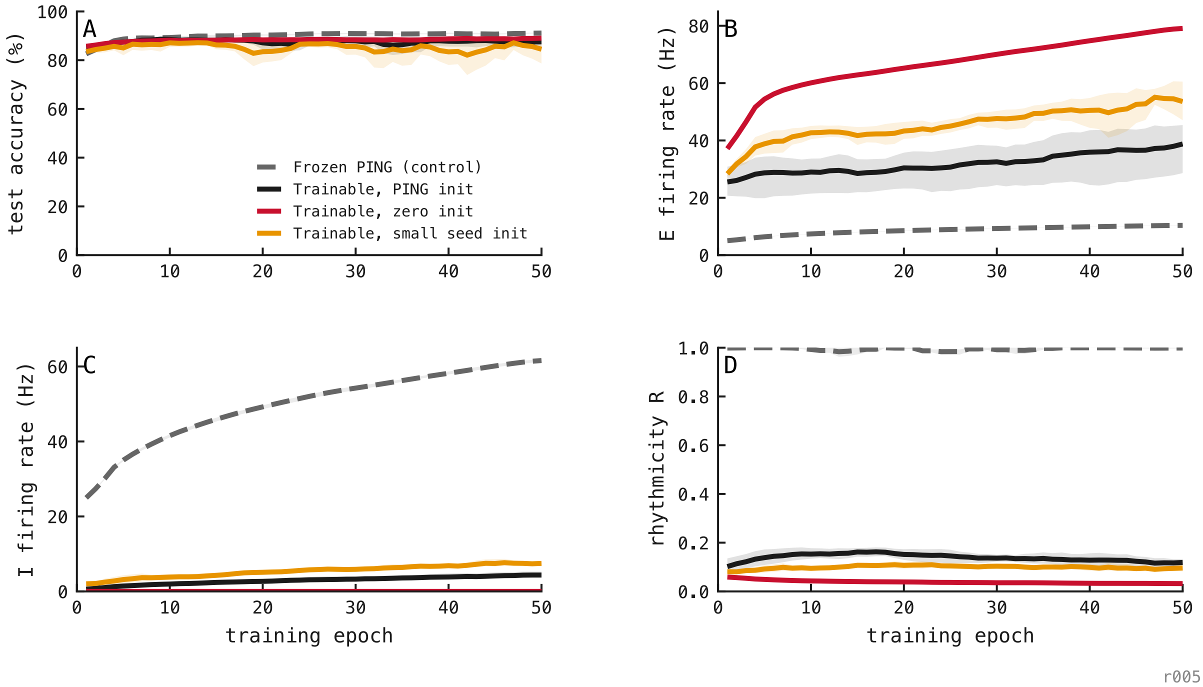

Figure 2: Per-epoch metrics from the runs themselves, coloured by init: PING init (black), zero init (red), small-seed init (amber), frozen-PING control (grey dashed); 10% of MNIST, 50 epochs, 3 seeds each. **Accuracy**: all reach ≈ 85–90% and track each other; losing effective PING costs nothing. **E rate**: the trainable runs climb to ≈ 37–75 Hz as inhibition releases the excitatory population, while the frozen control stays gated near 10 Hz. **I rate**: every trainable run drops to ≈ 0 within a few epochs as effective E→I recruitment weakens or remains absent, while the frozen control's I rate *rises* to ≈ 57 Hz as its readout trains. **Pingness**: the exp054 lobe–trough contrast holds near ≈ 0.99 for the frozen control while every trainable init collapses to ≈ 0.1–0.2. The residual floor is the metric's known low-rate inflation under shared input (exp054), not a surviving rhythm: the frozen-vs-trainable gap is the signal. Functional loop activity dies early, from every start, with no accuracy penalty.

The saved matrices identify which side of the recurrent loop changes. In the canonical-initialisation seed-42 checkpoint, 76.2% of E→I entries are zero after training, compared with 1.1% of I→E entries; their final means are 0.000451 and 0.008705, respectively. Thus the silent inhibitory population is associated chiefly with sparse E→I recruitment, while the I→E conductance remains predominantly positive.

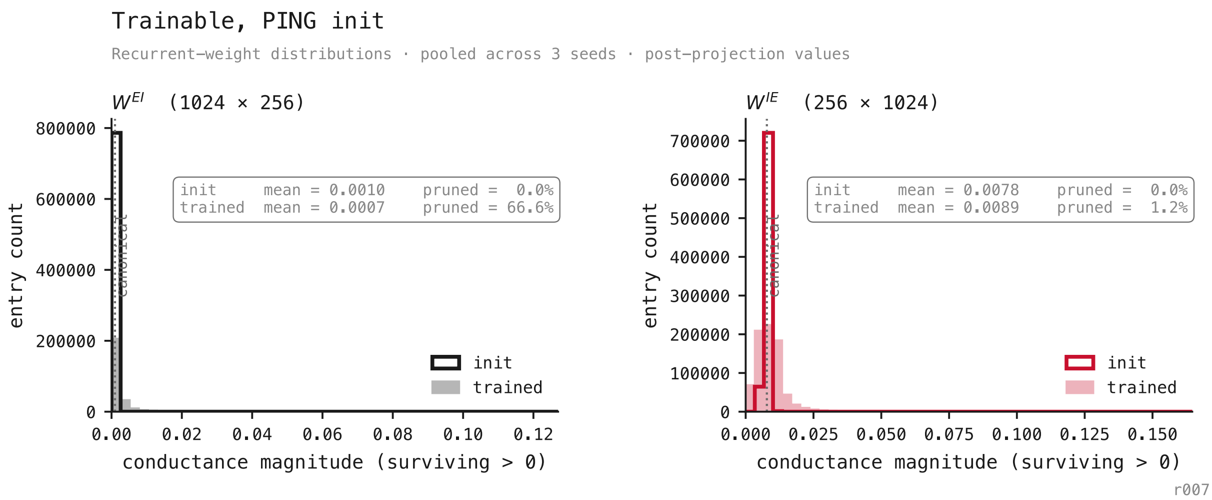

Figure 3: **The two recurrent matrices do not collapse symmetrically.** Initial and trained conductance distributions for the canonical-initialisation condition, pooled across seeds 42–44; each panel reports its own mean and zero fraction. The checkpoint-level distinction is already clear in seed 42: 76.2% of W^{EI} entries are zero after training, versus 1.1% of W^{IE} entries. The loss of PING therefore tracks weakened E→I recruitment and near-silent I activity, not I→E weights crossing into a negative sign cone.

Phase portrait — there is only one attractor under training

The four per-epoch panels above tell the story but separate the two state variables that matter. Putting *E rate* and *pingness* on perpendicular axes shows the trajectories of training directly, and the geometry rules out a misreading: it is *not* the case that PING and COBA are two attractors of the same dynamics, with the architecture deciding the basin. Under training there is **one** attractor (the COBA corner), and the PING state only persists because freezing the loop zeroes its gradients.

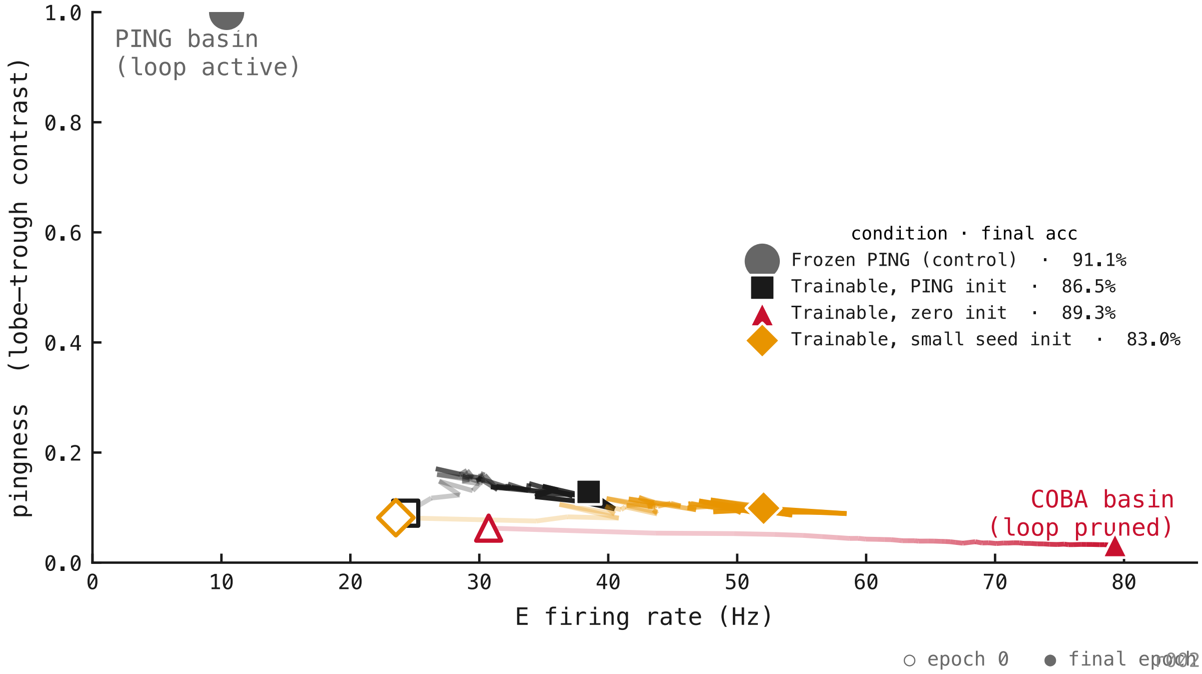

Figure 4: Each trainable trajectory is the per-epoch mean across 3 seeds, alpha-ramped along the epoch axis so the direction of flow reads off the line itself; open markers are epoch 1 (the first logged epoch), filled markers are epoch 50. **Frozen PING (grey)**: sits in the upper PING basin from start to finish at pingness ≈ 0.98; drifts a little in E rate (≈ 4 → 10 Hz) as the feedforward and readout weights train, but the loop weights themselves can't move by construction. **Trainable PING init (black)**: was initialised with canonical biophysical loop weights, but by the time the first metric is logged (after epoch 1) rhythmicity has already collapsed to ≈ 0.17. The trajectory then walks rightward along the COBA floor with the others. **Trainable zero init (red)** and **small-seed init (amber)**: start at the floor and stay there, E rate climbing as the feedforward layer learns the task. Every legend entry lands at 83–91% accuracy, so the move costs nothing. The reading is sharper than the time-series panels suggest: the empty band between pingness ≈ 0.2 and ≈ 0.85 is the *whole story*, because gradient descent flies through it within one epoch and the metric never catches it mid-flight. There is no separatrix between PING and COBA under training because there is no slow trajectory through the gap. The freeze isn't preventing slow erosion; it's preventing instant collapse.

Accuracy–rate trajectory — same destination, different spike economy

Figure 4 puts pingness on its own axis. Putting pingness on the *colour* axis instead and replaying training as (E rate, accuracy) trajectories (the exp025 accuracy–rate frontier given a time axis) gives another reading of the same data: every condition reaches the same final test accuracy, but along very different routes, and only one of them is doing PING while it gets there.

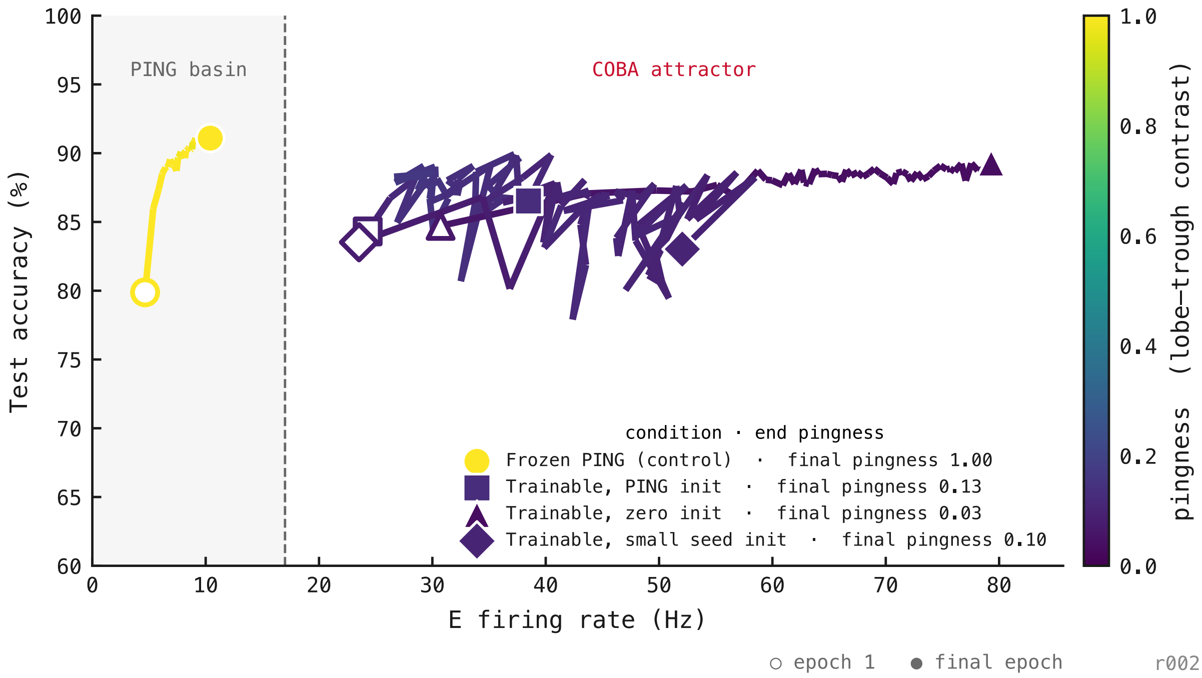

Figure 5: Each trajectory is the per-epoch mean across 3 seeds; per-segment colour is that epoch's pingness on a viridis 0→1 scale (colourbar at right). Open markers are epoch 1, filled markers are epoch 50. The frontier *with time* tells three things at once. **Same destination**: all four conditions land at ≈ 83–91% accuracy by epoch 50, regardless of init or whether the loop trains. **Different routes**: the frozen control climbs accuracy almost vertically with little change in E rate (≈ 4 → 10 Hz); the trainable conditions all bend right and up, the biggest one (zero init) reaching final accuracy at E ≈ 75 Hz, ≈ 8× the frozen control's rate. **Only one of them is still PING**: the frozen control's line is bright yellow because pingness stays high (≈ 1.0) the whole way; every trainable line is dark purple because rhythmicity had already collapsed to ≈ 0.1 by the first logged epoch. Reading the colour bar: most of pingness 0.2–0.85 is unused space, because no trajectory crosses it slowly enough to be logged. Reading the geometry: this is the manuscript's accuracy–rate frontier (exp025, §2.3) animated, and the architecture is the only thing keeping a trained network on the sparse, rhythmic side of it. The dashed divider at ≈ 17 Hz marks the two basins explicitly: frozen PING holds the left, every trainable run ends in the right, at the same accuracy.