

Perturbational equivalence: a testable criterion for replicating cognition.

V1.0 · OCTOBER 2026 · ARCAS DYNAMICS · HYPOTHESIS · NEUROSCIENCE + AI ARCHITECTURE

Brain models are judged by how closely they match the brain under ordinary conditions. Yet every trained network has a continuum of exact twins that match it on every observational metric and still respond differently when a part of them is silenced. We propose a necessary condition for replication: intervening on matched components must change the model's behaviour the way it changes the brain's, within the variability between individuals.

“The grain of the perturbation family sets the grain of the equivalence class.”

Hypothesis and synthesis paper. No new experimental data are reported; every model output is an estimate under illustrative parameters, not a finding.

01

The argument.

Why matching the brain under ordinary conditions cannot settle whether a model replicates it.

01 Observation is not mechanism

Distinct mechanisms can produce the same unperturbed activity. Among millions of three-cell circuit models, many different parameter sets gave similar output.

02 Every network has twins

Rotating a model's units leaves its representational geometry, its linear predictivity and its behaviour exactly unchanged.

03 Perturbation separates them

Silencing a local group of units does not survive the rotation, so local perturbations tell twins apart. A global change in gain does not.

04 A testable criterion

A model is perturbationally equivalent if intervening on matched components changes its behaviour as it changes the brain's, within the variation between individuals.

02

Live model.

A toy version of the paper's simulation: the same units in two layouts. Watch what each loses when a patch is silenced. The model runs in the browser at <https://arcasdynamics.com/research/perturbational-equivalence>.

Fig. 2a

Left, a topographic model in which units preferring the same category sit together. Right, its permuted twin: the same units and readout, shuffled. A silencing disk visits 16 sites. Switch to global gain to scale every response instead.

Fig. 2b

The paper's results. Every variant has the same RSA, 0.712, yet under local silencing the permuted and rotated twins exceed the tolerance. Under a global gain change, all four pass.

Status

Fig. 2a is a schematic. Fig. 2b shows the paper's model estimates under illustrative parameters, medians of 10 draws.

03

The equations.

01 EFFECT OF A PERTURBATION

$$\Delta_X(p) = b_X(p) - b_X(0)$$

The change in behaviour, across every task condition, when perturbation p is applied to system X .

02 BETWEEN-INDIVIDUAL VARIANCE

$$\sigma^2(p) = \frac{1}{n} \sum_i \|\Delta_i(p) - m_{-i}(p)\|^2$$

m_{-i} is the mean effect over every individual except i .

$$D_M = \frac{1}{|P_{\text{test}}|} \sum_p \frac{1}{n} \sum_i \frac{\|\Delta_M(p) - m_{-i}(p)\|^2}{\sigma^2(p)} \leq \tau$$

A typical individual scores about 1. The model passes if its distance is at most τ , the 95th percentile of the individual scores.

$$W'^T r' = W^T Q^T Q r = W^T r$$

$$Q^T E_U Q = E_U \Leftrightarrow \text{silencing } U \text{ is unchanged}$$

Rotating a model's units leaves its geometry, predictivity and behaviour unchanged, but not the effect of silencing them.

What it builds on.

The scale of today's brain models, and the limits of observing them, as the paper cites them.

139,255

neurons in the adult fly connectome, joined by 54.5 million synapses

DORKENWALD ET AL., 2024, NATURE · [HTTPS://DOI.ORG/10.1038/S41586-024-07558-Y](https://doi.org/10.1038/s41586-024-07558-y)

91%

of 164 testable predictions correct for a whole-brain model built on that connectome

SHIU ET AL., 2024, NATURE · [HTTPS://DOI.ORG/10.1038/S41586-024-07763-9](https://doi.org/10.1038/s41586-024-07763-9)

20 million+

three-cell circuit models, in which many parameter sets gave similar network activity

PRINZ ET AL., 2004, NATURE NEUROSCIENCE · [HTTPS://DOI.ORG/10.1038/NN1352](https://doi.org/10.1038/NN1352)

3,510

transistors lesioned one by one in a microprocessor, without revealing how the chip worked

JONAS & KORDING, 2017, PLOS COMPUTATIONAL BIOLOGY · [HTTPS://DOI.ORG/10.1371/JOURNAL.PCBI.1005268](https://doi.org/10.1371/journal.pcbi.1005268)

Predictions.

Each one names the result that would refute it.

P1 Ties come apart

Refuted if observational score, at its noise ceiling, accounts for all between-model variance in D.

STATE: UNTESTED

P2 Commitment matters

Refuted if permuted or rotated twins reach the same D on held-out sites.

STATE: UNTESTED

P3 Degeneracy surfaces

Refuted if the spread across matched models is no larger under perturbation than at baseline.

STATE: UNTESTED

P4 Families do not transfer

Refuted if rankings under input perturbations predict rankings under internal ones up to the noise ceiling.

STATE: UNTESTED

P5 D predicts new interventions

Refuted if selection by observational score predicts a held-out perturbation type equally well or better.

STATE: UNTESTED

06

How to test it.

First test: Primate IT reanalysis

Published muscimol inactivation data from inferior temporal cortex, compared against ventral-stream models across a range of observational scores, with each model's exact twins as negative controls. No new animals.

TESTS P1, P2, P4

Experiment 2: Connectome-constrained fly models

Silence and activate each identified cell type in silico, in a 50-model ensemble of the fly visual system and in a whole-brain model, and ask whether the spread across models grows under perturbation.

TESTS P3

Experiment 3: A held-out perturbation in mouse

Optogenetic silencing at one set of sites to fit models and excitation at another to test them, comparing models chosen by perturbational distance with models chosen by observational score.

TESTS P5

Where it could be wrong.

Observational fit may be enough in practice.

If high observational scores reliably come with a pass, the criterion adds cost without telling models apart. That is the refutation condition for P1, and we would accept it.

The mapping may be doing the work.

The criterion is only as strong as the rule that places model units on the brain, and different declared rules could give different verdicts.

Biological perturbations are imperfect.

Muscimol spreads by an uncertain amount, optogenetic effects can reach beyond their target, and circuits compensate.

Necessary, not sufficient.

Passing for one family of perturbations licenses no claim about another, and the criterion does not cover recovery after a lesion.

How to cite.

```
@misc{arcas2026perturbational,  
  author      = {Arcas Dynamics},  
  title       = {Perturbational equivalence: a testable criterion for replicating cognition},  
  year        = {2026},  
  month       = oct,  
  note        = {Hypothesis paper 02, version 1.0, CC BY 4.0},  
  howpublished = {\url{https://arcasdynamics.com/research/perturbational-equivalence}}  
}
```

© 2026 Arcas Dynamics. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0): <https://creativecommons.org/licenses/by/4.0/>

<https://arcasdynamics.com/research/perturbational-equivalence>

