RESPONSE CALIBRATION FOR NEURAL POPULATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models of neural dynamics predict activity well, but whether their learned operators recover the directed effects of perturbations is a separate question. We present PC-Flow, a perturbation-conditioned flow-matching model of population activity whose response operator is calibrated to randomized single-cell interventions. The operator is a finite-dose, stimulated-minus-blank mean contrast, equivalently a dose-averaged input Jacobian of the integrated flow. It estimates directed effective influence through recurrent and hidden pathways, rather than individual synaptic weights. In controlled recurrent circuits, calibration reduces response NRMSE from 0.317 to 0.120, compared with 0.190 for a calibrated Gaussian forecaster and 0.183 for a shared response-only network. Gains persist across the tested recurrence and observation conditions; a mechanism-matched recurrent model remains stronger when the full state is known. Matched objective controls locate the gain in response supervision, while joint predictive scores test the additional distributional information. PC-Flow also predicts held-out simultaneous perturbations: at eight sources, its full-response error is 0.049, versus 0.078 for calibrated Gaussian and 0.089 for exact-column superposition. Direct interaction-residual tests distinguish nonlinear prediction from improved single-source estimates. These results support response-calibrated generative models as estimators of directed perturbation effects; biological validation remains open.

1 INTRODUCTION

A neural population model is often used twice: first to forecast activity, then to predict what would change under a stimulus. These uses place different demands on its accuracy. Dominant population fluctuations can determine a forecast score while a smaller stimulus-induced change remains poorly estimated. For circuit analysis, that change is often the quantity of interest: which cells respond, how strongly, and on what timescale?

Virtual perturbations already read effective influence from learned dynamics (Luo et al., 2025). Recurrent-model gradients predict optogenetic responses (Sourmpis et al., 2026), and the effectome relates perturbations to propagation through a connectome (Pospisil et al., 2024). We build on this line of work by aligning a predictor's training target with a specified experimental contrast. The relevant distinction is between matching observed activity, matching a local derivative, and matching the effect of a finite pulse. These objectives coincide in special linear settings but can constrain different responses in nonlinear circuits.

Our target is a stimulus-minus-blank mean difference divided by the positive input dose. Source, pulse duration, observation channels, horizon and pretreatment context are explicit. Random assignment identifies the contrast on the experimental support; the response includes direct, indirect and recurrent paths. Figure 1 illustrates why a delayed response can connect cells without a direct synapse. It also separates a response mean from the conditional distribution of the stimulated activity.

Response calibration supervises this finite contrast while retaining a forecast objective. It can use repeated trials or randomized group moments from continuing stimulation. We compare it across model families instead of treating a flow architecture as essential. Perturbation-Calibrated Flow (PC-Flow) is the calibrated flow implementation; correlated Gaussian and recurrent forecasters receive

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.

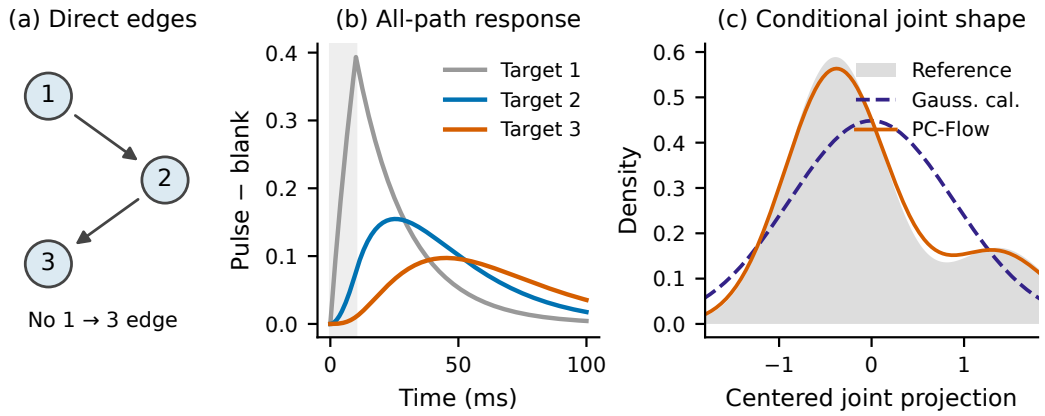

Figure 1: **Response means and predictive distributions carry different information.** (a,b) Stimulating cell 1 in a three-cell chain produces a delayed response in cell 3 without a direct edge. Shading marks the 10 ms pulse; curves are unit-amplitude linear responses (Appendix B.7). (c) Centered distributions of one population projection can differ even after their means agree. The reference, Gaussian and flow predictions use the same pretreatment context and stimulus.

the same response target. A shared response-only network and a mechanistic tanh rate model test whether representation sharing or a matched dynamical form explains the advantage.

The study addresses three linked questions. First, when does finite-response supervision improve an experimental readout beyond prediction loss or local-derivative supervision? Second, does a model recover state-dependent responses and useful joint predictive information, beyond a population-average contrast? Third, how do those gains change with repetition, partial observation, dose and simultaneous stimulation? The comparisons include evaluation-only noise references, circuit-level paired uncertainty and direct tests of interaction residuals. This focuses the contribution on experimentally specified training targets and the conditions under which they help, rather than on a new estimator of a squared mean.

2 Related Work

Perturbations and learned neural dynamics. Recurrence complicates connectivity inference from activity (Das & Fiete, 2020); targeted stimulation measures propagated influence (Daie et al., 2021; Randi et al., 2023). The effectome framework combines instrumental-variable identification with connectomic priors (Pospisil et al., 2024). Virtual perturbation of a fitted surrogate already underlies Neural Perturbational Inference (Luo et al., 2025). Biologically informed recurrent models predict optogenetic responses and micro-perturbation gradients, and show that unperturbed predictive fit need not determine perturbation fidelity (Sourmpis et al., 2026). CURBD estimates regional source currents from data-constrained recurrent models (Perich et al., 2026), while trial matching constrains distributions of neural activity (Sourmpis et al., 2023). These are direct precedents for learning dynamics and interrogating their responses. Our narrower contribution is an objective that aligns a specified finite positive-pulse readout with empirical randomized contrasts, evaluated against direct response models and calibrated recurrent and Gaussian predictors. It does not introduce virtual perturbation or infer anatomical adjacency.

Response supervision and predictive distributions. Sobolev training supplements function values with derivative targets (Czarnecki et al., 2017). For differentiable mean maps, a finite-dose contrast averages derivatives over an input interval: finite-pulse observations alone do not provide derivative ground truth at zero. Our calibration uses such observed contrasts, with experimental precision weights. The independent-half product is a classical construction for estimating squared expectations, related to order-two U-statistics and double sampling (Hoeffding, 1948; Baird, 1995); it removes computational bias rather than trial noise. The generative component builds on neural ODEs, flow matching, rectified flow and stochastic interpolants (Chen et al., 2018; Lipman et al.,

2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023; Tong et al., 2024). TSFlow already applies conditional flow matching to probabilistic forecasting (Kollovieh et al., 2025). In contrast to flows over connectivity distributions (Kim et al., 2026), our flow models future observations conditional on history and supplied input. Response calibration also applies to Gaussian and recurrent models; a flow advantage therefore requires evidence beyond a better mean contrast. CRPS, energy and variogram scores assess complementary forecast properties (Gneiting & Raftery, 2007; Scheuerer & Hamill, 2015).

Direct effect estimation and perturbation generalization. Input–output system identification provides excitation and finite-sample foundations (Willems et al., 2005; Oymak & Ozay, 2019); low-rank response estimation already supports active neural photostimulation (Wagenmaker et al., 2024). Heterogeneous-treatment-effect learners likewise predict contrasts directly from covariates (Künzel et al., 2019). Our direct-response and SVD-shrinkage baselines retain these alternatives. Sequential randomization identifies effects averaged over specified pretreatment groups (Boruvka et al., 2018), subject to consistency and treatment support (Hernán & Robins, 2020; Peters et al., 2017); it alone does not recover hidden-state-specific effects. Beyond neuroscience, CPA predicts cellular responses across perturbation contexts and doses, and GEARS predicts outcomes of unseen multigene perturbations (Lotfollahi et al., 2023; Roohani et al., 2024). Those transcriptomic tasks motivate testing perturbation generalization but have different observations and interventions. Here, dose-matched combined pulses test whether a fitted response extends beyond its calibrated single-source conditions; success is not a certificate of causal mechanism recovery.

3 Calibrating a model's experimental response

The target is a finite experimental contrast. Let H contain the 100 ms of recorded activity immediately before stimulation, and let $u \in \mathbb{R}_+^q$ give the amplitudes of q addressable input channels. A nonzero component delivers a 10 ms current pulse. We stack ten subsequent population observations into $Y \in \mathbb{R}^{10m}$, where m neurons are recorded. At physical horizon h, define

$$\begin{aligned} \mu_j(h, H, u) &= \mathbb{E}[y_j(h) \mid H, \mathrm{do}(u)], \\ D^{\delta}_{ji}(h, H) &= \frac{\mu_j(h, H, \delta e_i) - \mu_j(h, H, 0)}{\delta}. \end{aligned} \tag{1}$$

Each column measures the net response per unit input, including propagation through unobserved neurons and recurrent paths. Dose, pulse duration and readout are part of its definition. Randomized assignment identifies this contrast on the delivered treatment support. We evaluate its average across histories and, separately, its variation across observable states.

A shared target for different predictors. Write $G_\theta(\epsilon; H, u)$ for a conditional forecast sample, with base randomness independent of treatment. Its stimulated-minus-blank mean defines D^{δ}_{θ} by the same equation. Repeated randomized trials provide an empirical contrast $d^{\delta}_i = (\overline{Y}_{+,i} - \overline{Y}_0)/\delta$. Response calibration adds

$$\mathcal{L}_{\mathrm{cal}} = \mathbb{E}_{H,i,\delta} \frac{1}{10m} \sum_{\ell} w_{i\ell} \big(D^{\delta}_{\theta,\ell i}(H) - d^{\delta}_{\ell i}(H) \big)^2, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda \mathcal{L}_{\mathrm{cal}}. \tag{2}$$

Here ℓ indexes a neuron and future time. Losses use observations standardized on training data; reported responses undo that transformation. Independent pilot trials estimate contrast variances $\widehat{\nu}_{i\ell}$, pooled across histories. We use $w_{i\ell} \propto (\widehat{\nu}_{i\ell} + \eta)^{-1}$, with $\eta = \max(\mathrm{median}\, \widehat{\nu}, 10^{-4})$, and normalize weights to mean one per dose. These frozen weights reduce the influence of noisy coordinates without changing the history distribution. Appendix B.3 gives the variance and selection rules.

This objective applies to recurrent, Gaussian and flow forecasters. A tanh rate RNN tests a model with the generator's state-equation form; its weights and initial-state inference are learned from the same observed trials. A shared response-only network instead learns the contrasts directly. The generative comparison asks whether calibration can supply these responses while retaining a useful forecast distribution.

Notation and one calibration update. H: pretreatment history; d: fixed empirical contrast; δ: pulse amplitude; h: physical horizon; s: flow transport time; ϵ: model randomness.

1. Sample four history–source–dose groups, keeping their pilot weights fixed.
2. Draw eight independent bases. For each base, predict pulse and blank futures with the same base and form their difference divided by δ.
3. Average four contrasts in each independent half to obtain A, B. Average $w(A-d)(B-d)$ over output coordinates and sampled contexts. With an analytic conditional mean, use its squared residual directly.
4. Add the calibration loss to the ordinary prediction loss every fourth update, multiplying its λ weight by four.

For observable groups, each half independently samples histories and bases from the same fixed group pool, with the group-mass normalization in Appendix B.2; both subtract the same fixed group target. Analytic conditional means still require independent history batches or complete pool enumeration.

Table 1: **Response calibration uses paired inputs and independent residual means.** Pairing occurs within a model draw; independence occurs between the two means.

Calibrating the mean without an extra variance penalty. For one context and coordinate, form K paired differences $\Delta_k = [G_\theta(\epsilon_k; H, \delta e_i) - G_\theta(\epsilon_k; H, 0)]/\delta$. Sharing the base within a pair supplies a numerical coupling. Squaring their sample mean adds $\mathrm{Var}(\Delta_k)/K$ to the expected squared mean error. That extra term penalizes variation in the paired contrast, even when its mean is correct. Instead, split the independent pairs into two equal groups with means A and B:

$$\widehat{\ell}_{\mathrm{cal}} = w(A-d)(B-d), \qquad \mathbb{E}_\epsilon[\widehat{\ell}_{\mathrm{cal}} \mid d] = w(D_\theta^\delta - d)^2. \tag{3}$$

This is a classical unbiased construction for a squared expectation, closely related to order-two U-statistics (Hoeffding, 1948). The gradient is also unbiased when differentiation and expectation can be interchanged. This removes computational bias; empirical targets retain trial noise, and estimator variance need not decrease. Appendix A gives exact variances and an all-pairs alternative. We use eight pairs during calibration, analytic means when available, and 256 pairs for sampled response readout. Single-source and full-response comparisons share one endpoint cache.

The flow instantiation. Perturbation-Calibrated Flow (PC-Flow) uses a conditional flow over the complete future trajectory (Lipman et al., 2023; Tong et al., 2024). For independent $\epsilon \sim \mathcal{N}(0, I)$ and $s \sim \mathcal{U}[0, 1]$,

$$I_s = (1-s)\epsilon + sY, \qquad \mathcal{L}_{\mathrm{FM}} = \frac{1}{10m}\mathbb{E}\|v_\theta(I_s, s; H, u) - (Y - \epsilon)\|_2^2. \tag{4}$$

Generation integrates $\partial_s z_s = v_\theta(z_s, s; H, u)$ from $z_0 = \epsilon$. Transport time s differs from physical response time h, which indexes output coordinates. The uncalibrated CFM uses the identical network and prediction loss. A matched Gaussian transformer uses the same population backbone with a diagonal-plus-rank-eight covariance head; this separates the calibration objective, attention architecture and forecast family.

Calibration without restoring a simulator state. Continuing experiments seldom revisit an identical history. We therefore also match responses within fixed observable groups $g = G(H)$, defined from pretreatment activity. At eligible trial t, let $A_t = 0$ denote blank and $A_t = i$ denote a dose-δ pulse, with known positive assignment probabilities p_0, p_i. For a fixed pretreatment baseline $b(H_t)$, define

$$Z_{ti} = \frac{1}{\delta}\left[\frac{\mathbf{1}\{A_t = i\}}{p_i} - \frac{\mathbf{1}\{A_t = 0\}}{p_0}\right](Y_t - b(H_t)). \tag{5}$$

Sequential randomization makes its conditional expectation the current pulse-versus-blank response, averaging hidden states under the acquisition policy (Boruvka et al., 2018). The baseline cancels in expectation and can reduce noise. We minimize group residuals formed by averaging $D_{\theta,i}^\delta(H_t) - Z_{ti}$ over *all* eligible histories in each group. Pulse and blank predictions therefore use a common history distribution. Both Monte Carlo halves independently sample histories and base noise from that same fixed pool; reusing a random history batch would add a state-variation penalty. Appendix B.2 specifies the acquisition, normalization and weights. This experiment evaluates observable-group responses; a fully observed control separately tests individual-history responses.

Reading the directed response operator. The generated endpoint integrates the learned velocity over transport time. Its input Jacobian differentiates this complete trajectory, including the dependence of intermediate flow states on the intervention. Let P_h select the population readout at physical horizon h, and let $S_1^{(i)}(u) = \partial_{u_i} G_\theta(\epsilon; H, u)$. Under the differentiability and integrability conditions in Appendix A,

$$D^{\delta}_{\theta,:,i}(h, H) = \frac{1}{\delta} \int_0^{\delta} \mathbb{E}_\epsilon[P_h S_1^{(i)}(a e_i)]\, da. \tag{6}$$

Thus the experimental response has an equivalent dose-averaged Jacobian readout. Differentiating only the velocity's explicit input argument omits propagation through the generated trajectory. Supervising the derivative at zero, as in a local Sobolev objective, also targets a different quantity in a nonlinear system (Czarnecki et al., 2017); the oracle control evaluates both targets, while ordinary fitting receives no simulator derivatives.

Direction runs from a stimulated source to an observed response. Recurrent propagation can connect two cells without a direct edge, so this operator generally differs from synaptic adjacency. The linear pulse calculation in Appendix A makes that distinction explicit. For multi-source inputs, we compare full predictions with sums of *dose-matched* columns. They agree at one source; independent references and interaction-residual scores determine what the full model adds for larger sets.

4 Experiments

4.1 One response definition, complementary controls

The primary assay contains 24 independent excitatory–inhibitory rate circuits, eight at each recurrent gain $0.35, 0.60, 0.85$. Each has 80 neurons, 64 observed outputs and 20 addressable excitatory sources. Fitting uses blank and single-source 10 ms pulses at doses $0.1, 0.2, 0.4$; evaluation queries dose 0.2 at 10, 30, 60 and 100 ms. This weak-input assay is nearly additive: its main challenge is estimating small responses from noisy trials. Separate stronger circuits test state dependence and nonlinear combinations. Acquisition, model definitions and selection rules appear in Appendix B.

The controls distinguish calibration, architecture and response sharing. CFM versus PC-Flow isolates the objective; Gaussian transformer and GRU pairs repeat this comparison. ResponseNet shares the attention representation but learns responses directly. Pooled contrasts, SVD shrinkage and local response ridge test simpler alternatives. A latent tanh rate RNN, with and without calibration, matches the generator's state-equation form while learning its weights and hidden initial states. It receives no true connectivity or hidden-state labels.

Response NRMSE averages horizon-specific matrix errors after averaging over the same 48 held-out histories; self-responses are excluded. Forecast CRPS uses four separate held-out histories per circuit, with blank and two source conditions at all 640 future coordinates. Three fits average within a circuit before 2,000 paired circuit bootstrap resamples. Evaluation-only repeated contrasts and an approximate oracle forecast quantify reference difficulty (Table 11). An eight-repeat response estimate at the test histories has error 0.328; infinite repeats at those same histories give zero by definition. This privileged measurement is a noise-budget reference, not an additional trained baseline.

4.2 Calibrate the response, then assess the model family

PC-Flow reduces response error from 0.317 to 0.120 (62.2%); its paired difference from CFM is -0.197 [-0.222, -0.174]. Calibrated Gaussian reaches 0.190, and GRU reaches 0.194. The mechanism-matched rate RNN improves from 0.248 to 0.164. Calibration therefore benefits several predictive families, while PC-Flow retains a measurable advantage in this partially observed assay. ResponseNet reaches 0.183; the paired PC-Flow difference is -0.064 [-0.084, -0.046]. SVD shrinkage reaches 0.314, selecting full rank in all primary circuits.

The repeat curve supplies a separate budget comparison at fixed $\lambda = 0.3$; low-repeat performance cannot be attributed to selecting calibration off. At two fitting repeats per condition, PC-Flow reaches 0.230; at 32 it reaches 0.078. More repetitions also improve direct response learning. These curves support a benefit under limited response supervision, but do not isolate whether forecasting,

Table 2: **Response calibration across predictive families.** Primary assay; brackets give 95% circuit bootstrap intervals. Parameters are in millions. CRPS evaluates four held-out contexts per circuit; direct response estimators do not define predictive distributions.

Model	Params. (M)	Response ↓	CRPS ↓
Pooled contrast	–	0.330 [0.301, 0.362]	–
SVD shrinkage	–	0.314 [0.290, 0.339]	–
Response network	0.848	0.183 [0.163, 0.206]	–
GRU + cal.	0.198	0.194 [0.174, 0.216]	0.131 [0.129, 0.133]
Gaussian TF + cal.	0.849	0.190 [0.173, 0.207]	0.128 [0.127, 0.130]
Rate RNN	0.109	0.248 [0.225, 0.273]	0.125 [0.124, 0.127]
Rate RNN + cal.	0.109	0.164 [0.148, 0.179]	0.125 [0.124, 0.127]
CFM	0.848	0.317 [0.295, 0.341]	0.128 [0.126, 0.130]
PC-Flow	0.848	0.120 [0.109, 0.133]	0.128 [0.126, 0.129]

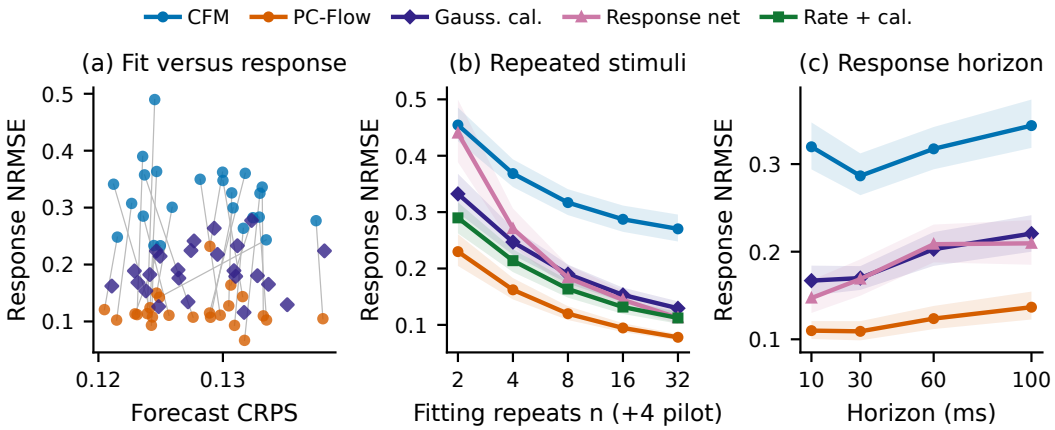

Figure 2: **Prediction accuracy, response accuracy and trial budget.** (a) Circuit-level forecast and response scores. (b) Nested fitting repetitions per condition, with four separate pilot repeats. (c) Response error at each horizon. Bars and bands resample circuits; short-lag errors remain visible rather than being hidden by longer-lag responses.

regularization or optimization causes that benefit. Figure 2 reports horizon-specific results alongside the information-budget curve.

The paired CRPS difference from CFM is -0.0005 [-0.0015, 0.0004]. PC-Flow's excess above the approximate oracle is 0.003, so forecast quality is not judged only by near-equal absolute scores. Neither a small CRPS change nor a small response error establishes the other.

Objective controls belong to this mean-response comparison. At the same eight-pair budget, naive-square minus split-half error is 0.017 [0.013, 0.022]. The all-pairs U-statistic gives difference 0.0000 [-0.0010, 0.0013] from split halves. Both unbiased estimators target the same mean loss; split halves are an implementation choice, not a new statistical principle. Their variance comparison depends on residual size (Appendix A). Weighting, source-label, checkpoint, calibration-strength and independently varied training/readout sample counts appear in Appendix B.8. An equal-time CFM control reaches 0.264 [0.244, 0.285]; this comparison does not rank model families under equal compute.

Table 3 makes the objective comparison explicit. Shuffling source labels degrades response estimates even though the forecast trials are unchanged. Uniform weights remain competitive, so the gain does not depend on the particular variance weights. Changing the validation criterion improves uncalibrated predictors modestly; it does not reproduce the improvement from supervising the intervention contrast.

Table 3: **Which changes improve the intervention readout?** Objective, weighting and checkpoint controls use the same primary trials. Brackets give 95% circuit bootstrap intervals. Default PC-Flow has error 0.120; the equal-time row gives CFM more optimization without extra trials.

Setting	Response NRMSE ↓
PC-Flow, CRPS-only checkpoint	0.146 [0.133, 0.162]
Naive squared contrast	0.137 [0.125, 0.151]
Shuffled source labels	2.844 [2.413, 3.302]
Uniform contrast weights	0.151 [0.142, 0.162]
CFM, CRPS only	0.335 [0.311, 0.361]
CFM, CRPS + 0.1 response	0.317 [0.295, 0.341]
GaussianTF, CRPS only	0.302 [0.274, 0.332]
GaussianTF, CRPS + 0.1 response	0.285 [0.257, 0.315]
CFM, equal training time	0.264 [0.244, 0.285]
All-pairs U-statistic, 8 pairs	0.120 [0.108, 0.134]

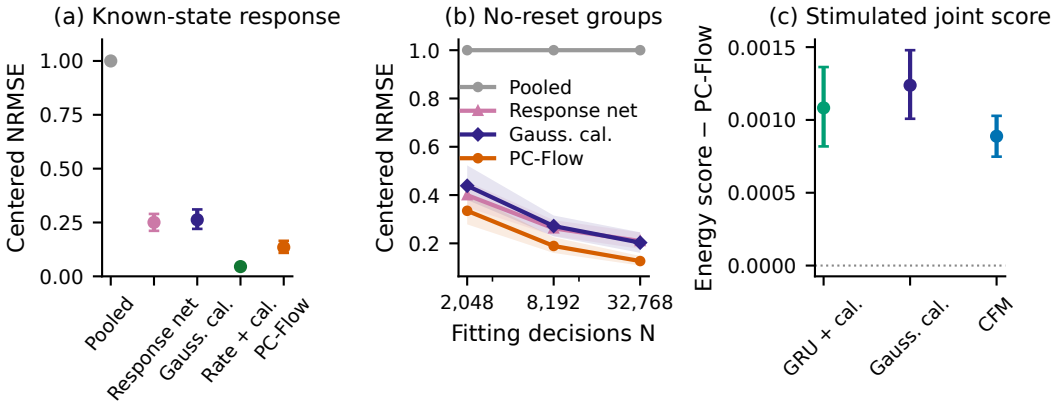

Figure 3: **Conditioning and joint forecasting answer different questions.** (a) Centered response error at known states. (b) Centered group-response error during continuing randomized acquisition; the pooled reference is one. (c) Stimulated energy-score differences relative to PC-Flow on a fixed 16-dimensional projection; positive values favor PC-Flow. These assays share twelve circuit draws and are not independent cohorts.

4.3 Test conditioning before averaging it away

Twelve stronger recurrent circuits, accepted by an equilibrium-stability criterion at gain 1.3, support complementary conditioning tests. In a fully observed noiseless control, preparatory currents produce different initial states, models are refitted, and the final pretreatment observation reveals the complete state. Centered response error removes each model's own mean across states; a state-independent predictor has error exactly one. PC-Flow reaches raw error 0.099 and centered error 0.135. The calibrated rate RNN reaches centered error 0.045, outperforming PC-Flow and showing the value of the known dynamical form when hidden-state uncertainty is removed.

A separate oracle-label control on these states compares local-derivative supervision with finite-dose supervision on the same flow backbone (Table 12). At dose four, finite-response error is 0.195 under local-derivative supervision and 0.038 under finite-contrast supervision. The control measures both targets rather than treating the derivative at zero as a finite pulse. The integrated-derivative identity in Equation 6 remains exactly the finite contrast; it is not a distinct prediction rule.

The same twelve circuits also support continuing randomized stimulation with no training-state resets. Pretreatment PC1 tertiles define fixed observable groups; logged assignment probabilities identify their pulse-minus-blank moments. At 32,768 fitting decisions, PC-Flow reaches raw group error 0.063 and centered error 0.127; calibrated Gaussian reaches centered error 0.203. Figure 3(b) reports all three acquisition budgets. Here the estimand averages histories and hidden states within a group under the acquisition policy, whereas the noiseless control tests individual states.

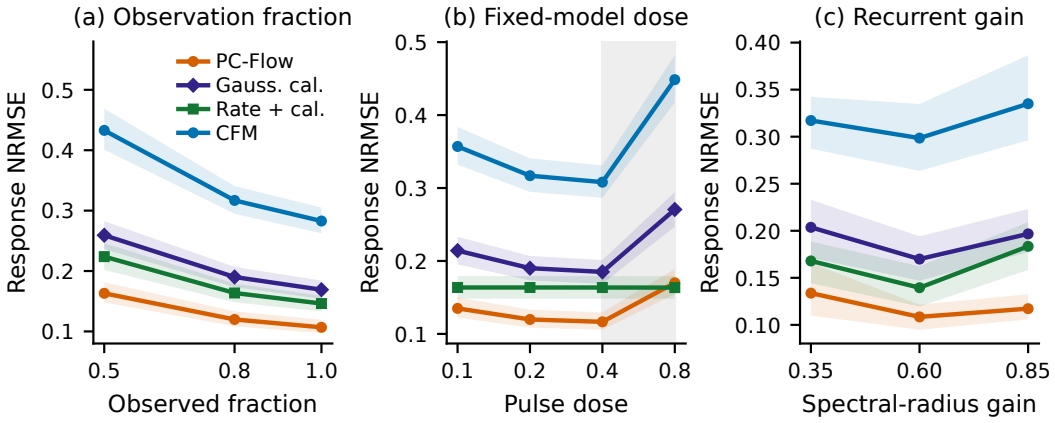

Figure 4: **Report operating limits explicitly.** Response error against (a) observed fraction, (b) pulse dose and (c) recurrent gain. Only dose 0.8 lies outside fitting support. Observation subsets retain all source cells. Gain panels contain eight independent circuits each; other panels use 24. Table 8 supplies numerical intervals.

4.4 A FORECAST PROVIDES INFORMATION BEYOND MEAN RESPONSES

Conditional forecast samples are scored at three fixed histories per continuing-assay circuit, keeping blank and stimulation separate. Energy and variogram scores evaluate eight unstimulated cells at two horizons, retaining population and temporal dependence (Gneiting & Raftery, 2007; Scheuerer & Hamill, 2015). Stimulated energy scores are 0.1625 for PC-Flow and 0.1638 for calibrated Gaussian; their paired Gaussian-minus-PC difference is 0.0012 [0.0010, 0.0015]. A Gaussian with reference-matched moments and complementary variogram scores bound the interpretation (Table 7). The result concerns this projection, not recovery of the complete population law. A direct response estimator can solve the mean task without supplying this distributional information.

4.5 VISIBILITY, DOSE AND CIRCUIT STRUCTURE EXPOSE FAILURE MODES

Figure 4 reports every observation fraction and dose. With 40 of 80 cells observed, PC-Flow reaches 0.163 versus 0.259 for calibrated Gaussian. At dose 0.8, outside the fitting support, their errors are 0.170 and 0.271; the mechanistic rate RNN remains competitive at 0.164. The observation-0.8 and dose-0.2 anchors reproduce the primary result exactly. Observation tests refit each model; dose tests keep the fitted model fixed.

Calibration helps across the three recurrent gains. On pairs with no direct edge but a directed indirect path, PC-Flow reaches 0.587 versus 1.554 for CFM. Direct-support and excitatory/inhibitory-target breakdowns use separate reference norms (Table 9). These are response audits against known anatomy, not connectivity-recovery claims. Separately refitted spiking models reach 0.255 for PC-Flow and 0.412 for CFM; VARX remains strong in the well-specified linear control at 0.100 (Table 5).

4.6 COMBINATIONS REQUIRE BOTH ACCURATE COLUMNS AND INTERACTIONS

The stronger cohort uses single-source training doses $0.2, 0.8, 4$ and tests unseen source sets at matched doses. At dose four and eight sources, true-column superposition has error 0.089. PC-Flow's estimated columns give 0.107, whereas its full response gives 0.049; calibrated Gaussian's full response gives 0.078. Thus nonlinear prediction improves on an exact-column additive reference within this assay.

Full-response accuracy alone can hide cancellation between column and interaction errors. We therefore score $\|\widehat{q} - q\|/\|q\|$ directly, normalizing masked errors within each source set and horizon, then averaging equally within each circuit. At eight sources, PC-Flow gives 0.058 and calibrated Gaussian gives 0.150. Smaller sets have weaker interactions and can have larger relative errors;

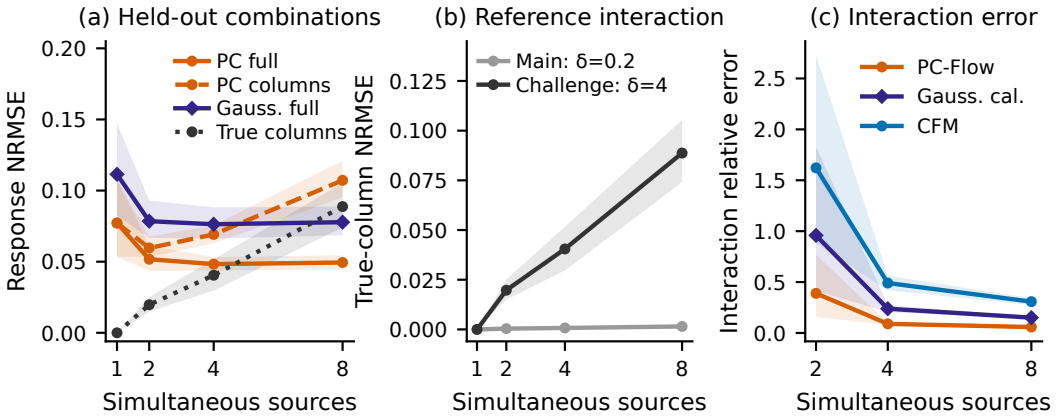

Figure 5: **Measure the interaction rather than infer it from full prediction.** (a) Dose-four full and column-sum errors, including calibrated Gaussian. (b) True nonadditivity in weak and strong assays. (c) Relative error of the predicted interaction residual for two, four and eight sources. Bands resample the same twelve stronger circuits. Table 10 gives numerical intervals.

Figure 5(c) displays that difficulty. At one source both residuals are exactly zero, so their relative error is undefined, while full and column predictions agree exactly.

Public-recording check. A bounded retrospective analysis uses twelve public recordings from Randi et al. (2023), held out by acquisition date. On supported source–target pairs, group-response RMSE is 0.10157 for a regularized history predictor and 0.10171 after adding group-mean calibration; zero change gives 0.10175. This small check supplies no improvement and does not validate PC-Flow. Missing matched blanks and limited cross-date coverage prevent identifying the manuscript's causal contrast (Appendix C).

5 Discussion

Response calibration makes the intervention query an explicit training target. Its benefits across flow, Gaussian, recurrent and mechanistic predictors support that design choice; differences among calibrated families determine which model is useful for a given assay. The known-state rate RNN control shows why a suitable dynamical form remains valuable. Predictive distributions supply an additional capability whose value must be assessed with joint scores, rather than inferred from mean-response accuracy.

Observable-group effects offer a practical target under continuing randomization, while pointwise effects require stronger information about state. Likewise, single-source accuracy does not identify arbitrary multi-source interactions. Evaluating the interaction residual directly prevents improvements in column error from being mistaken for a learned interaction law.

The present dimensions and timings also matter: neuron attention scales quadratically in observed population size, and sampled flow contrasts are expensive. Gaussian analytic means and direct-response models offer much cheaper readouts (Appendix B.4). A 32-pair PC-Flow contrast takes 21.84 s per context and source on the profiled CPU, versus 0.027 s for an analytic Gaussian contrast. These microbenchmarks quantify readout cost, not end-to-end training. The results do not establish equal-compute superiority across model families. The bounded public-recording check gives no positive calibration evidence and leaves optical artifacts, targeting uncertainty and causal identification unresolved. Biological validation requires a suitable acquisition protocol and substantially broader recording coverage.

AI USE STATEMENT

Large language models assisted with drafting and revising the manuscript, including its organization, language, and presentation of technical arguments. They also supported editorial consistency checks across the text, equations, tables, and figure descriptions. Responsibility for the scientific claims, interpretation, and final wording rests with the authors.

REPRODUCIBILITY STATEMENT

Appendix B specifies circuit and intervention parameters, observation models, data splits, predictor architectures, optimization, and model-selection rules. Appendix A states the assumptions behind the response identities and calibration estimators; Appendix B.5 defines scores, aggregation, uncertainty, and numerical tolerances. Appendix C identifies the public recordings and details inclusion criteria, preprocessing, and held-out evaluation.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pp. 30–37. Morgan Kaufmann, 1995. doi: 10.1016/B978-1-55860-377-6.50013-X. URL https://leemon.com/papers/1995b.pdf.

Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A. Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113 (523):1112–1121, 2018. doi: 10.1080/01621459.2017.1305274.

Nicolas Brunel. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *Journal of Computational Neuroscience*, 8(3):183–208, 2000. doi: 10.1023/A:1008925309027.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/758a06618c69880a6cee5314ee42d52f-Abstract.html.

Kayvon Daie, Karel Svoboda, and Shaul Druckmann. Targeted photostimulation uncovers circuit motifs supporting short-term memory. *Nature Neuroscience*, 24(2):259–265, 2021. doi: 10.1038/s41593-020-00776-3.

Abhranil Das and Ila R. Fiete. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nature Neuroscience*, 23(10):1286–1296, 2020. doi: 10.1038/s41593-020-0699-2.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. doi: 10.1002/qj.2270.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. URL https://miguelhernan.org/whatifbook.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Timothy Doyeon Kim, Ulises Pereira-Obilinovic, Yiliu Wang, Eric Shea-Brown, and Uygar Sümbül. Identifying connectivity distributions from neural dynamics using flows. *arXiv preprint arXiv:2603.26506*, 2026. URL https://arxiv.org/abs/2603.26506v2. Version 2, 29 May 2026.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with Gaussian process priors for probabilistic time series forecasting. In *International Conference on Learning Representations*, pp. 95435–95457, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ee1a1ecc92f35702b5c29dad3dc909ea-Abstract-Conference.html.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Jay Shendure, Jose L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

Zixiang Luo, Kaining Peng, Zhichao Liang, Shengyuan Cai, Chenyu Xu, Dan Li, Yu Hu, Changsong Zhou, and Quanying Liu. Mapping effective connectivity by virtually perturbing a surrogate brain. *Nature Methods*, 22(6):1376–1385, 2025. doi: 10.1038/s41592-025-02654-x.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of LTI systems from a single trajectory. In *2019 American Control Conference (ACC)*, pp. 5655–5661, 2019. doi: 10.23919/ACC.2019.8814438.

Matthew G. Perich, Charlotte Arlt, Sofia Soares, Siyan Zhou, Manuel Beiran, Aaron S. Andalman, Tyler Benster, Megan E. Young, Clayton P. Mosher, Juri Minxha, Eugene Carter, Ueli Rutishauser, Peter H. Rudebeck, Christopher D. Harvey, Karl Deisseroth, and Kanaka Rajan. Inferring brain-wide interactions using data-constrained recurrent neural network models. *Neuron*, 2026. doi: 10.1016/j.neuron.2026.07.016. Advance online publication, 7 August 2026.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017. URL https://mitpress.mit.edu/9780262037310/elements-of-causal-inference/.

Dean A. Pospisil, Max J. Aragon, Sven Dorkenwald, Arie Matsliah, Amy R. Sterling, Philipp Schlegel, Szi-chieh Yu, Claire E. McKellar, Marta Costa, Katharina Eichler, Gregory S. X. E. Jefferis, Mala Murthy, and Jonathan W. Pillow. The fly connectome reveals a path to the effectome. *Nature*, 634(8032):201–209, 2024. doi: 10.1038/s41586-024-07982-0.

Francesco Randi, Anuj K. Sharma, Sophie Dvali, and Andrew M. Leifer. Neural signal propagation atlas of *Caenorhabditis elegans*. *Nature*, 623(7986):406–414, 2023. doi: 10.1038/s41586-023-06683-4.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. doi: 10.1175/MWR-D-14-00269.1.

Christos Sourmpis, Carl C. H. Petersen, Wulfram Gerstner, and Guillaume Bellec. Trial matching: capturing variability with data-constrained spiking neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74787–74798, 2023. doi: 10.52202/075280-3269.

Christos Sourmpis, Carl C. H. Petersen, Wulfram Gerstner, and Guillaume Bellec. Biologically informed cortical models predict optogenetic perturbations. *eLife*, 14:RP106827, 2026. doi: 10.7554/eLife.106827.3.

Marcel Stimberg, Romain Brette, and Dan F. M. Goodman. Brian 2, an intuitive and efficient neural simulator. *eLife*, 8:e47314, 2019. doi: 10.7554/eLife.47314.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Andrew Wagenmaker, Lu Mi, Marton Rozsa, Matthew S. Bull, Karel Svoboda, Kayvon Daie, Matthew D. Golub, and Kevin Jamieson. Active learning of neural population dynamics using two-photon holographic optogenetics. In *Advances in Neural Information Processing Systems*, volume 37, pp. 31659–31687, 2024. doi: 10.52202/079017-0994.

Jan C. Willems, Paolo Rapisarda, Ivan Markovsky, and Bart L. M. De Moor. A note on persistency of excitation. *Systems & Control Letters*, 54(4):325–329, 2005. doi: 10.1016/j.sysconle.2004.09.003.

A RESPONSE IDENTITIES AND NUMERICAL CHECKS

We suppress fixed history and readout indices. Conditional means average hidden states given recorded history; repeated trials from one full simulator snapshot instead estimate that snapshot's response. Their aggregation uses the history distribution in Appendix B.

Finite differences and flow sensitivity. Assume base noise independent of u, a continuously differentiable velocity, a unique flow solution on $[0, 1]$, and an integrable bound on endpoint input derivatives along $u = ae_i$, $0 \le a \le \delta$. For $S_s^{(i)} = \partial_{u_i} z_s$, differentiating the ODE gives

$$\partial_s S_s^{(i)} = \partial_z v_\theta S_s^{(i)} + \partial_{u_i} v_\theta, \qquad S_0^{(i)} = 0. \tag{7}$$

Applying the fundamental theorem of calculus to the physical-time projection P_h and interchanging expectation and dose integration yields Equation 6. This equates two readouts of the same model; randomization supplies the target's experimental interpretation. Endpoint pairing is a numerical coupling, not an individual physiological counterfactual.

A local derivative $J_i(H) = \partial_a \mu(H, ae_i)|_{a=0}$ differs from $D_i^\delta(H)$: for $\mu(a) = \alpha a + \beta a^2$, they equal α and $\alpha + \beta\delta$. Dose integration changes the computation, not the finite-response objective. The oracle control in Appendix B.6 compares these different labels; it tests target alignment, not superiority under identical experimental acquisition.

Independent halves for mean calibration. For a fixed empirical target d, independent model means A, B satisfy $\mathbb{E}[(A - d)(B - d)] = (\mathbb{E}A - d)^2$ when both have the same expectation. For a differentiable sampler with an integrable bound on the loss gradient in a parameter neighborhood, differentiation and expectation can be interchanged, giving an unbiased gradient. Squaring a single finite-sample mean instead adds its variance. The paired-contrast variance can change through the coupling or through either conditional distribution: the extra penalty need not reduce both marginal variances.

More precisely, let $K \ge 2$ be even, Δ_k independent and identically distributed paired contrasts, $r = \mathbb{E}\Delta_k - d$, $\sigma^2 = \operatorname{Var}\Delta_k$, and $\mu_j = \mathbb{E}[(\Delta_k - \mathbb{E}\Delta_k)^j]$. For unweighted losses, define $T_{\text{split}} = (A - d)(B - d)$ and $T_{\text{sq}} = (\bar{\Delta} - d)^2$. With finite fourth moment,

$$\begin{aligned} \mathbb{E}T_{\text{split}} &= r^2, & \operatorname{Var}(T_{\text{split}}) &= \frac{4r^2\sigma^2}{K} + \frac{4\sigma^4}{K^2}, \\ \mathbb{E}T_{\text{sq}} &= r^2 + \frac{\sigma^2}{K}, & \operatorname{Var}(T_{\text{sq}}) &= \frac{4r^2\sigma^2}{K} + \frac{4r\mu_3 + 2\sigma^4}{K^2} + \frac{\mu_4 - 3\sigma^4}{K^3}. \end{aligned} \tag{8}$$

Each half-mean has variance $2\sigma^2/K$; expanding their independent product gives the first variance. For the second, expand $(r + \bar{X})^2$, using $\mathbb{E}\bar{X}^3 = \mu_3/K^2$ and $\mathbb{E}\bar{X}^4 = 3\sigma^4/K^2 + (\mu_4 - 3\sigma^4)/K^3$, where $X = \Delta - \mathbb{E}\Delta$. A fixed weight multiplies these variances by w^2. Their leading K^{-1} terms coincide for fixed nonzero r. At $r = 0$ with Gaussian contrasts, the split estimator instead has exactly twice the variance of the squared estimator. Unbiasedness does not guarantee smaller variance or finite-sample mean squared error.

The classical all-pairs U-statistic (Hoeffding, 1948), with $R_k = \Delta_k - d$, is

$$T_U = \frac{(\sum_k R_k)^2 - \sum_k R_k^2}{K(K-1)}, \qquad \mathbb{E}T_U = r^2, \qquad \operatorname{Var}(T_U) = \frac{4r^2\sigma^2}{K} + \frac{2\sigma^4}{K(K-1)}. \tag{9}$$

Separate products with disjoint indices from those sharing one index to obtain the variance. Averaging over balanced splits yields T_U, removing partition randomness without additional endpoints in $O(K)$ operations. Both variances require only a finite second moment; neither estimator removes target noise.

For observable groups, each half samples histories and model noise independently from the same fixed empirical pool and subtracts the same target. Reusing a random history batch adds covariance from variation of its model mean, even with independent noise draws. Complete pool enumeration also avoids this term. These statements are conditional on the data, not finite-data generalization guarantees.

Randomization in a continuing experiment. Condition on pretreatment history and any earlier assignments. Under the positive logged assignment probabilities in Eq. (5),

$$\mathbb{E}[Z_{ti} \mid H_t] = \frac{\mathbb{E}[Y_t \mid H_t, A_t = i] - \mathbb{E}[Y_t \mid H_t, A_t = 0]}{\delta} = D_i^{\delta}(H_t). \tag{10}$$

The baseline cancels. Multiplying by $\mathbf{1}\{G(H_t) = g\}$ and averaging identifies the group moment without requiring independent adjacent trials or hidden-state recovery, under the pretreatment eligibility, randomization and outcome-window conditions in Appendix B.2.

A finite-width linear pulse. Consider $dX_r = (AX_r + Bu\mathbf{1}_{[0,d]}(r))\,dr + \Sigma\,dW_r$ and $y(r) = CX_r + \eta_r$, with a treatment-independent initial-state distribution and zero-mean observation noise. Variation of constants cancels the common initial-state and noise means, yielding

$$D^{\delta}(h) = C \int_0^{\min(h,d)} e^{A(h-r)} B\,dr. \tag{11}$$

This finite-horizon identity does not require stability or an invertible A. For $C = I$, $B = I/\tau$ and $h < d$,

$$D^{\delta}(h) = \frac{h}{\tau} I + \frac{h^2}{2\tau} A + E_h, \qquad \|E_h\|_2 \le \frac{h^3 \|A\|_2^2 e^{h\|A\|_2}}{6\tau}. \tag{12}$$

The bound follows by integrating $\|e^{At} - I - tA\|_2 \le t^2 \|A\|_2^2 e^{t\|A\|_2}/2$. With $A = (-I + W)/\tau$, each off-diagonal entry consequently has leading term $W_{ji}h^2/(2\tau^2)$. Filtering, binning and partial observation apply additional observation operators; the short-lag coefficient is specific to the stated fully observed linear model. Appendix B.7 checks the finite pulse against independent numerical integration.

What a composition improvement measures. For a set S at dose a, write $r = \mu(a\mathbf{1}_S) - \mu(0)$, $r_{\mathrm{col}} = a\sum_{i\in S} D^a_{:,i}$, and $q = r - r_{\mathrm{col}}$. Define their model counterparts using the same endpoint cache. Then

$$\widehat{r} - r = (\widehat{r}_{\mathrm{col}} - r_{\mathrm{col}}) + (\widehat{q} - q). \tag{13}$$

Thus full-response accuracy depends on both column estimation and the model's departure from additivity. More explicitly, with $e_{\mathrm{col}} = \widehat{r}_{\mathrm{col}} - r$,

$$\|\widehat{r} - r\|_2^2 - \|e_{\mathrm{col}}\|_2^2 = 2e_{\mathrm{col}}^{\mathsf{T}}\widehat{q} + \|\widehat{q}\|_2^2. \tag{14}$$

A learned correction can help by canceling column error, even if the true q is small. This is why the experiments report the reference superposition residual separately. At $|S| = 1$, both q and $\widehat{q}$ are exactly zero. Single-source data alone cannot determine arbitrary interactions: adding $c\,u_1u_2v(H)$ to a mean model leaves every single-source response unchanged. The stronger-coupling assay tests a specified generalization regime rather than inferring an unrestricted interaction law.

B Experimental protocol and implementation details

B.1 Rate circuits and repeated-snapshot acquisition

Each of 24 independent circuits has $N = 80$ rate-deviation variables, 64 excitatory and 16 inhibitory presynaptic types, and off-diagonal connection probability 0.1. In the convention $W_{ji} : i \to j$, Dale's sign is attached to a column. Nonzero unsigned weights are $\max(Z, 0.05)$, $Z \sim \mathcal{N}(1, 0.25^2)$; inhibitory columns are multiplied by -4. The diagonal is zero, $b_j \sim \mathcal{N}(0.18, 0.16^2)$, and positive rescaling sets $\rho(W) = g$, with eight circuits at each $g \in \{0.35, 0.60, 0.85\}$. Physical time is in milliseconds:

$$dx = \tau^{-1}[-x + W\tanh x + b + u(t)]\,dt + 0.05\,\tau^{-1/2}dB, \qquad \tau = 20\ \mathrm{ms}. \tag{15}$$

B has independent standard Brownian components. These variables are normalized deviations, not firing rates in Hz. Twenty observed excitatory neurons receive additive square pulses lasting 10 ms. Observations select the first 52 excitatory and 12 inhibitory neurons and add independent Gaussian noise with standard deviation 0.02. Local stability depends on $[-I + W\,\mathrm{diag}(\tanh'(x))]/\tau$.

Histories contain ten observations at $-90, -80, \ldots, 0^-$ ms, with the last measured before input delivery; futures contain ten observations at $10, \ldots, 100$ ms. Euler–Maruyama with step 0.25 ms specifies stochastic trial generation. Training-only means and standard deviations standardize observations for every method and split. Forecast scores retain these units; response scores undo output standardization. Doses remain in generator-current units.

After a 2 s transient, snapshots separated by 500 ms are allocated to training/validation/test in counts $96/24/48$; no state or overlapping history crosses splits. Repeated trials restore the full snapshot and observed history, then generate independent future process and measurement noise. The 61 conditions are blank and 20 single-source pulses at doses $0.1, 0.2, 0.4$, presented in randomized order. Each training or validation condition has eight response and four separate variance-pilot trials. Training acquisition therefore costs $96 \cdot 61 \cdot (8 + 4) = 70{,}272$ trials per circuit. Every estimator receives the same trials. Section B.2 removes this assay's reset requirement. Hidden state is never a model input. Primary responses average over the same 48 test histories before applying a matrix norm; these averages do not establish posterior response accuracy at every partially observed history.

B.2 Continuous randomized sessions without resets

This assay uses the twelve stable gain-1.3 circuits defined in Section B.6, independently drawn from the primary 24 circuits. It retains the rate equation and observation noise, adding shared drive $0.6vc(t)$ inside the drift. Entries of v are independent standard normals rescaled to unit root-mean-square, fixed within a circuit. Independently of B,

$$dc = -c\,dt/(500\,\text{ms}) + \sqrt{2/(500\,\text{ms})}\,dB_0.$$

Thus c has stationary standard deviation one. Each independent session starts x at the autonomous deterministic equilibrium and $c(0) \sim \mathcal{N}(0, 1)$, discards 10 s, then evolves continuously. The first six source channels are addressable. Onset intervals are independent uniforms on $[500, 1000]$ ms; no later pulse enters an outcome window. At each onset, blank has probability $p_0 = 1/4$, and each dose-0.4 single-source pulse has $p_i = 1/8$. Assignments are independent of pretreatment information and future noise, with logged probabilities. There is no restoration between trials or equilibrium requirement.

A separate 2,048-trial pilot fixes observation standardization and fits PC1 to flattened pretreatment histories, freezing tertile cutpoints, group masses $\widehat{p}_g$ and variance weights across acquisition budgets. Current-trial futures do not enter grouping. Each fitting budget $n \in \{2048, 8192, 32768\}$ uses two independent sessions of $n/2$ trials. Validation and test each use two further independent 4,096-trial sessions. Total acquisition is $n + 2048 + 8192 + 8192$, before evaluation-only reference simulation. All splits and random streams are disjoint.

Let $G(H_t)$ be the frozen group and $b_h(H_t)$ the last pretreatment observation. For $\delta = 0.4$, define

$$\begin{aligned} Z_{ti,h} &= \delta^{-1}\left[\frac{\mathbf{1}(A_t = i)}{p_i} - \frac{\mathbf{1}(A_t = 0)}{p_0}\right][Y_{t,h} - b_h(H_t)], \\ \widehat{d}_{gih} &= \frac{\sum_{t=1}^n \mathbf{1}\{G(H_t) = g\} Z_{ti,h}}{n\widehat{p}_g}, \qquad m_{\theta,gih} = \frac{\sum_{t=1}^n \mathbf{1}\{G(H_t) = g\} D^{\delta}_{\theta,:,i}(h, H_t)}{n\widehat{p}_g}. \end{aligned} \tag{16}$$

Both sums use the union of fitting-session histories, irrespective of assigned arm. Pretreatment subtraction cancels in expectation. Sequential randomization identifies pulse-minus-blank group effects under consistency, positive assignment probabilities, pretreatment-only eligibility and no additional intervention within the outcome window (Boruvka et al., 2018). Hidden state and earlier treatments are averaged under this acquisition policy; history need not reconstruct them.

The empirical objective is $(3q\,10m)^{-1} \sum_{g,i,h} \|m_{\theta,gih} - \widehat{d}_{gih}\|^2_{W_{gih}}$, averaging uniformly over three groups, $q = 6$ sources, all ten forecast horizons and m outputs; frozen weights have mean one across these coordinates. Two independent estimates of m_θ sample histories and generator bases independently from the *same* empirical history pool; both subtract the same fixed $\widehat{d}$. Their residual product is unbiased for this empirical objective conditional on the data. We do not claim an unbiased population squared-error estimator. Sampling uniformly within a group includes its empirical mass divided by $\widehat{p}_g$ to preserve the stated moment. Pilot weights use variances of $\mathbf{1}_g Z/\widehat{p}_g$ divided by

n, with the floor below; these are working weights, not a claim of serial independence. Test comparisons average predictions and separate simulator references over the same histories within each group; state restoration is confined to this evaluation oracle. This tests observable-group effects, not pointwise conditional effects.

B.3 Matched predictors and optimization

Attention models. CFM, PC-Flow, Gaussian transformer and its calibrated version use four blocks, each applying neuron attention, time attention and a feed-forward map, with width 128, four heads, feed-forward width 256, pre-layer normalization and SiLU activations (Vaswani et al., 2017). A shared $10 \to 64 \to 128$ temporal MLP embeds each neuron's history. Tokens concatenate history, 16-dimensional neuron identity and physical-time embeddings, local stimulus amplitude and a scalar future query. A projection of the complete stimulus and a $16 \to 128 \to 128$ transport-time embedding are added globally. Flow uses the interpolant as its query; mean heads use zero queries and transport time. Joint future attention accesses generated coordinates, not measured futures at inference. CFM uses standard Gaussian base noise, a straight conditional path and 16 Heun steps; 32 steps test discretization.

Gaussian transformer predicts the full future mean and covariance $LL^\top + \mathrm{diag}(s^2)$, with rank-eight L and $s = \mathrm{softplus}(d) + 10^{-4}$, and optimizes Gaussian negative log likelihood per coordinate. Its calibrated version adds the same finite-contrast objective, using analytic mean contrasts. ResponseNet uses the same backbone with a deterministic direct-response head and source/dose input, sharing parameters across sources and horizons; it has no forecast likelihood. At $m = 64, q = 20$, CFM/PC-Flow/ResponseNet each have 847,809 parameters; either Gaussian transformer has 848,970. Their heads output one and ten values per token, respectively.

Other baselines. Pooled averages training contrasts per source/dose, ignoring history. LowRank is SVD shrinkage of their $(4m) \times q$ matrix at the four scored horizons. Row variance is the source-averaged sample variance of four pilot contrasts; whiten by $\sqrt{\max(\nu_{\mathrm{row}}, 0.05 \ \mathrm{median}\, \nu_{\mathrm{row}}, 10^{-12})}$. Exclude source-own entries from fitting and validation. Initialize missing entries to zero; each of 20 completion iterations takes a rank-r SVD projection, restores observed entries and updates missing ones from the projection. A final rank-r projection has singular values multiplied by α. Choose $r \in \{4, 8, 12, 16, 20\}$, clipped to source count and deduplicated, and $\alpha \in \{0.65, 0.8, 0.9, 1\}$ by masked whitened mean squared error against an independent eight-trial validation mean, then undo row whitening. Full rank is eligible. Fits are separate per dose, including dose four in the strong-input challenge.

LocalResponse fits source/dose-specific ridge regressions with intercepts from 16 training-history principal components to complete response vectors. In the continuous assay, direct predictors fit the corresponding randomized group moments; Pooled/LowRank use the unconditional inverse-probability response estimate.

Ridge VARX uses ten observation lags and planned stimulus. Its penalty grid is $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, with Gaussian residual covariance estimated during training. NARX replaces its mean by a two-layer width-256 SiLU network with diagonal Gaussian innovations (267,904 parameters). GRU has two width-128 layers and Gaussian output innovations (197,760 parameters) (Cho et al., 2014); GRU+calibration adds the response objective. All recursive predictors receive identical histories and correctly timed inputs; direct response models supply only contrasts and column sums.

Mechanistic rate RNN. RateRNN uses 80 latent states and the drift $[-x + W_\theta \tanh x + b_\theta + Bu]/20$ in physical units. It learns a dense off-diagonal W_θ, background drive, diagonal process and observation noise scales, and a $640 \to 128 \to 160$ SiLU history encoder. The encoder outputs a mean correction and positive diagonal scale for the initial-state distribution; observed coordinates are initialized around the last recorded activity after undoing standardization. Input and observation selectors match the assay, but true weights, synaptic signs and hidden snapshots are not supplied. Euler–Maruyama uses a 0.5 ms step, compared with 0.25 ms in the data protocol. Four independent generated futures estimate the fair trajectory energy score in Appendix B.5; this is trial-distribution matching, not an exact latent-SDE likelihood. RateRNN+calibration adds the same eight-pair response loss. Both variants have 109,312 nominal trainable parameters in the primary assay and use

the common fitting/selection rules. In the fully observed noiseless control, the last observation fixes the complete initial state and noise is zero; the encoder and noise parameters are frozen.

Calibration and selection. For primary repeated snapshots, independent pilot variances give $\nu^{\delta}_{i\ell} = (s^2_{+,i\ell}/8 + s^2_{0,\ell}/8)/\delta^2$, pooled across training histories; other repetition budgets use their actual counts. Add the median ν^{δ}, floored at 10^{-4} in standardized-response units squared, before inversion. Normalize inverse weights to mean one per dose and freeze them. Shared blanks induce correlations; diagonal weights do not assert independence. Four snapshot/source/dose groups enter every fourth update, with a factor four preserving the expected λ weight. Eight coupled endpoint pairs split into independent halves of four estimate the model residual product. The continuous assay samples four group/source combinations with the same eight-pair allocation.

Learned predictors use AdamW, batch size 32, learning rates $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}\}$, weight decay 10^{-4}, gradient clipping at one and at most 30,000 updates. Checkpoints every 1,000 updates use patience five; three fits per circuit share splits. Default calibrated comparisons fix $\lambda = 0.3$, including every point of the repetition curve. A separate sensitivity analysis evaluates $\lambda \in \{0, 0.1, 0.3, 1\}$; this sweep does not select a different weight at each repetition budget. Primary selection minimizes validation CRPS in training-standardized observation units plus 0.1 times validation response NRMSE, using the same source-own mask and horizon aggregation as evaluation. Direct models use response NRMSE alone. A second analysis selects forecasters by CRPS alone with identical candidate sets. Both selectors use the same validation trials.

Validation uses 32 endpoint samples split $16 + 16$; primary response readout uses 256 samples. Analytic Gaussian means require no Monte Carlo approximation. Ablations vary calibration samples $4/8/16$ at fixed 256-sample readout, and readout $8/32/128/256$ at a fixed checkpoint. Heun $16/32$ varies separately. Uniform weights and shuffled source targets leave raw forecast labels unchanged.

B.4 Compute cost and the equal-time control

Table 4 uses random tensors of shape $32 \times 64 \times 10$, an Apple M1 Pro CPU, two threads, PyTorch 2.6.0 and macOS 14.4.1 (arm64). Each mode has one warm-up and three timed optimizer updates. Scheduled cost is 0.75 times the ordinary-update median plus 0.25 times the calibration-update median. Flow uses 16 Heun steps with activation checkpointing during calibration. Readout timings use medians for one history/source contrast with 32 shared-noise pairs, versus 256 pairs in the response benchmark, or analytic/direct means where available.

Table 4: **CPU step and readout microbenchmark.** Update and readout times are seconds; memory is whole-process peak RSS. Ordinary updates exclude calibration; superscript a denotes analytic/direct readout. These are random-tensor measurements, not end-to-end fitting times.

Model	Ordinary (s)	Scheduled (s)	Contrast (s)	RSS (GiB)
CFM	0.970	0.970	21.2796	1.082
PC-Flow	0.787	18.882	21.8389	1.302
Gaussian TF + cal.	0.859	0.909	0.0268^a	1.081
ResponseNet	0.864	0.864	0.0139^a	1.107
GRU + cal.	0.009	0.017	0.0046	0.270

PC-Flow's scheduled update costs $18.882/0.970 \approx 19.5$ times CFM's; Gaussian calibration costs $0.909/0.859 \approx 1.06$ times its ordinary update. Each 16-step Heun trajectory requires 32 velocity evaluations.

The equal-time control gives CFM PC-Flow's training-and-validation wall-time allowance, removing its 30,000-update limit and patience stopping rule. Data, readout and validation criterion remain fixed; checkpoints are evaluated every 1,000 updates and the best eligible checkpoint is retained when time expires. Extra optimization supplies no additional trials.

Scaling the response readout. With $T = 10$ bins, factorized neuron and time attention costs $O(Tm^2 + mT^2)$ per block, besides token projections. The primary assay has $m = 64, q = 20$; it does not establish throughput for thousands of cells. Evaluating 48 histories, 20 sources and 256

pairs requires 960 context–source queries. Multiplying the measured 32-pair time by eight and by 960 projects 46.6 CPU hours per circuit and fit under serial execution. Sharing each history's blank trajectories across sources reduces the trajectory count by $21/40$, projecting 24.5 hours. These are arithmetic projections, not observed runtimes or measurements of batching or device-acceleration benefits.

B.5 Response, distribution and composition scores

Responses and uncertainty. For the primary response average, E_D averages masked relative Frobenius errors at $h = 10, 30, 60, 100$ ms, excluding the source's own row. Model and reference contrasts average over identical histories before each norm. Denominators are reference norms; signed correlation concatenates masked entries across horizons.

State and group metrics instead retain that axis. Let D_{ah} and $\widehat{D}_{ah}$ denote a state's response or an observable group's mean response, M the fixed source-own mask, and ω_a fixed weights summing to one. Define $\bar{D}_h = \sum_a \omega_a D_{ah}$ and $\widehat{\bar{D}}_h = \sum_a \omega_a \widehat{D}_{ah}$. The raw and centered errors are

$$
\begin{aligned}
E_{\mathrm{raw}} &= \frac{1}{4} \sum_h \sqrt{\frac{\sum_a \omega_a \|M \odot (\widehat{D}_{ah} - D_{ah})\|_F^2}{\sum_a \omega_a \|M \odot D_{ah}\|_F^2}}, \\
E_{\mathrm{het}} &= \frac{1}{4} \sum_h \sqrt{\frac{\sum_a \omega_a \|M \odot [(\widehat{D}_{ah} - \widehat{\bar{D}}_h) - (D_{ah} - \bar{D}_h)]\|_F^2}{\sum_a \omega_a \|M \odot (D_{ah} - \bar{D}_h)\|_F^2}}.
\end{aligned}
\tag{17}
$$

Use $\omega_a = 1/24$ for the fully observed states and $1/3$ for the three observable groups. Squared errors are summed before the relative square root; response matrices are not averaged first. Centering removes each model's own weighted mean and the reference mean separately. History-independent Pooled, LowRank and linear VARX therefore have $E_{\mathrm{het}} = 1$ whenever reference heterogeneity is nonzero, regardless of their mean-response accuracy. Zero heterogeneity leaves this metric undefined. Horizon-specific raw and centered errors accompany their averages.

Three initialization scores average within each circuit; 2,000 paired bootstrap resamples of independent circuits give percentile 95% intervals for scores and method differences, including CRPS. State, continuous-session and strong-composition controls reuse their twelve circuit draws with separate trial laws and refits; their circuit counts are not pooled. Matrix entries and trial bins are not independent replicates.

Joint predictive scores. The primary CRPS check uses four held-out histories in each of 24 rate circuits, with blank and source-zero/source-one pulses as three equally weighted conditions. An independent autonomous session uses a 2 s transient and snapshots at 2.0, 2.5, 3.0 and 3.5 s; histories have the same ten lags and measurement noise as the primary protocol. No preparatory drive is added to these contexts. It averages fair scalar scores across all $10m = 640$ standardized future coordinates, using 64 model draws and 32 independent reference outcomes per condition. This same-history, same-stimulus forecast diagnostic is distinct from the response metric over all 48 test histories and twenty sources. An approximate oracle uses 256 independent draws from the reference conditional law, scored against the same 32 outcomes. PC-minus-oracle differences retain circuit pairing; finite-sample fair scores can fluctuate on either side of their population values. A separate joint-distribution diagnostic uses twelve continuous-assay circuits and three fixed histories, one representing each pretreatment group, with blank and first-source stimulation scored separately at each history, using 128 model draws and 256 independent reference outcomes per condition. Its projection fixes eight observed cells (indices 2–9, excluding the stimulated source) at 30 and 60 ms, hence $d = 16$ coordinates. For model draws X_k and one reference outcome y,

$$
\widehat{\mathrm{ES}} = K^{-1} \sum_k \|X_k - y\|_2 - [2K(K-1)]^{-1} \sum_{k \neq l} \|X_k - X_l\|_2.
$$

Divide by $\sqrt{d}$ for a fixed dimension scale, and average over reference outcomes. Scalar-coordinate versions give fair CRPS (Gneiting & Raftery, 2007; Ferro, 2014). This energy score assesses the specified joint projection. The complementary variogram score uses $p = 1/2$ and equal weights summing to one over its 120 distinct coordinate pairs (Scheuerer & Hamill, 2015). With $a_{jl} = |y_j -$

$y_l|^p$ and $v_{kjl} = |X_{kj} - X_{kl}|^p$, its unbiased ensemble estimator is $\sum_{j<l} w_{jl}[(a_{jl} - \bar{v}_{jl})^2 - s^2_{v,jl}/K]$. The correction removes finite-sample variance bias; estimates may be negative. Variogram scoring is proper but not strictly proper.

Scores average within each condition before averaging conditions; stimulated-only summaries are also reported. A covariance diagnostic compares conditional covariance at the same fixed history and stimulus, with off-diagonal Frobenius error divided by the reference norm; near-zero denominators remain unresolved. A Gaussian with matched reference mean and covariance separates errors in those moments from distributional shape. These diagnostics do not certify the full 640-dimensional law or identify per-history effects from the unpaired acquisition data.

Composition. Six source sets at each size $k \in \{1, 2, 4, 8\}$ are fixed before outputs are generated. Dose a is applied to every source, and columns are read at that dose. Compare full mean contrast, $a \sum_{i \in S} \widehat{D}^a_{:,i}$, and reference single-source column sum against compound reference. All use identical histories, horizons and targets outside S; relative Euclidean errors average equally over sets and horizons. A single cache of blank and single-source endpoint means constructs full and column-sum predictions at $k = 1$, making their errors identical to numerical precision. Direct models enter column-sum comparisons. Compound trials never enter training or selection. For the interaction score, compute $q = \Delta^S - \sum_{i \in S} \Delta^i$ and its model counterpart from the same endpoint caches. Within each circuit and set size, compute the Euclidean norm of masked interaction error divided by the corresponding reference interaction norm for each of six sets and four horizons, then average these 24 ratios equally. This gives weak and strong interaction queries equal weight and can amplify errors on small residuals. Grid refinement checks each nonzero reference residual individually; a zero denominator would remain undefined rather than being regularized into a score. At $k = 1$, both residuals vanish exactly and the relative ratio is undefined.

Noise-budget and structural controls. An evaluation-only contrast averages eight independently measured stimulated and blank outcomes at each of the same 48 evaluation histories. It is normalized and masked exactly like the primary response metric. An infinite-repeat mean at these histories equals the defined reference; this zero-error limit does not describe a pooled estimate from a different training-history distribution. Neither control adds trials to model fitting. The same primary predictions are stratified by recurrent gain and by output-pair anatomy. Direct pairs have $W_{ji} \neq 0$; indirect pairs have no direct edge but a directed path from source to target. Source-own pairs are excluded. Excitatory/inhibitory strata refer to the target cell. Every stratum uses its own reference Frobenius norm, averaged across horizons; weak indirect responses can therefore have large relative errors even when absolute errors are small. Eight circuits support each gain stratum; all 24 support the anatomical strata.

Reference precision. Evaluation-only pulse/blank simulations couple exogenous future noise and remove known zero-mean observation noise. For K_H fixed histories and R paired futures, the estimated squared Monte Carlo error is $V_h = (K_H^2 R)^{-1} \sum_k \operatorname{tr} \widehat{\operatorname{Cov}}_r(Z_{kr,h})$. An excluded 128-pair pilot chooses a power-of-two $R \geq 256$ targeting $\sqrt{V_h}/\|\bar{D}_h\| \leq 1\%$ at every scored horizon. Production rechecks precision; failures double the budget to 2^{20} pairs per history and remain unresolved if still failing. Zero pilot norms or variances require expansion. An independent production batch supplies an agreement diagnostic reporting both Monte Carlo errors. These are estimated precision checks, not confidence guarantees or history-sampling uncertainty.

Interaction precision is checked directly on $Y^S - Y^0 - \sum_{i \in S}(Y^i - Y^0)$ under coupled exogenous noise. Its Monte Carlo standard error must be at most 5% of the interaction-residual norm and 0.1% of the compound-response norm. Unresolved small residuals retain absolute uncertainty; precise individual columns alone do not establish interaction precision.

B.6 State dependence and nonlinear composition controls

Twelve additional circuits have $\rho(W) = 1.3$. Equilibrium integration starts at zero, uses RK4 at 0.5 ms for at least 2 s, and extends in 500 ms increments to at most 8 s, targeting residual below 10^{-9}. Accept only residual below 10^{-7} and negative local-Jacobian spectral abscissa; seeds and rejected draws are recorded, with no filtering on interaction size. This establishes local equilibrium stability, not global contraction. Prepare disjoint $96/24/24$ training/validation/test states by applying dense

currents drawn independently as $\mathcal{N}(0, 0.75^2 I)$ for 100 ms from equilibrium, then releasing them. At exactly observed $x(0^-)$, compare blank and dose-0.2 pulses on the first six sources. Models are refitted under this preparation law. This control removes process and observation noise, observes all 80 states and uses RK4 at 0.5 ms. Its deterministic per-state references use Equation 17.

These circuits also define the stronger composition assay, with equilibrium backgrounds, all twenty source channels and single-source training doses $0.2, 0.8, 4$. Its main compound challenge uses dose four, sizes $1, 2, 4, 8$, and dose-matched columns. Report reference superposition remainder, estimated-column error and full-model error. High dose tests a specific departure from additivity; local stability and single-source accuracy do not guarantee interaction accuracy.

Local derivative versus finite-dose labels. An oracle-label diagnostic refits the same flow backbone on the fully observed, noiseless state assay. The three objectives are ordinary forecasting, forecasting plus exact local pulse-amplitude derivatives at zero, and forecasting plus exact finite-dose contrasts. Fits at doses 0.2 and 4 are separate and use the same source/output coordinates and state splits. These labels contain different information: this control tests target alignment, not acquisition efficiency. Ordinary benchmark training receives neither true derivatives nor hidden weights. Both local-derivative and finite-dose errors use the state-resolved raw score with source-own entries excluded.

For $J = \partial x/\partial a|_{a=0}$, the oracle integrates

$$\dot{J} = \tau^{-1}\big[-J + W\,\mathrm{diag}(1 - \tanh^2 x)J + B\mathbf{1}_{[0,10)}(t)\big], \qquad J(0) = 0,$$

along the blank trajectory; columns of B select the six sources. Independent finite-pulse trajectories give the dose-normalized contrast. RK4 uses 0.5 ms steps with an exact pulse-offset boundary. Central differences at amplitudes 10^{-4} and $5 \cdot 10^{-5}$ and a 0.25 ms refinement at two prespecified states per circuit give maximum relative discrepancy below $2.2 \cdot 10^{-8}$.

B.7 Observation stress, spiking refits and linear controls

Observation fractions $1, 0.8, 0.5$ refer to the 80 underlying neurons. Nested observed sets are all 80 cells, the primary 64-cell set, and 40 cells comprising excitatory indices 0–31 and inhibitory indices 64–71. All retain the twenty source cells. Every method is refitted with the same snapshot split; scores use the observed target rows at each fraction, with source-own entries removed. The 0.8-fraction result is exactly the primary assay. The repetition curve uses $2, 4, 8, 16, 32$ fitting repetitions and four shared separate pilot repetitions per condition. The fixed-model dose stress test uses continuous-input forecasters and ResponseNet at doses $0.1, 0.2, 0.4, 0.8$; only 0.8 is outside the fitting/selection support. Its 0.2-dose result is again the primary assay. Dose-specific direct estimators remain in same-dose composition comparisons. Observation, dose and acquisition budgets vary separately; Appendix B.8 reports their numerical results.

The LIF assay refits models on 16 independent 200-neuron circuits, with 160 excitatory cells, connection probability 0.1 and 1.5 ms delays (Brunel, 2000; Stimberg et al., 2019). Log-normal jump currents have coefficient of variation 0.3 and means $0.015/0.060$ nA for excitation/inhibition, decaying over $5/10$ ms. Membrane parameters are $C = 200$ pF, $\tau_m = 20$ ms, rest/reset -65 mV, threshold -50 mV and refractory duration 2 ms. Independent 1.5 kHz and shared 100 Hz external Poisson drives produce 0.04 nA excitatory jumps. Discard 2 s and integrate at 0.1 ms. Observe 160 cells, including 16 excitatory sources, with 10 ms count bins and uniform $[0, 1)$ dequantization. Ten-ms pulses use $0.1, 0.2, 0.4$ nA, primarily 0.2 nA. Snapshot counts remain $96/24/48$, with 16 fitting and eight pilot trials per condition. References couple external Poisson streams and use integer counts; endogenous spikes are not forced to match. Gaussian baselines model dequantized counts. This is separate training on spiking observations, not transfer of rate-model parameters.

Linear controls. The three-cell illustration has $W_{21} = W_{32} = 0.85$, other weights zero, $\tau = 20$ ms, linear activation and a pulse to cell one. An eight-state numerical check uses a stable coupling with spectral radius 0.65: a block matrix exponential evaluates the square-pulse response without assuming invertible drift, and RK4 steps $0.5/0.25$ ms and finite differences check it independently. The 0.5 ms relative error is 9.95×10^{-9}.

Sixteen fully observed linear Gaussian circuits use the primary weight/drive distributions with gain 0.65, linear activation and no measurement noise. Exact 10 ms Gaussian transitions include inte-

Table 5: **Spiking refits and the linear limit.** Response errors and circuit-level intervals for separately fitted models.

Model	LIF response	Linear response
Ridge VARX	0.528 [0.471, 0.587]	0.100 [0.090, 0.111]
Response network	0.305 [0.279, 0.333]	0.117 [0.109, 0.125]
GRU + cal.	0.331 [0.298, 0.367]	0.135 [0.121, 0.151]
Gaussian TF + cal.	0.273 [0.240, 0.316]	0.127 [0.114, 0.141]
CFM	0.412 [0.355, 0.472]	0.133 [0.123, 0.143]
PC-Flow	0.255 [0.221, 0.291]	0.138 [0.124, 0.152]

grated constant drive, input and process covariance; pulses align with the grid. Analytic conditional means give references. This one-lag transition lies within the ten-lag VARX family.

B.8 Additional comparisons and sampling diagnostics

Table 6 includes every primary baseline and both composition readouts. Direct estimators supply only column sums.

Table 6: **All estimators under the common primary readout.** Response intervals use independent circuits; composition columns report the eight-source error at dose 0.2. Dashes indicate outputs the estimator does not define.

Model	Response	CRPS	Columns ($k = 8$)	Full ($k = 8$)
Pooled contrast	0.330 [0.301, 0.362]	–	0.379	–
SVD shrinkage	0.314 [0.290, 0.339]	–	0.355	–
Response network	0.183 [0.163, 0.206]	–	0.212	–
Ridge VARX	0.383 [0.348, 0.421]	0.134	0.438	0.438
NARX	0.271 [0.248, 0.297]	0.132	0.311	0.311
GRU	0.317 [0.297, 0.336]	0.132	0.365	0.365
GRU + cal.	0.194 [0.174, 0.216]	0.131	0.222	0.223
Gaussian TF	0.285 [0.257, 0.315]	0.128	0.330	0.330
Gaussian TF + cal.	0.190 [0.173, 0.207]	0.128	0.219	0.219
CFM	0.317 [0.295, 0.341]	0.128	0.364	0.364
PC-Flow	0.120 [0.109, 0.133]	0.128	0.138	0.138
Local response ridge	0.267 [0.241, 0.296]	–	0.305	–
Rate RNN	0.248 [0.225, 0.273]	0.125	0.207	0.207
Rate RNN + cal.	0.164 [0.148, 0.179]	0.125	0.137	0.136

Table 7: **Joint predictive scores under stimulation.** The continuing-assay projection has 16 coordinates. Paired energy-score differences retain circuit pairing; positive differences favor PC-Flow. Variogram scores assess pairwise increments and complement the energy score.

Model	Stim. ES	Stim. VS	ES minus PC-Flow
GRU + cal.	0.1636	0.0181	0.0011 [0.0008, 0.0014]
Gaussian TF + cal.	0.1638	0.0181	0.0012 [0.0010, 0.0015]
CFM	0.1634	0.0181	0.0009 [0.0007, 0.0010]
PC-Flow	0.1625	0.0181	–
Gaussian, matched moments	0.1638	0.0181	0.0013 [0.0009, 0.0018]

Table 3 and Figure 6 report objective, checkpoint and sampling controls specified in Appendix B.3. Readout-pair comparisons use nested noise draws at one checkpoint; training-pair comparisons share the primary 256-pair readout.

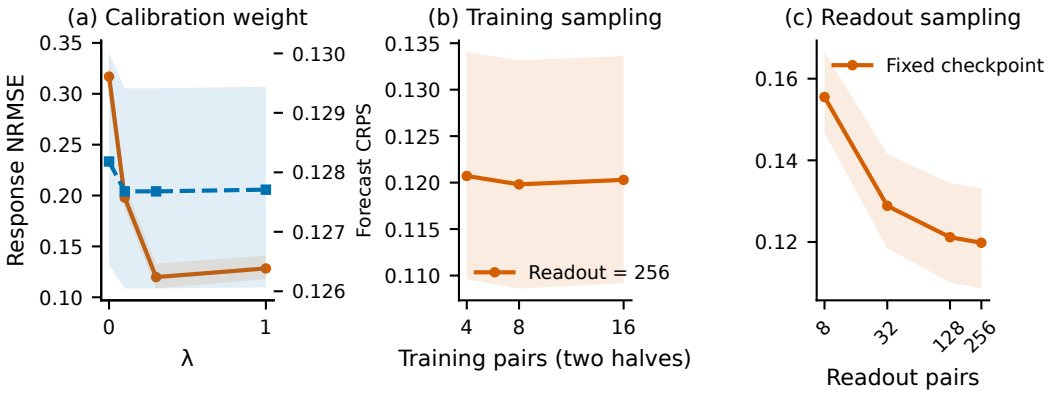

Figure 6: **Separate objective strength from Monte Carlo precision.** (a) Response error and CRPS versus calibration weight. (b) Training pairs at fixed 256-pair readout. (c) Readout pairs at one checkpoint. Intervals resample circuits.

Table 8: **Observed fraction and dose stress, fully reported.** Fractions refer to all 80 cells. Each observation setting is refitted; doses share a fixed fitted model. Only 0.8 is outside training support. Brackets give 95% circuit bootstrap intervals.

Setting	PC-Flow	Gaussian + cal.	Rate RNN + cal.
Observed 1.0	0.107 [0.097, 0.119]	0.169 [0.154, 0.184]	0.146 [0.133, 0.159]
Observed 0.8	0.120 [0.109, 0.133]	0.190 [0.173, 0.207]	0.164 [0.148, 0.179]
Observed 0.5	0.163 [0.148, 0.181]	0.259 [0.236, 0.283]	0.224 [0.202, 0.246]
Dose 0.1	0.135 [0.122, 0.150]	0.214 [0.195, 0.233]	0.164 [0.148, 0.179]
Dose 0.2	0.120 [0.109, 0.133]	0.190 [0.173, 0.207]	0.164 [0.148, 0.179]
Dose 0.4	0.117 [0.106, 0.130]	0.185 [0.169, 0.201]	0.164 [0.148, 0.179]
Dose 0.8	0.170 [0.155, 0.189]	0.271 [0.247, 0.294]	0.164 [0.148, 0.179]

Table 9: **Circuit gain and structural response strata.** Direct and indirect refer to known anatomical support; cell type refers to the target. Each row has its own reference norm. Gain rows contain eight circuits each; anatomy rows use all 24.

Stratum	PC-Flow	CFM	Gaussian + cal.
Gain 0.35	0.134 [0.110, 0.166]	0.317 [0.287, 0.342]	0.204 [0.171, 0.233]
Gain 0.60	0.108 [0.095, 0.121]	0.298 [0.264, 0.335]	0.170 [0.147, 0.194]
Gain 0.85	0.117 [0.106, 0.133]	0.335 [0.296, 0.387]	0.197 [0.177, 0.223]
Direct support	0.054 [0.049, 0.058]	0.165 [0.151, 0.181]	0.089 [0.081, 0.097]
Indirect reachable	0.587 [0.476, 0.713]	1.554 [1.309, 1.837]	0.881 [0.740, 1.036]
Excitatory targets	0.119 [0.108, 0.132]	0.316 [0.294, 0.340]	0.189 [0.172, 0.207]
Inhibitory targets	0.124 [0.112, 0.140]	0.325 [0.303, 0.347]	0.196 [0.179, 0.212]

Table 10: **Full response and interaction-residual error at dose four.** Interaction errors normalize within each set and horizon, then average equally. The one-source interaction is exactly zero and has no defined relative error. Full predictions and columns share their one-source endpoints.

Sources	PC full	Gaussian full	PC interaction	Gaussian interaction
1	0.077 [0.054, 0.108]	0.111 [0.082, 0.147]	–	–
2	0.052 [0.043, 0.061]	0.078 [0.067, 0.093]	0.389 [0.159, 0.768]	0.959 [0.413, 1.832]
4	0.048 [0.044, 0.053]	0.076 [0.066, 0.088]	0.090 [0.079, 0.103]	0.238 [0.205, 0.275]
8	0.049 [0.044, 0.055]	0.078 [0.069, 0.089]	0.058 [0.052, 0.066]	0.150 [0.140, 0.162]

Table 11: **Noise-budget references and unbiased mean-loss alternatives.** Evaluation-only repeat controls receive measurements at the test histories. CRPS uses the same four contexts and reference outcomes as Table 2. Objective differences use paired circuits at the same eight-pair training budget; negative differences favor the first named estimator.

Control	Mean [95% CI]
Independent evaluation mean, 8 repeats	0.3285 [0.2990, 0.3597]
Evaluation-history oracle, infinite repeats	0.0000 [0.0000, 0.0000]
Naive minus split calibration	0.0173 [0.0128, 0.0223]
All-pairs U-statistic minus split	0.0000 [-0.0010, 0.0013]
Approximate oracle CRPS	0.1249 [0.1233, 0.1268]
PC-Flow CRPS minus oracle	0.0028 [0.0021, 0.0034]

Table 12: **Oracle supervision distinguishes local and finite-dose targets.** The same flow backbone is fitted separately for each dose with forecast-only, local-derivative or finite-contrast supervision. Both response targets are evaluated on held-out known states. This is a target-alignment control with privileged labels, not an equal-acquisition comparison.

Oracle control, dose	Local J^0 error	Finite D^δ error
CFM, 0.2	0.175 [0.161, 0.193]	0.179 [0.164, 0.196]
Oracle local derivative, 0.2	0.034 [0.031, 0.037]	0.038 [0.035, 0.041]
Oracle finite contrast, 0.2	0.126 [0.116, 0.139]	0.033 [0.031, 0.035]
CFM, 4.0	0.175 [0.161, 0.193]	0.284 [0.269, 0.298]
Oracle local derivative, 4.0	0.034 [0.031, 0.037]	0.195 [0.186, 0.204]
Oracle finite contrast, 4.0	0.126 [0.116, 0.139]	0.038 [0.036, 0.041]

C PUBLIC FLUORESCENCE-DATA ILLUSTRATION

Data and prespecified inclusion. We fitted regression models to Randi et al. (2023)'s recording exports from OSF (`https://doi.org/10.17605/OSF.IO/E2SYT`). Before inspecting outcomes, we selected IDs 0–8 in integer order. These covered three dates; a metadata-only amendment added the first subsequent export on each of three new dates: IDs 9, 12 and 13. The twelve recordings contain 121,067,129 bytes.

Each export contains fluorescence, timestamps, neuron labels, stimulation-volume indices, targeted-neuron indices and an original recording path. We preserved label-line indexing and retained only exact, unique identities in the authors' canonical anatomical-neuron list. Partial, duplicated and uncertain labels were excluded without merging sides or converting AWCON/AWCOFF to anatomical left/right identities. Negative or invalid target indices and stimulated self-outputs were excluded. We applied no autoresponse or downstream-amplitude selection. Genotype, manual versus automatic targeting, event-level quality and individual worm identity are not established by these files. The supplied sampling interval is 0.5 s; the paper describes 500-ms WT illumination and 300-ms mutant illumination, but the exports do not supply event-specific durations. These signals therefore differ from the rate benchmark's 10-ms current pulses.

Observed response and predictor. For each eligible event and output, let F_0 be its finite-sample mean fluorescence from -12 to -2 s before the logged onset. We require $F_0 > 0$ and at least 80% finite observations in that window. The target is the mean of $F/F_0 - 1$ from 2 to 10 s, also requiring 80% finite observations. Other stimulation onsets must lie outside the full $[-12, 10]$ s window. The delayed readout excludes the reported pulse interval but does not establish absence of residual artifacts. Post-window missingness is recorded separately from pretreatment eligibility and can still induce measurement selection.

The predictor uses five pretreatment samples at $-10, -8, -6, -4, -2$ s from both the output and the targeted neuron, normalized by their respective F_0. Missing history samples are interpolated from pretreatment observations only. Training-only means and standard deviations standardize these ten features, with a standard-deviation floor of 0.01. Additional terms are source, target and source–target-pair offsets. Unknown training identities receive no offset. The ordinary model minimizes trial-output squared error with ridge regularization. Its calibrated counterpart adds the uniformly pair-weighted squared error of training pair-mean residuals. Both objectives use identical trials and relative-fluorescence units. This is moment reweighting in a linear model, not a distinct causal estimator. Pretreatment subtraction is not a no-stimulation counterfactual.

Held-out evaluation. Each outer fold holds out an entire recording date. Inner leave-date-out validation selects ridge strength from $\{0.001, 0.01, 0.1, 1, 10\}$. The calibrated family additionally selects its group-loss weight from $\{0.1, 0.3, 1, 3\}$; this grid excludes zero, and the ordinary model supplies the separate zero-weight control. Selection minimizes trial RMSE plus group-mean RMSE, equally averaged over inner validation dates. The evaluation mask retains only source–target pairs seen on at least two training dates, independently of held-out response amplitudes. Group errors first average repeated events within each recording and pair, then equally weight these recording–pair means. Final summaries equally weight the six held-out dates. Group NRMSE divides group RMSE by the measured group-response root-mean-square, floored at 0.01; no response-magnitude threshold selects pairs.

The 824 stimulation entries yield 264 events with valid target identities and pretreatment histories. Of 7,866 pretreatment-eligible trial-output measurements, 567 lack sufficient post-window samples, leaving 7,299 analyzed measurements. Only 1,078 measurements (14.77%) satisfy cross-date pair support, forming 726 recording–pair means. Table 13 reports the actual fitted predictions. Calibration does not improve this subset: its group-RMSE change is $+0.000142$, with descriptive paired date-resampling interval $[+0.000021, +0.000298]$. Both fitted predictors remain close to the zero-change baseline. We performed no outcome-based cohort revision or retuning after observing this result.

Table 13: **Actual held-out public-data prediction.** RMSE uses relative-fluorescence units; lower is better. Results average six held-out dates on the common-support subset. This comparison fits ordinary and group-calibrated ridge models, not PC-Flow.

Predictor	Trial RMSE	Group RMSE	Group NRMSE
Ordinary ridge	0.10992	0.10157	0.99292
Ridge + group calibration	0.10995	0.10171	0.99400
Zero post-minus-pre change	0.11054	0.10175	1.00000

Limits and reproducibility. The 2,000 paired resamples give descriptive six-date intervals, not independent-worm population confidence intervals. Matched blanks, logged assignment propensities and verified animal identifiers are unavailable. The exports lack a preprocessing manifest; upstream scripts support spike removal, causal smoothing and photobleaching correction, limiting this to retrospective prediction.