WHICH EXPERIMENT WOULD CHANGE YOUR MIND? MEASUREMENT-AWARE EXPERIMENTAL DESIGN FOR CONNECTOME-CONSTRAINED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Connectome-constrained models combine maps of synaptic wiring with task-driven learning to predict neural activity. These networks can solve the same task while making incompatible predictions about neural activity. However, it's imperative to ask the question: what recording would distinguish them? We formulate the joint choice of cell type and stimulus as Bayesian experimental design over trained networks, accounting for what calcium imaging can measure. The framework integrates out recording-specific gain and offset analytically, marginalizes uncertain voltage-to-calcium transforms at inference, and searches 38,870 candidate recordings subject to genetic targeting constraints. In a benchmark of 91 networks calibrated using 313 published calcium recordings, one selected recording identifies the generating network in 63.7% of single-trial and 80.8% of five-trial measurements. A conventional T5b motion experiment reaches 50.8% and 63.9%, and the first recording of a standard battery reaches 37.1% and 55.4%. For single-trial measurements, stimulus optimization on the selected target adds 12.3 percentage points over a generic edge, demonstrating the value of joint design. A protocol fixed on the original ensemble reaches 90% identification within 65 seconds of scheduled recording in two retrained ensembles. Retrospective tests on published recordings improve held-out prediction over random conditions. These results turn disagreement among connectome-constrained models into concrete recording choices, with identification evaluated within an explicit finite hypothesis set.

1 INTRODUCTION

A connectome specifies the wiring of a neural circuit, but leaves many aspects of its dynamics undetermined. Connectome-constrained networks learn the missing time constants, resting potentials, and synaptic gains by solving a task (Litwin-Kumar & Turaga, 2019; Lappalainen et al., 2024), and this approach can predict neural activity across a sensory system, but it also leaves an experimental opportunity: networks with similar task performance can make sharply different predictions about the same neuron. In the 50-network ensemble of Lappalainen et al. (2024), at least ten networks disagree with the others about the sign of a flash response in 32 of 65 cell types (Appendix A). Therefore, we can conclude that task performance alone does not decide which response a fly will produce.

Resolving that ambiguity requires choosing *where to record and what stimulus to present*, and each choice commits a genetic driver line, a fly preparation, and imaging time. The largest difference between simulated voltages doesn't need to be the most informative experiment: calcium imaging smooths fast dynamics, rectifies responses, and introduces uncertain fluorescence gain and baseline. A useful recording must separate the networks after these transformations, at the noise level of the experiment.

We formulate this joint choice as measurement-aware Bayesian experimental design, where each trained network is a candidate hypothesis. We predict the calcium recording under every hypothesis, integrate out its unknown gain and offset, and select the cell-type–stimulus pair with the highest expected information gain (EIG). A fast screen makes this calculation practical across 38,870 can-

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.

didate recordings, and genetic targeting and time constraints then turn the scores into a recording protocol.

The paper makes three contributions:

1. **A tractable measurement-aware design framework.** A Gaussian observation model integrates calibration uncertainty analytically, while a finite mixture handles uncertain imaging transforms at inference. The resulting likelihood supports joint selection of recording targets and stimuli at the scale of a 45,669-neuron connectome model.
2. **A data-calibrated identification benchmark.** Across 91 trained networks, one selected recording identifies its generating network substantially more often than a conventional motion experiment, a connectivity-based target, or the first recording of a standard battery (Table 1).
3. **Feasible protocols and tests on published measurements.** A frozen protocol transfers to two retrained ensembles, and retrospective analyses test whether selected conditions help predict responses to unobserved stimuli.

The identification benchmark deliberately separates experimental design from inference: the generating network is withheld during design and restored to the candidate set for identification. This tests whether an experiment distinguishes represented hypotheses. Protocol transfer and published-data prediction address complementary questions about the usefulness of the selected experiments.

Related work. EIG is an established criterion for discriminating mechanistic models (Lindley, 1956; Box & Hill, 1967; Cavagnaro et al., 2010); for a finite ensemble it is the BALD objective (Houlsby et al., 2011). Observation-aware experimental design has also compared brain models through a hemodynamic measurement model (Daunizeau et al., 2011). Our setting brings this principle to task-trained connectome models: the hypotheses share measured anatomy and similar task performance, while differing in the dynamics that a neural recording could resolve. The design objective values observations that separate those candidate explanations after the imaging process.

Neural experiment design has optimized stimuli and recording choices for receptive-field estimation, connectivity mapping, and population responses (Lewi et al., 2009; Shababo et al., 2013; Cowley et al., 2017; Wagenmaker et al., 2024). Several recent approaches use experiments to distinguish alternatives directly. Controversial stimuli separate competing models of perception (Golan et al., 2020); discriminative stimuli identify functional cell-type clusters (Burg et al., 2024); and optimized stimulus distributions distinguish hypotheses about probabilistic neural codes (Kuo & Walker, 2026). These studies motivate choosing inputs where alternatives make different predictions. Here the decision also includes which genetically accessible cell type to observe, and the relevant differences are those that survive uncertain calcium measurement. Our max-min discrepancy baseline provides a direct comparison with a strong disagreement-based rule.

Within connectome-constrained modeling, Beiran & Litwin-Kumar (2025) show how recordings from selected neurons can constrain predictions of unmeasured activity. We address a complementary decision as them: which cell-type–stimulus pair best identifies a member of an existing ensemble of task-trained networks? Joint selection matters because an informative neuron may require a particular stimulus to expose its distinguishing response. Analytic marginalization of gain and offset makes this search tractable, while genetic targeting and time budgets turn ranked designs into feasible protocols. Our contribution is this combination of measurement-aware joint design, data-calibrated evaluation, and protocol transfer. The benchmark separately tests stimulus selection, target selection, and acquisition objectives, making the source of each empirical gain explicit. For more extended comparisons, see Appendix K.

2 Measurement-aware experimental design

The design problem has three steps: predict what each network would produce in the microscope, score how well a recording would distinguish those predictions, and assemble high-scoring recordings into a feasible protocol. A *recording* denotes one cell-type–stimulus pair, repeating it reduces noise, while adding a different pair supplies a new experimental condition.

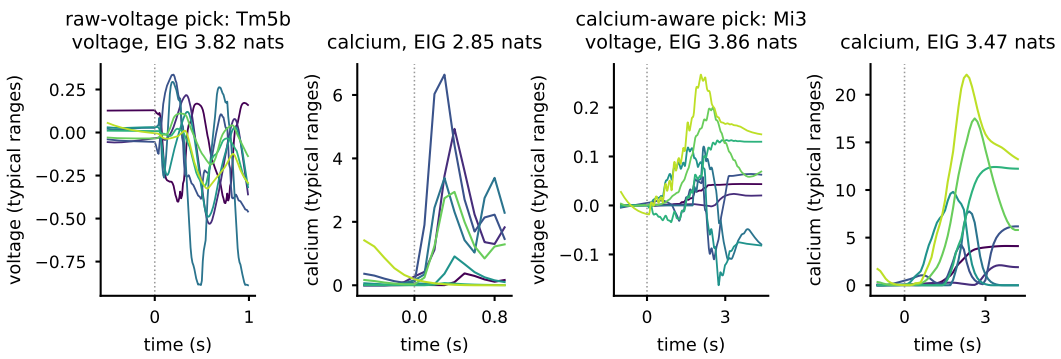

Figure 1: **Design for the measured response.** Eight network predictions for a voltage-selected grating on Tm5b (left) and a calcium-selected moving edge on Mi3 (right). Voltage-space EIG is similar (3.82 and 3.86 nats); after the calcium transform it is 2.85 and 3.47 nats. Each space is normalized separately, so vertical axes are relative ranges rather than comparable physical amplitudes. The designs last 0.9 and 4.2 s: the edge carries more information per recording, whereas the grating carries more per second. Schedule selection accounts for recording time.

2.1 From network predictions to candidate recordings

The `flyvis` ensemble contains 50 networks trained independently on the same optic-flow task (Lappalainen et al., 2024) - each has 45,669 neurons, approximately 1.5×10^6 synapses, and 65 cell types. In the network, connectivity is fixed; each network learns 734 parameters specifying time constants, resting potentials, and gains between connected cell types. We augment the published ensemble with two independently retrained ensembles of 21 and 20 networks for the primary 91-network benchmark. For any candidate set of size K, member identity k has prior $p(k) = 1/K$.

A design $d = (c, s)$ pairs a cell type c with one of 598 stimuli: flashes, impulses, moving edges, moving bars, drifting gratings, or naturalistic clips. The full catalog contains $65 \times 598 = 38{,}870$ designs. Network k predicts a central-column voltage trace $f_k(d)$, which we map to the measured space:

$$x_k(d) = \mathrm{DS}_{\delta t}(g(f_k(d)) * h_{\mathrm{ind}}) \in \mathbb{R}^T. \tag{1}$$

Here g is a voltage-to-calcium nonlinearity, h_{ind} describes indicator kinetics, and $\mathrm{DS}_{\delta t}$ averages camera frames. The primary transform rectifies relative to each network's resting level. We use GCaMP6f rise and decay times of 50 and 200 ms and a 10 Hz frame rate (Chen et al., 2013; Pirogova & Borst, 2023; Strother et al., 2017; Fisher et al., 2015). The filter starts at its pre-stimulus steady state, avoiding an artificial initialization transient. Predictions are baseline-relative, as in $\Delta F/F$ measurements.

Calibration from published recordings. We normalize each cell type by its median response range across networks and 60 sampled stimuli, excluding the generating network during design. Analysis of 313 published GCaMP6f recordings from 15 cell types (Drews et al., 2020) gives single-trial noise $\sigma_1 = 0.63$ in these units, an interquartile range of 0.40–1.23, and lag-one correlation 0.73. Averaging five trials gives $\sigma_5 = 0.28$. Both the simulated observations and the likelihood use this temporal correlation. This normalization assumes comparable relative signal quality across types; Appendix B specifies the constants and noise conventions.

2.2 Integrating calibration and transform uncertainty

The same predicted calcium response can yield different fluorescence scales across recordings. We model these differences with a gain a and offset b:

$$y = ax_k + b\mathbf{1} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, R), \quad a \sim \mathcal{N}(\mu_a, \tau_a^2), \quad b \sim \mathcal{N}(\mu_b, \tau_b^2). \tag{2}$$

For independent noise $R = \sigma^2 I_T$; the calibrated benchmark uses the AR(1) covariance. Integrating out a and b gives an exact likelihood under these Gaussian priors:

$$y \mid k, d \sim \mathcal{N}(\mu_a x_k + \mu_b \mathbf{1}, \Sigma_k), \qquad \Sigma_k = R + \tau_a^2 x_k x_k^\top + \tau_b^2 \mathbf{1}\mathbf{1}^\top. \tag{3}$$

The covariance has a rank-two update to R. Whitening and the Woodbury identity reduce each likelihood evaluation to $O(T)$ work (Appendix C). Bayes' rule then updates the probability of every candidate network.

We use $\mu_a = 1$ and $\mu_b = 0$. The controlled known-transform setting uses $\tau_a = 0.3$ and $\tau_b = 0.5$; the uncertain-transform benchmark widens the gain prior to $\tau_a = 1$ and generates nonnegative gains. Appendix B specifies the generative and inference priors. The published-data analysis uses $\tau_a = \tau_b = 1$.

The imaging nonlinearity is also uncertain. Let m index five transforms: rectification at simulator zero, three rectified power laws (exponents 1.5, 2, and 2.5), and rectification relative to rest. With a uniform prior,

$$p(y \mid k, d) = \frac{1}{5} \sum_{m=1}^{5} p(y \mid k, d, m). \tag{4}$$

The transform is shared across a fly's recordings, so multi-recording inference marginalizes it after combining their conditional likelihoods. In the primary uncertain-transform experiments, *selection uses the resting-level transform favored by the published data, while inference uses this mixture*. Fully mixture-based acquisition and other transform choices are evaluated in Appendix D.

2.3 Scoring and selecting experiments

We select the recording expected to reduce uncertainty about network identity most:

$$U(d) = I(\mathsf{k}; Y_d) = \mathbb{E}_{k\sim p,\, y\sim p(\cdot\mid k,d)} \left[\log p(y \mid k, d) - \log \sum_j p(j) p(y \mid j, d) \right]. \tag{5}$$

This EIG measures distinguishability in the recording space, including noise and calibration uncertainty (Lindley, 1956; Chaloner & Verdinelli, 1995). If every network predicts the same distribution of recordings, observing the response cannot change their relative probabilities and EIG is zero. A valuable design instead produces observations that favor different subsets of the ensemble, depending on which network generated them. For a uniform prior, the maximum possible gain is $\ln K$ nats: enough information to identify one of K candidates. Thus a strong response is useful only to the extent that its measured features distinguish hypotheses.

Every density in Eq. 5 is evaluated analytically. Monte Carlo estimates only the outer expectation, avoiding a nested likelihood estimator (Foster et al., 2019; Rainforth et al., 2024).

To search the catalog efficiently, we first compare predictions after their best positive-gain and offset alignment:

$$\Delta_{jk}(d) = \min_{a\geq 0, b} \frac{\|a x_j + b\mathbf{1} - x_k\|^2}{\sigma^2} = \frac{\|x_k^c\|^2 \{1 - (\rho_{jk}^+)^2\}}{\sigma^2}, \tag{6}$$

where x_k^c is centered and ρ_{jk}^+ is the correlation clamped at zero. Mixture-entropy scores built from these distances (Kolchinsky & Tracey, 2017) shortlist 1,200 designs for full EIG evaluation. The screen removes arbitrary positive scaling; the final likelihood retains the finite gain prior and member-dependent covariance. Appendix C evaluates estimator precision and screening accuracy.

2.4 From scores to feasible protocols

A split-GAL4 line restricts indicator expression to a cell type. Available lines isolate 40 of the 65 modeled types; a pan-T4/T5 line reaches eight additional subtypes jointly (Nern et al., 2025). The primary uncertain-transform first-recording benchmark searches these 48 accessible types, with a 40-type restriction tested separately. Multi-recording protocols use only the 40 individually targetable types and one line per fly.

Schedules allocate up to 30 minutes per fly (Turner et al., 2022; Fisher et al., 2015), charging five stimulus presentations plus inter-trial intervals per recording (Arenz et al., 2017; Salazar-Gatzimas et al., 2016). Lazy greedy selection ranks each addition by its conditional information gain per second (Minoux, 1978; Leskovec et al., 2007). With a fixed transform and conditionally independent

Table 1: **A single selected recording identifies the generating network more often than standard experimental choices.** Recovery (%) among 91 candidates. "Unknown" means that inference marginalizes five transforms; the simulated fly uses the resting-level transform. One recording contains either one trial or the average of five trials. Brackets give the 95% bootstrap interval of ***EIG's paired margin***, in percentage points. Stars mark Holm-adjusted $p < 0.05$; all margins over the PageRank and T5b baselines have $p < 0.01$.

	Transform unknown		Transform known	
Selection rule	1 trial	5 trials	1 trial	5 trials
EIG	**63.7**	**80.8**	**81.2**	**94.2**
Max-min discrepancy	61.1 $[-3, 8]$	78.8 $[-2, 5]$	65.7* $[9, 21]$	88.1* $[3, 9]$
PageRank target	51.6* $[5, 19]$	70.1* $[5, 17]$	66.2* $[7, 22]$	85.7* $[4, 13]$
T5b preferred edge	50.8* $[5, 21]$	63.9* $[10, 24]$	60.2* $[12, 30]$	76.6* $[10, 25]$
Standard battery†	37.1* $[20, 34]$	55.4* $[19, 32]$	44.7* $[29, 44]$	67.7* $[20, 34]$
Random recording	24.9* $[30, 47]$	37.2* $[36, 51]$	27.2* $[46, 61]$	42.3* $[44, 60]$

†First recording only, not the complete battery.

recordings, information gain is monotone submodular (Krause & Guestrin, 2005). Shared transform uncertainty and per-fly targeting constraints require separate treatment; we use the procedure as a practical schedule heuristic and check diminishing returns empirically (Appendix C).

3 One recording distinguishes candidate networks

Evaluation. Our primary benchmark pools the 50 published networks and 41 independently retrained networks, and each serves in turn as the simulated fly. Design selection excludes the generating network - identification includes all 91 candidates. Success means that the generating network is the unique posterior mode. We average 16 noise and nuisance draws per generating network, compare rules with paired sign-flip tests over networks, and apply Holm correction within each condition. Here, confidence intervals bootstrap generating networks.

The unknown-transform condition generates responses with the resting-level transform, widens gain uncertainty, and marginalizes the five transforms at inference. It differs from the supporting 50-network experiment that draws the generating transform from a posterior-weighted mixture. Also, the known-transform condition provides a controlled comparison with reduced gain uncertainty ($\tau_a = 0.3$) and all 65 types eligible. Offset uncertainty is $\tau_b = 0.5$ in both conditions. Therefore, the two conditions vary more than transform knowledge alone. Appendix D gives the complete configurations, selection seeds, and 50-network results.

Baselines. We compare against a preferred-direction OFF edge on T5b, a conventional motion experiment (Maisak et al., 2013), the highest-PageRank target with standard stimuli, the first recording of a standard flash-and-edge battery, beginning with a randomly chosen flash on a routinely recorded type, and a random recording. Max-min discrepancy chooses the design maximizing the smallest pairwise separation in Eq. 6, providing a strong model-based comparator motivated by controversial-stimulus design (Golan et al., 2020). All rules use the same inference likelihood within a condition. See Appendix D for comparisons on expected posterior accuracy, expected MAP accuracy, and tempered EIG.

These comparisons test distinct experimental choices. The T5b baseline asks whether design improves on a targeted motion experiment, and PageRank compares against a structurally prominent target paired with standard stimuli. The generic-edge control below holds EIG's target fixed and changes only its stimulus. Because inference is shared, differences in recovery reflect which observation was acquired. The max-min comparison then asks how much the acquisition objective matters once both rules use model predictions.

Identification at measured noise. With the transform unknown at inference, EIG identifies the generating network in 63.7% of single-trial and 80.8% of five-trial recordings (Table 1). The gains over the T5b preferred edge are 12.9 and 16.9 percentage points - over the PageRank target they are

12.1 and 10.7 points. Both comparisons remain significant after multiplicity correction. With the transform known, recovery reaches 81.2% and 94.2%, and the gains over these two baselines span approximately 8–21 points. Thus the improvement extends beyond random or broadly sampled experiments to targeted conventional choices.

EIG and max-min discrepancy perform similarly when the transform is unknown - when it is known, EIG leads max-min by 15.5 and approximately 6 points. These results support measurement-aware selection as a family of useful strategies - see more in Appendix D about the conditions in which their objectives differ.

Selecting the stimulus adds value beyond selecting the neuron. In the same 91-network single-trial benchmark, a generic moving edge on EIG's selected target recovers 51.4% with the transform unknown, compared with 63.7% for the jointly selected stimulus and target, while with the transform known, the corresponding values are 70.2% and 81.2%. The 12.3- and 11.0-point gains show why neuron selection alone is insufficient. Selecting the largest predicted response reaches only 39.6% and 8.7%, respectively: strong activation need not distinguish hypotheses. These controls use the same inference procedure, isolating the value of the recording choice (each Holm $p \leq 0.002$).

Measurement modeling guides the choice of experiment. The calcium transform changes which differences remain observable (Fig. 1). In the supporting 50-network benchmark, calcium-aware EIG exceeds voltage-space EIG by 14.4 points under posterior-weighted transform uncertainty. Many voltage objectives lie near their $\ln K$ information ceiling, and modeling the indicator reduces this saturation. Tempering the voltage objective recovers part of the gain, and simple discrepancy-based objectives can also be competitive. Full ablations and matched-time comparisons appear in Appendix D.

4 Informative targets and transferable protocols

The identification benchmark establishes the value of selecting a recording. We next ask where the information lies across the visual system, and whether those choices can be turned into a reusable protocol. These analyses use the original ensemble for design, with retrained ensembles reserved for the transfer test.

An atlas of experimental informativeness. We evaluate the full catalog on the original 50-network ensemble at $\sigma = 1.0$, where differences between designs remain visible below the information ceiling. The median design carries 1.3 nats, standard flashes and edges carry 2.0, and the best design for a cell type carries about 3. The best achievable EIG varies 1.6-fold between the 5th and 95th percentiles of cell types and 3.1-fold across the full range. A moving edge is the best stimulus for 61 of 65 types, but its parameters must be chosen for the target. Therefore, the atlas provides a map of *cell-type–stimulus combinations* and not just simply a list of neurons to record.

A frozen protocol transfers across retrainings. The schedule selects three individually targetable types—TmY4, Tm5b, and Mi10—with one genetic line per fly (Appendix E). We freeze the targets and stimuli, then apply the protocol unchanged to two independently retrained ensembles containing 21 and 20 networks. Each retrained ensemble supplies its own hypotheses at inference. This tests transfer of the *experimental design*, separately from identifying a network outside the hypothesis set.

The first selected recording identifies 76% of these 41 generating networks from a single trial and 90% from five. Relative to standard stimuli on the same targets, the paired gains are 9.5 points [2.9,16.1] and 6.2 points [0.8,11.8]. Recovery reaches 90% after 65 seconds of scheduled recording, compared with 80–120 seconds for the standard alternatives (Fig. 3). These times charge five presentations and intervals per recording and exclude preparation time. The comparison isolates how the selected sequence uses the available recording budget.

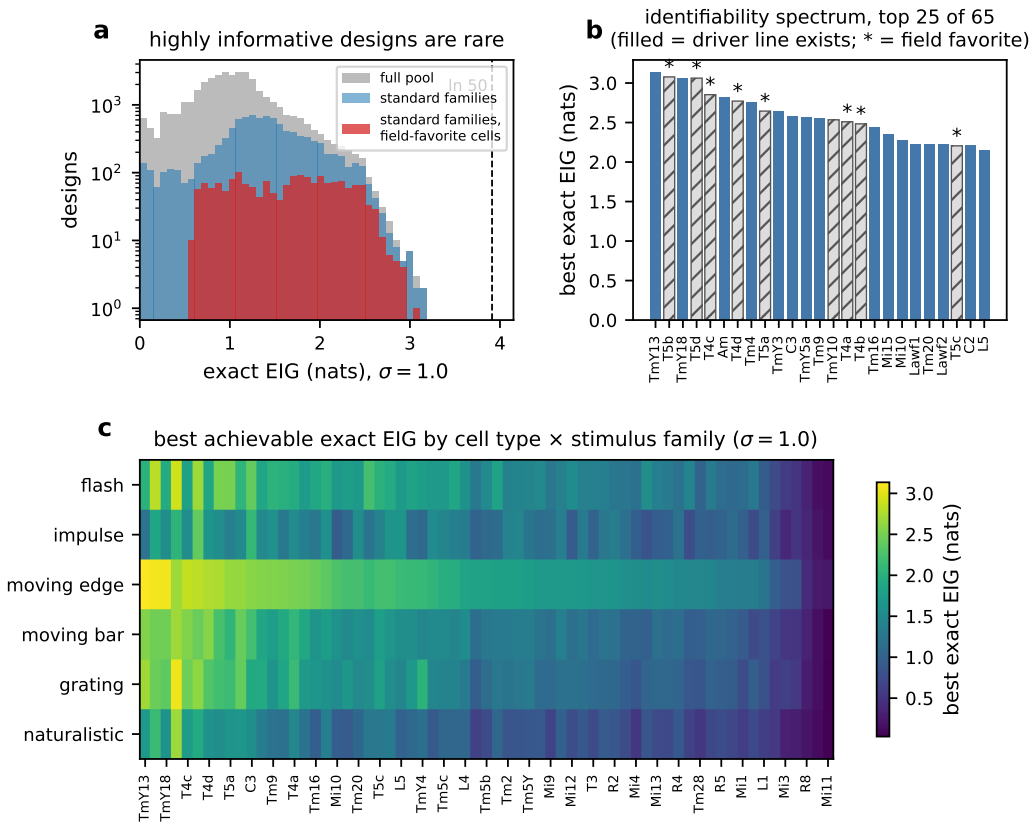

Figure 2: **The joint design space reveals informative recording choices.** Full-estimator EIG across 38,870 designs at $\sigma = 1.0$ in the 50-network ensemble. **(a)** Distribution across all designs, standard stimulus families, and standard families on routinely recorded types. **(b)** The 25 highest-ranked cell types by best-design EIG; filled markers indicate driver availability and stars indicate routinely recorded types (Currier et al., 2023). **(c)** Best EIG for each cell type and stimulus family. This atlas uses a higher noise setting to expose differences that approach the ceiling at five-trial noise.

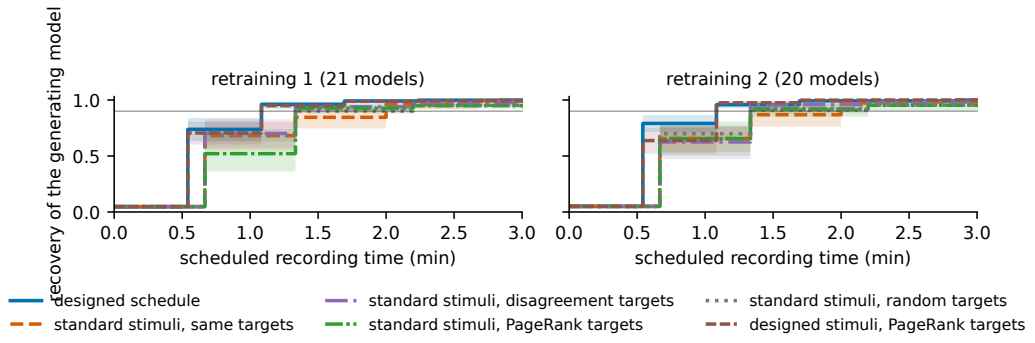

Figure 3: **A frozen protocol transfers to retrained ensembles.** Identification as scheduled recording time accumulates in two independent retrainings. Each ensemble is its own candidate set. Recovery is scored at single-trial noise; time charges five trials plus intervals. Controls vary targets or substitute a standard flash-and-edge battery. Bands bootstrap generating networks; the horizontal line marks 90% recovery. Preparation time is excluded. Appendix E reports paired comparisons at matched budgets.

5 SELECTED CONDITIONS PREDICT PUBLISHED RESPONSES

Protocol transfer tests whether recording choices generalize across trained ensembles. Published recordings let us ask whether selected observations also help predict biological responses to stimuli that were not used for conditioning. We evaluate this retrospectively on calcium recordings from two laboratories, keeping acquisition conditions separate from the conditions used to score prediction (Appendix I).

Contrast responses across 15 cell types. On 313 recordings under a 42-condition contrast protocol (Drews et al., 2020), one EIG-selected condition improves held-out R^2 over a random condition by 0.057 [0.045,0.068]. A standard contrast battery and voltage-selected conditions perform similarly to calcium-aware selection, coming within 0.007 and 0.004 R^2, respectively, at every tested count. These data establish value relative to random acquisition while showing that a well-chosen conventional battery can already be informative.

Generalization across grating conditions. We next select among half of a 36-condition grating grid and predict responses in the other half (Ramos-Traslosheros & Silies, 2021). For T5, EIG-selected gratings outperform random choices at every acquisition count; a standard frequency sweep leads at one grating and performs similarly from two onward. For Tm9, selected gratings improve held-out R^2 over the sweep by 0.065–0.111 with two to six observations. T5 and Tm9 were recorded in the same flies and provide complementary cell-type results within one dataset.

The benefit of calcium-aware scoring is cell-type dependent: it improves Tm9 prediction relative to voltage selection, whereas voltage selection performs better for T5 at several budgets. Accordingly, the biological endpoint is the comparison between acquisition rules, rather than absolute fit or recovery of a uniquely correct model. Figure 4 shows the learning curves; Appendix I provides uncertainty estimates and fit diagnostics; Appendix H reports a separate evaluation on 110 published tuning records.

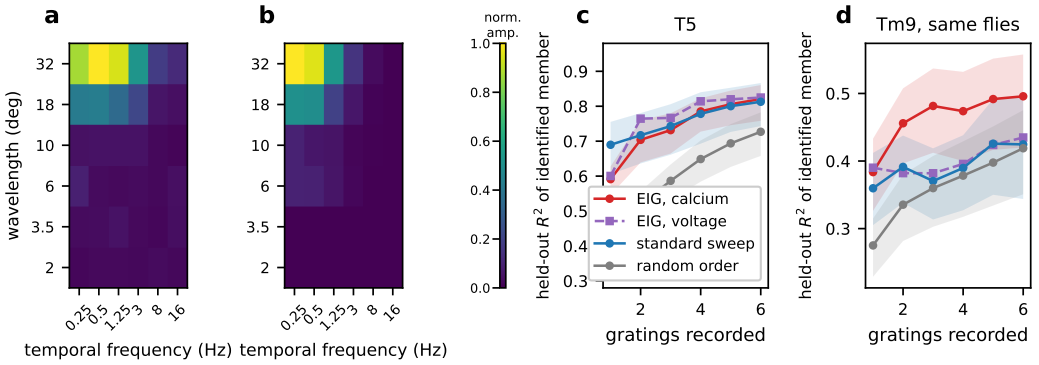

Figure 4: **Selected gratings predict unobserved biological responses. (a,b)** Recorded and predicted T5 spatiotemporal response maps. **(c,d)** Held-out R^2 after conditioning on selected gratings for T5 and Tm9. EIG-selected conditions outperform random choices in both cell types; the gain over a standard sweep depends on cell type and acquisition budget. Acquisition uses $\sigma = 0.1$. Bands summarize ROI resampling; Appendix I specifies the uncertainty calculation. The two cell types were recorded in the same flies.

6 DISCUSSION

Connectome-constrained learning produces testable alternatives about circuit dynamics. **This work makes those alternatives actionable**: it assigns value to a recording in the space the microscope observes, searches jointly over cell types and stimuli, and produces a protocol constrained by available genetic lines and recording time. The 91-network benchmark shows that a single selected recording can substantially outperform targeted conventional choices. The frozen-protocol and retrospective analyses then show that these choices remain useful across retraining and on measured responses.

From model differences to observable differences. The measurement model specifies which aspects of disagreement an experiment can resolve. Indicator kinetics and frame averaging determine which temporal differences survive, while gain and offset uncertainty determine how much to trust amplitude and baseline differences. Integrating these effects into the likelihood gives a common basis for comparing otherwise very different recordings. The joint-selection controls isolate a second source of value: even after choosing an informative cell type, choosing its stimulus improves identification. Together, these results make the observation process and the experimental choice central parts of model evaluation.

Reading the atlas as an experimental guide. The atlas and the schedule answer complementary questions. Moving edges maximize information per recording for most cell types, whereas the frozen schedule opens with gratings because their shorter duration gives more information per second (Appendix E). An experiment's value therefore depends on both its target and the available recording budget. The atlas can guide the choice of genetic line before a preparation is made; within that preparation, the schedule prioritizes stimuli that distinguish the hypotheses. This connects a survey of model disagreement to the decisions an experimenter can make.

Toward prospective model refinement. A prospective study could begin with the frozen targets and stimuli, calibrate the observation model using pilot recordings, and evaluate the resulting posterior on a predeclared set of held-out conditions. The framework also permits later stimulus choices to use the updated distribution over networks, concentrating measurements on unresolved alternatives. Transfer across independent retrainings supports testing the selected protocol beyond the particular optimization runs used to construct it. The retrospective analyses supply the complementary biological endpoint: whether acquired observations improve prediction of responses that have yet to be measured. A prospective experiment can therefore assess both how much the models are separated and what is learned about the circuit.

The framework also makes the scientific target explicit. Identifying an ensemble member, distinguishing a wiring hypothesis, and predicting an unmeasured response are different objectives. The same likelihood can support class-level design by replacing member information with information about a wiring class; Appendix G evaluates this extension. More broadly, the approach provides a concrete interface between task-trained circuit models and the experiments needed to refine them.

For a structural question, several independently trained networks can represent each candidate wiring hypothesis. Marginalizing over those networks treats variation across training runs as uncertainty within the hypothesis. The acquisition objective can then prioritize recordings that separate wiring hypotheses, even when substantial differences remain among networks sharing the same wiring. This gives the experimental question a direct role in defining useful information: resolving parameter variability and testing a structural explanation can lead to different recording choices.

The framework's components can also change with the experimental setting. An updated connectome changes the candidate predictions; a different indicator changes the observation model; new driver lines change the feasible recording set. These changes can enter the same design calculation, provided that the corresponding predictions and measurement likelihood are available. The fly visual system supplies a concrete test of this approach at scale. Its broader use is as a repeatable process for translating increasingly detailed circuit models into increasingly specific experimental questions.

Scope and next steps. Identification assumes the generating network is represented in the inference ensemble. Separate out-of-ensemble tests show that confident identification need not improve prediction over standard acquisition (Appendix J). Observation-model adequacy also matters: the robustness analyses examine indicator kinetics, saturation, and response-dependent noise, including repairs through the likelihood and acquisition objective (Appendix D). The biological analyses are retrospective. Prospectively testing the frozen protocol is the next step toward using model disagreement to guide experimental neuroscience.

Connectome-constrained models can thus contribute both predictions of neural activity and explicit proposals for the measurements that would most effectively refine those predictions.

REPRODUCIBILITY STATEMENT

The models, connectome, and driver-line annotations are public data. The code, the measurement corpus, and every result file cited in the appendices are provided as supplementary material, with the likelihood implementation, validation tests, experiment drivers, and figure scripts. Appendix B specifies every observation-model constant and its source, Appendix D the benchmark, Appendix E the protocol transfer, Appendix H the biological prediction analysis, and Appendix I the calcium time-series analysis. Building the response cache takes about 40 minutes on one RTX Blackwell 6000, and training each perturbed-connectome network takes about 2.9 GPU-hours. With the caches in place, the leave-one-out benchmark takes 10 to 40 minutes per condition and seed on one GPU, and the biological, wiring-class, and protocol-transfer analyses a few minutes each on four CPU threads.

STATEMENT ON GENERATIVE AI USE

Generative AI tools (specifically, those available from both Anthropic and OpenAI)) assisted with literature review, experimental design, code implementation, analysis, figure preparation, and manuscript revision. AI-assisted code was checked against reference calculations, budget and data-separation assertions, and the repository's validation tests. Experiment configurations and numerical outputs are included with the code. We, the authors authors are responsible for the scientific claims, citations, and final submission.

REFERENCES

A. Arenz, M. S. Drews, F. G. Richter, G. Ammer, and A. Borst. The temporal tuning of the Drosophila motion detectors is determined by the dynamics of their input elements. *Current Biology*, 27(7):929–944, 2017. doi: 10.1016/j.cub.2017.01.051.

A. C. Atkinson and V. V. Fedorov. The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70, 1975.

M. Beiran and A. Litwin-Kumar. Prediction of neural activity in connectome-constrained recurrent networks. *Nature Neuroscience*, 28(12):2561–2574, 2025. doi: 10.1038/s41593-025-02080-4.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented Bayesian active learning. In *International Conference on Artificial Intelligence and Statistics*, 2023.

George E. P. Box and William J. Hill. Discrimination among mechanistic models. *Technometrics*, 9 (1):57–71, 1967.

M. F. Burg, T. Zenkel, M. Vystrčilová, J. Oesterle, L. Höfling, K. F. Willeke, J. Lause, S. Müller, P. G. Fahey, Z. Ding, K. Restivo, S. Sridhar, T. Gollisch, P. Berens, A. S. Tolias, T. Euler, M. Bethge, and A. S. Ecker. Most discriminative stimuli for functional cell type clustering. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024.

Daniel R. Cavagnaro, Jay I. Myung, Mark A. Pitt, and Janne V. Kujala. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, 22(4):887–905, 2010.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3): 273–304, 1995. doi: 10.1214/ss/1177009939.

T.-W. Chen, T. J. Wardill, Y. Sun, S. R. Pulver, S. L. Renninger, A. Baohan, E. R. Schreiter, R. A. Kerr, M. B. Orger, V. Jayaraman, L. L. Looger, K. Svoboda, and D. S. Kim. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300, 2013. doi: 10.1038/nature12354.

B. R. Cowley, R. C. Williamson, K. Acar, M. A. Smith, and B. M. Yu. Adaptive stimulus selection for optimizing neural population responses. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

T. A. Currier, M. M. Pang, and T. R. Clandinin. Visual processing in the fly, from photoreceptors to behavior. *Genetics*, 224(2):iyad064, 2023. doi: 10.1093/genetics/iyad064.

Jean Daunizeau, Kerstin Preuschoff, Karl Friston, and Klaas Stephan. Optimizing experimental design for comparing models of brain function. *PLoS Computational Biology*, 7(11):e1002280, 2011.

C. DiMattina and K. Zhang. Active data collection for efficient estimation and comparison of nonlinear neural models. *Neural Computation*, 23(9):2242–2288, 2011. doi: 10.1162/NECO_a_00167.

Michael S. Drews, Aljoscha Leonhardt, Nadezhda Pirogova, Florian G. Richter, Anna Schuetzenberger, Lukas Braun, Etienne Serbe, and Alexander Borst. Dynamic signal compression for robust motion vision in flies. *Current Biology*, 30(2):209–221.e8, 2020. doi: 10.1016/j.cub.2019.10.035.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2021.

Y. E. Fisher, J. C. S. Leong, K. Sporar, M. D. Ketkar, D. M. Gohl, T. R. Clandinin, and M. Silies. A class of visual neurons with wide-field properties is required for local motion detection. *Current Biology*, 25:3178–3189, 2015. doi: 10.1016/j.cub.2015.11.018.

A. Foster, M. Jankowiak, E. Bingham, P. Horsfall, Y. W. Teh, T. Rainforth, and N. Goodman. Variational Bayesian optimal experimental design. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

A. Foster, D. R. Ivanova, I. Malik, and T. Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3384–3395. PMLR, 2021.

T. Golan, P. C. Raju, and N. Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117 (47):29330–29337, 2020. doi: 10.1073/pnas.1912334117.

L. N. Groschner, J. G. Malis, B. Zuidinga, and A. Borst. A biophysical account of multiplication by a single neuron. *Nature*, 603:119–123, 2022. doi: 10.1038/s41586-022-04428-3.

N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. arXiv:1112.5745, 2011.

Desi R. Ivanova, Adam Foster, Steven Kleinegesse, Michael U. Gutmann, and Tom Rainforth. Implicit deep adaptive design: Policy-based experimental design without likelihoods. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Madhura D. Ketkar, Burak Gür, Sebastian Molina-Obando, Maria Ioannidou, Carlotta Martelli, and Marion Silies. First-order visual interneurons distribute distinct contrast and luminance information across ON and OFF pathways to achieve stable behavior. *eLife*, 11:e74937, 2022. doi: 10.7554/eLife.74937.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. Kolchinsky and B. D. Tracey. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017. doi: 10.3390/e19070361. See Correction: Entropy 2017, 19(11), 588, doi:10.3390/e19110588; corrected version at arXiv:1706.02419v4.

A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI 2005)*, pp. 324–331. AUAI Press, 2005.

P.-C. Kuo and E. Y. Walker. An information-theoretic framework for optimizing experimental design to distinguish probabilistic neural codes. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 2026.

J. K. Lappalainen, F. D. Tschopp, S. Prakhya, M. McGill, A. Nern, K. Shinomiya, S. Takemura, E. Gruntman, J. H. Macke, and S. C. Turaga. Connectome-constrained networks predict neural activity across the fly visual system. *Nature*, 634(8036):1132–1140, 2024. doi: 10.1038/s41586-024-07939-3.

J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pp. 420–429, 2007. doi: 10.1145/1281192.1281239.

J. Lewi, R. Butera, and L. Paninski. Sequential optimal design of neurophysiology experiments. *Neural Computation*, 21(3):619–687, 2009. doi: 10.1162/neco.2008.08-07-594.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069.

A. Litwin-Kumar and S. C. Turaga. Constraining computational models using electron microscopy wiring diagrams. *Current Opinion in Neurobiology*, 58:94–100, 2019. doi: 10.1016/j.conb.2019.07.007.

M. S. Maisak, J. Haag, G. Ammer, E. Serbe, M. Meier, A. Leonhardt, T. Schilling, A. Bahl, G. M. Rubin, A. Nern, B. J. Dickson, D. F. Reiff, E. Hopp, and A. Borst. A directional tuning map of Drosophila elementary motion detectors. *Nature*, 500(7461):212–216, 2013. doi: 10.1038/nature12320.

M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In J. Stoer (ed.), *Optimization Techniques: Proceedings of the 8th IFIP Conference on Optimization Techniques, Würzburg, 1977*, volume 7 of *Lecture Notes in Control and Information Sciences*, pp. 234–243. Springer, 1978. doi: 10.1007/BFb0006528.

A. Mishra, E. Serbe-Kamp, A. Borst, and J. Haag. Voltage to calcium transformation enhances direction selectivity in Drosophila T4 neurons. *Journal of Neuroscience*, 43(14):2497–2514, 2023. doi: 10.1523/JNEUROSCI.2297-22.2023.

A. Nern, F. Loesche, S.-y. Takemura, L. E. Burnett, M. Dreher, et al. Connectome-driven neural inventory of a complete visual system. *Nature*, 641(8065):1225–1237, 2025. doi: 10.1038/s41586-025-08746-0.

Antony M. Overstall and James M. McGree. Bayesian decision-theoretic design of experiments under an alternative model. *Bayesian Analysis*, 17(4), 2022.

N. Pirogova and A. Borst. Contrast normalization affects response time-course of visual interneurons. *PLOS ONE*, 18(6):e0285686, 2023. doi: 10.1371/journal.pone.0285686.

T. Rainforth, A. Foster, D. R. Ivanova, and F. Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. doi: 10.1214/23-STS915.

Giordano Ramos-Traslosheros and Marion Silies. The physiological basis for contrast opponency in motion computation in Drosophila. *Nature Communications*, 12:4987, 2021. doi: 10.1038/s41467-021-24986-w.

P. Rupprecht, S. Carta, A. Hoffmann, M. Sandhu, J. Zhang, J. Nakai, K. Bengtsson Bernander, R. W. Friedrich, and F. Helmchen. A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. *Nature Neuroscience*, 24:1324–1337, 2021. doi: 10.1038/s41593-021-00895-5.

E. Salazar-Gatzimas, J. Chen, M. S. Creamer, O. Mano, H. B. Mandel, C. A. Matulis, J. Pottackal, and D. A. Clark. Direct measurement of correlation responses in Drosophila elementary motion detectors reveals fast timescale tuning. *Neuron*, 92(1):227–239, 2016. doi: 10.1016/j.neuron.2016.09.017.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT '92)*, pp. 287–294. ACM, 1992. doi: 10.1145/130385.130417.

Ben Shababo, Brooks Paige, Ari Pakman, and Liam Paninski. Bayesian inference and online experimental design for mapping neural microcircuits. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Sabina J. Sloman, Daniel M. Oppenheimer, Stephen B. Broomell, and Cosma Rohilla Shalizi. Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias, 2022.

J. A. Strother, S.-T. Wu, A. M. Wong, A. Nern, E. M. Rogers, J. Q. Le, G. M. Rubin, and M. B. Reiser. The emergence of directional selectivity in the visual motion pathway of Drosophila. *Neuron*, 94(1):168–182.e10, 2017. doi: 10.1016/j.neuron.2017.03.010.

M. H. Turner, A. Krieger, M. M. Pang, and T. R. Clandinin. Visual and motor signatures of locomotion dynamically shape a population code for feature detection in Drosophila. *eLife*, 11:e82587, 2022. doi: 10.7554/eLife.82587.

A. Wagenmaker, L. Mi, M. Rózsa, M. S. Bull, K. Svoboda, K. Daie, M. D. Golub, and K. Jamieson. Active learning of neural population dynamics using two-photon holographic optogenetics. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

E. Y. Walker, F. H. Sinz, E. Cobos, T. Muhammad, E. Froudarakis, P. G. Fahey, A. S. Ecker, J. Reimer, X. Pitkow, and A. S. Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 22(12):2060–2065, 2019. doi: 10.1038/s41593-019-0517-x.

H. H. Yang, F. St-Pierre, X. Sun, X. Ding, M. Z. Lin, and T. R. Clandinin. Subcellular imaging of voltage and calcium signals reveals neural processing in vivo. *Cell*, 166(1):245–257, 2016. doi: 10.1016/j.cell.2016.05.031.

Table 2: **Terms used throughout.** The field's default is the T5b preferred-edge baseline; the routine battery is the standard battery, evaluated after its first recording in the recovery tables.

Resting-level transform	Rectification relative to each member's resting level, the transform the published traces favor (posterior 0.62); GCaMP6f kinetics (50/200 ms) started at steady state, 10-Hz frames
Zero transform	Rectification at the simulator's zero with the pre-stimulus level subtracted (sensitivity analysis)
Single-trial, five-trial noise	$\sigma = 0.63$ and 0.28 with lag-one correlation 0.73 in the simulated fly and in the likelihood
Model-based objectives	EIG; expected accuracy; expected MAP accuracy; max-min discrepancy; tempered EIG (Section 2.3)
Standard stimuli	Full-field flashes and moving edges
Field's default experiment	T5b, then T4c and the other T4/T5 subtypes, each with the preferred-direction edge of its polarity first (OFF for T5, ON for T4)
PageRank target	Tm5b, the type with the highest connectome PageRank, with standard stimuli
Disagreement	The recording with the largest mean pairwise distance between members' calcium traces
Routine practice	The routine battery and the field's default experiment
Routine battery	Standard stimuli on the twelve types most often recorded (T4a–d, T5a–d, Mi1, Tm3, L1, L2), flashes before edges, types in random order
Random recording	One design drawn uniformly from the candidate pool: 598 stimuli $\times$ 65 types (38,870) with the transform known, $\times$ 48 driver-accessible types (28,704) in the realistic benchmark

A Ensemble degeneracy

The 50 trained members reach closely matched task losses (5.14 to 5.68) yet disagree qualitatively: for 32 of the 65 cell types at least 10 members respond to a full-field ON or OFF flash with the sign opposite to the others (31 for a local flash), and task error orders the members only weakly by agreement with published tuning (Spearman -0.40). This spread is the design problem.

B Observation model and noise levels

Each simulated voltage trace is rectified relative to the member's resting level, $g(v) = \max(v - v_{\text{rest}}, 0)$ with v_{rest} the mean of the first 0.5 s before the stimulus, filtered with the causal indicator kernel $h(t) \propto (1 - e^{-t/\tau_{\text{rise}}})e^{-t/\tau_{\text{decay}}}$ (sampled every 10 ms, truncated after five decay constants, normalized to unit sum) with the filter at the steady state of the first sample, and averaged into 10-Hz frames; Eq. 2 then adds the recording's gain, offset, and noise. Starting the filter from zero instead adds a warm-up transient whose size depends on each member's resting level; it makes short stimuli look uninformative and lets the resting level, which no $\Delta F/F$ recording reports, identify members. With the steady start, the 1 s before the stimulus carries a median 0.50 nats, 14% of the information in EIG's first single-trial choices, so pre-stimulus settling does not drive identification. The zero transform rectifies at the simulator's zero, $g(v) = \max(v, 0)$, and subtracts each trace's first frame so that the resting calcium level is not observed; without the subtraction the resting level would stay in the trace as a constant that fixes each recording's gain. Table 3 lists every constant and its source. Effective cell-level filters can be slower than the GCaMP6f kinetics used here, and voltage-to-$\Delta F/F$ mappings in these non-spiking cells are rectified and supralinear with exponents of 1.3–2.5 (Yang et al., 2016; Mishra et al., 2023); both are tested below.

C Likelihood, estimator, and selection checks

Marginal likelihood in $O(T)$. With $U_k = [x_k, \mathbf{1}]$ and $D = \text{diag}(\tau_a^2, \tau_b^2)$, the covariance of Eq. 3 is $\Sigma_k = \sigma^2 I + U_k D U_k^\top$, and $A_k = D^{-1} + U_k^\top U_k/\sigma^2$,

$$r^\top \Sigma_k^{-1} r = \min_{a,b} \frac{\|y - a x_k - b\mathbf{1}\|^2}{\sigma^2} + \frac{(a-\mu_a)^2}{\tau_a^2} + \frac{(b-\mu_b)^2}{\tau_b^2}, \tag{7}$$

Table 3: Observation-model defaults and sources.

Quantity	Default	Source
Indicator rise / decay τ	50 / 200 ms	Chen et al. (2013); Pirogova & Borst (2023)
Imaging frame rate	10 Hz	Strother et al. (2017); Fisher et al. (2015)
Single-trial noise σ_1	0.63 of the typical response range (median), lag-one correlation 0.73	313 published traces, Appendix I
Trials per stimulus	5, 5 s apart	Arenz et al. (2017); Strother et al. (2017); Salazar-Gatzimas et al. (2016)
Session budget per fly	30 min	Turner et al. (2022); Fisher et al. (2015)
Nuisance widths (τ_a, τ_b)	$(0.3, 0.5)$ with the transform known; $(1.0, 0.5)$ in the realistic benchmark; $(1.0, 1.0)$ on real data	modeling choices; see text

Table 4: **Noise levels used in the paper.** AR(1) gives the lag-one noise correlation used in the simulated fly and in the likelihood; all values are in units of each cell type's typical response range.

Name	Meaning	σ	AR(1)	Where used
Single trial	median noise of 313 published traces in model units	0.63	0.73	Sections 3, 4
Five trials	single-trial noise averaged over five trials (Rupprecht et al., 2021)	0.28	0.73	Sections 3, 4
Quartiles	interquartile range of the published single-trial noise	0.40, 1.23	0.73	App. D
Noise sweep	independent noise from below to above the measured levels	0.30 to 1.5	–	App. D
Atlas and selection	level at which 50 hypotheses leave room below $\ln 50$; atlas, schedule selection, wiring classes, prediction outside the ensemble	1.0	–	Section 4, App. G, J
Low noise	many averaged trials; wiring classes, prediction outside the ensemble	0.10	–	App. G, J
Gratings	deposited amplitudes averaged over trials and periods (sensitivity 0.05 and 0.2)	0.10	–	App. I

with $r = y - \mu_a x_k - \mu_b \mathbf{1}$, the minimizing gain and offset solving a 2×2 system with matrix A_k, and $\log\det\Sigma_k = T\log\sigma^2 + \log\det D + \log\det A_k$. We evaluate this sum of non-negative terms rather than the textbook form $\|r\|^2/\sigma^2 - (U_k^\top r)^\top A_k^{-1} U_k^\top r/\sigma^4$, which cancels catastrophically when a member's trace exceeds the observation by many orders of magnitude, as for runaway members under power-law transforms. Gram terms are precomputed per design, so a measurement costs two inner products per member; randomized tests match dense evaluation to 10^{-7}. Correlated AR(1) noise is handled by whitening y, the traces, and the offset direction with the banded inverse Cholesky factor, after which the same form applies. The profiled discrepancy of Eq. 6 is the residual after unpenalized gain and offset alignment, with the gain clamped at zero for anticorrelated traces.

Screening. Candidates are screened with two Kolchinsky–Tracey mixture-entropy scores built from the symmetrized profiled discrepancies, $\underline{U}(d) = -\sum_i p(i)\log\sum_j p(j)\, e^{-\bar{\Delta}_{ij}/8}$ and the analogous score with $e^{-\bar{\Delta}_{ij}/2}$; the 1,200 best candidates are re-scored with the exact-integrand Monte Carlo estimator ($S = 256$ Monte Carlo draws of member and observation in total, shared by all candidates, for selection; 1024 for the atlas). On the atlas's full-estimator EIG at $\sigma = 1.0$, the screen's 1,200-design shortlist contains the 100 best of the 38,870 designs and 93% of the best 1,000. At $S = 256$ the Monte Carlo standard error of the top design's EIG is 0.02 to 0.06 nats (eight held-out flies, five estimator seeds each). A median of 6 shortlisted designs lie within two standard errors of the calcium-aware top design, against 20 for the voltage objective, which sits closer to the $\ln 49$ ceiling (3.74 against 3.51 nats); the selected design therefore varies with the estimator seed, while recovery does not (Appendix D).

Submodularity. With the transform fixed and recording-specific nuisance variables and noise independent across recordings, recordings are conditionally independent given the member. Batch EIG is then monotone submodular. We use lazy greedy selection; the per-fly driver-line constraints

do not inherit the usual unconstrained guarantee. Across 150 comparisons on 25 random chains at single-trial noise, four violate diminishing returns, by a mean of 0.0004 nats, within Monte Carlo error. When an unknown transform is shared across a fly's recordings and marginalized, conditional independence given the member alone no longer holds, so the submodularity argument does not apply. In the empirical check, the mixture objective violates diminishing returns in 9 of 480 comparisons, with mean violation 0.003 nats.

Calibration and robustness. Under the assumed model the randomized probability integral transform of the posterior is consistent with uniformity for the top design (Kolmogorov–Smirnov statistic 0.018; expected calibration error 0.0015). Rankings of the candidate designs by the screen, computed from 10, 20, and 35 members, correlate with the 50-member ranking at Spearman 0.89, 0.94, and 0.97, and tempering the member prior by task error keeps them at 0.98 and 0.84 for tempering factors 1 and 3.

D Recovery benchmark details

Each of the 50 members serves in turn as the fly; screening, selection, and normalization use the other 49; identification considers all 50; eight noise and nuisance draws per member give 400 trials per rule and condition, and a trial counts as recovered when the generating member is the unique posterior mode. Paired comparisons use a sign-flip permutation test over the 50 generating members (20,000 draws) and bootstrap intervals over members, resampling selection seeds as well when several are pooled; only EIG and its variants depend on the selection seed.

With the transform known. Over the 50 published members one EIG-selected recording recovers 83.3% from a single trial and 96.7% from five (five selection seeds), against 48.0 and 70.0% for the routine battery and 32.2 and 45.2% at random (each $p < 0.001$); the field's default experiment, T5b's preferred-direction OFF edge first (Maisak et al., 2013), reaches 68.5 and 79.5%, and the PageRank target comes within 5 points (not significant). Both gaps are resolved in the 91-network columns of Table 1.

When EIG and max-min discrepancy differ. EIG leads when amplitude separates members or when the hypothesis set is large: by 20.0 points under the zero transform, and by 15.5 and 6.0 with 91 hypotheses and the transform known, the condition in which the two look level with 50. When the transform is unknown an equivalence test places them within 5 points at five trials and leaves a single trial undecided (Table 1). Max-min discrepancy, which profiles out the gain, leads when gains are uncertain or the kinetics are wrong: by 7.0 points with a gain prior of width 10 and by 19.0 when the indicator decays twice as slowly. Expected accuracy tracks EIG throughout (Appendix D).

The calcium model mainly keeps the objective off its ceiling. The voltage selector computes EIG with the same noise, gain and offset prior, and normalization on the members' voltage; its choice is scored under the calcium model. The comparison is therefore within the simulator, where the scoring and generating observation models coincide; on published recordings they are close in the contrast protocol and their ordering differs between T5 and Tm9 in the grating protocol (Appendix I). Under realistic uncertainty it trails EIG by 14.4 points from a single trial ($p = 0.006$) and with the transform known by 10.5 ($p \leq 0.04$). Its objective sits near the $\ln K$ ceiling for many designs, and its first picks are fast gratings and natural clips whose modulation the indicator smooths away (Fig. 1). Tempering it, as tempered EIG tempers the calcium objective, recovers 47 to 75% of the gap under realistic uncertainty, and what remains is not significant; it still trails by 15.5 points from a single trial under the zero transform ($p = 0.013$). At five trials under that transform it is the one condition in which the voltage objective does not trail (88.0% against 87.5%), and there every objective-based rule but observation disagreement lands between 85.8 and 91.5%, so the column separates none of them (Appendix D).

Ablations. Ranking recordings by how much members' calcium traces or voltages disagree costs 18 to 55 points under both conventions, because runaway members (the few whose responses exceed the others' many times) dominate the mean pairwise distance (Table 5). Under the resting-level transform the pairwise screen alone matches the full estimator, and is nominally 3.0 and 1.8 points

Table 5: **Ablations: first-recording recovery (%) with one ingredient removed** at the measured noise under both conventions. *significantly worse than the full objective (paired permutation test over members, Holm-adjusted over the seven rows of each column, $p < 0.05$). The reference is the first selection seed (82.8% and 96.2% in the resting-level columns) where Table 6 pools five (83.3 and 96.7) and adjusts over twelve rules, so a row can be starred there and not here. The last two rows separate the choice of target from the choice of stimulus.

	Resting-level transform		Zero transform	
Selector	1 trial	5 trials	1 trial	5 trials
Full objective (EIG)	82.8	96.2	**85.0**	87.5
Pairwise screen instead of full estimator	**85.8**	**98.0**	73.5*	87.0
Profiled discrepancy, median (no bound)	73.2	88.2	58.0*	62.0*
Full objective on raw voltage	72.8	90.0	70.2*	**88.0**
Observation-space disagreement	54.8*	78.2*	46.8*	51.2*
Raw-voltage disagreement	36.8*	56.2*	30.2*	51.0*
Selected target, standard stimuli	77.5	87.8*	68.5*	74.8*
Selected stimuli, PageRank target	78.0	96.0	60.5*	82.5

ahead of it (neither significant): its screening score is a soft minimum over member pairs, and its top design is max-min discrepancy's first choice for every fly. Both profile out the gain, and under the zero transform they trail EIG by 11.5 and 20.0 points.

Sensitivity overview. The noise, hypothesis-set, and gain-prior tests quantify how each condition changes recovery and the margins between rules. EIG stays ahead of voltage selection, the field's default, and the routine battery at independent noise from $\sigma = 0.30$ to 1.5, at both quartiles of the measured noise, and with 145 hypotheses, significantly in 20 of these 24 comparisons (Table 7). Ranked by information per second, as the schedules are, it is never significantly behind any rule at matched recording time and, under the uncertain-transform condition with a resting-level generating fly, is ahead of all of them from 90 s on. Two mismatches hurt it and both are repairable in the likelihood: an indicator decaying twice as slowly as modeled drops it to 54.8% where max-min discrepancy holds at 73.8%, and marginalizing the decay constant lifts it to 76.5% for 1.6 points when the kinetics are as modeled; assuming the wrong rectification convention costs 16 to 34 points, and marginalizing the transform recovers about half of that from a single trial and 80% of it at five for a resting-level fly. Noise that grows with the response, as it does within the published ROIs, also changes relative performance: every rule that selects large responses loses ground, and at 91 networks EIG's leads over the PageRank target and the field's default fall to 9.1 and 4.2 points, neither significant. In this comparison, giving the likelihood the noise scaling does not improve recovery, while letting the objective select under it returns EIG to 60.2% and its lead over the PageRank target to 12.8 points ($p = 0.004$), leaving the field's default at 6.6, within what the benchmark resolves (Appendix D).

Robustness. Table 7 repeats the benchmark under other conditions. EIG leads voltage selection, the field's default, and the routine battery in 20 of the 24 noise and hypothesis-set comparisons, and max-min discrepancy is level with it or ahead. Restricted to the 48 driver-accessible types its leads over voltage selection and the field's default are no longer significant. EIG's first cell type with a random stimulus, or the battery's first edge, recovers 64.8% and 56.2% at five trials against 96.2% with the selected stimulus: the gain comes from choosing the stimulus for the type. EIG's first recordings evoke strong responses (median member peak 8.5 typical ranges) without leaning on runaway members (one member above ten times the median in 26% of them, against 64 to 100% for the routine battery, the field's default, and disagreement).

Recording time. Charging each recording five trials plus 5-s intervals, ranking recordings by EIG per second, the criterion of the schedules in Section 4, recovers 80.2% of generating members from single trials within 45 s and 97.5% within 90 s with the transform known. At each of 45, 60, 90, 120, and 180 s it is never significantly below any rule, and it is significantly ahead of the routine battery and random recordings at all five budgets and of the field's default at four (Holm-adjusted within

Table 6: **First-recording recovery (%) of the generating model at the measured noise** (single trial $\sigma = 0.63$, five trials $\sigma = 0.28$, measured noise correlation). Realistic: the fly's transform drawn from its posterior on the published traces, every rule scoring with the five-transform mixture, gain prior of width 1 with non-negative simulated gains, 48 driver-accessible types; EIG selects under the most probable transform (second row: under the mixture). Known: the transform as simulated, gains integrated out with prior width 0.3 (EIG pools five selection seeds under the resting-level transform). *significantly worse than EIG (paired permutation test over the 50 generating members, Holm-adjusted over the twelve rules of a column, $p < 0.05$; none is significantly better). The reference is this table's EIG row, which pools five selection seeds in the known columns, where Table 5 uses one seed and a family of seven — so identical percentages can be starred in one table and not the other. Brackets give 95% bootstrap intervals; bold marks each column's best rule; dashes, rules not run.

	Realistic		Known, resting-level		Known, zero	
Rule	1 trial	5 trials	1 trial	5 trials	1 trial	5 trials
EIG	67.5 [61, 74]	78.3 [72, 84]	83.3	96.7	**85.0**	87.5
EIG, transform mixture at selection	60.0 [54, 66]	77.5 [72, 83]	–	–	–	–
Expected accuracy	66.7 [60, 73]	79.5 [74, 85]	79.8	96.8	76.8	87.2
Expected MAP accuracy	66.4 [60, 73]	80.5 [76, 85]	82.0	96.2	82.2	86.8
Max-min discrepancy	**70.9** [65, 77]	**81.6** [76, 87]	**85.8**	**98.0**	65.0*	85.8
Tempered EIG	64.2 [57, 71]	78.2 [72, 83]	84.2	96.2	73.2	**91.5**
EIG on raw voltage	53.1 [46, 61]*	66.8 [61, 72]	72.8*	90.0*	70.2*	88.0
Tempered EIG on raw voltage	63.9 [57, 70]	72.2 [65, 79]	76.5	92.8	69.5*	89.2
PageRank target	61.4 [55, 67]	74.3 [69, 79]	78.5	92.2	62.7*	88.0
PageRank target, edges first	44.0 [36, 52]*	56.1 [47, 65]*	–	–	–	–
Field's default	56.7 [48, 65]	70.1 [62, 77]	68.5	79.5*	68.2*	79.8
Disagreement	42.6 [34, 51]*	59.7 [53, 66]*	54.8*	78.2*	46.8*	51.2*
Routine battery, random order	38.2 [30, 47]*	55.9 [48, 64]*	48.0*	70.0*	41.0*	58.5*
Routine battery, fixed order	46.9 [38, 56]*	61.6 [53, 69]*	–	–	–	–
Random	25.0 [18, 32]*	38.7 [30, 47]*	32.2*	45.2*	33.0*	45.0*

each budget). Under realistic uncertainty, with the fly rectifying at its resting level, it recovers 61.0% within 45 s and 89.8% within 90 s, and from 90 s on it is significantly ahead of every other rule.

What makes the realistic benchmark harder. The realistic condition changes three things at once, and each was also run alone from the same starting point (82.8% with the transform known, all 65 types and a gain prior of width 0.3; one trial, first selection seed). Drawing the fly's transform from its posterior and marginalizing it costs 6.5 points (76.3%), widening the gain prior to 1 with non-negative simulated gains costs 10.5 (72.2%), and restricting selection to the 48 driver-accessible types costs 2.5 (80.2%). Together they cost 15.3 points (67.5%, the same value the posterior-weighted column of Table 6 reports), which is 4.2 points less than the sum of the three taken alone: they overlap by about a fifth, and the gain prior, not the unknown transform, is the largest single ingredient.

If T4/T5 subtypes cannot be isolated. A pan-T4/T5 line labels the eight subtypes together, so we also ran the 91-network realistic benchmark with selection restricted to the 40 types a single line isolates on its own. The restriction costs EIG 2.1 points from a single trial (not significant) and 5.3 at five ($p = 0.06$), and EIG so restricted still beats the field's default, which keeps T5b, by 10.8 and 11.5 points ($p \leq 0.009$) and the PageRank target by 10.0 from a single trial ($p = 0.009$; 5.4 at five, not significant). Max-min discrepancy leans on those subtypes far more, falling to 43.5 and 63.3%.

Controls that use no objective. Three rules with no information-theoretic content, run against EIG in the same 91-network benchmark (one trial, realistic and known): the design with the largest median member peak, a generic moving edge on a cell type drawn at random from the driver-accessible list, and the same edge on EIG's own first cell type. They recover 39.6/8.7, 49.2/55.2, and 51.4/70.2% against EIG's 63.7/81.2 (realistic/known, each Holm $p \leq 0.002$). Amplitude alone is not what EIG exploits: the largest-response rule is the worst of the three and, with the transform

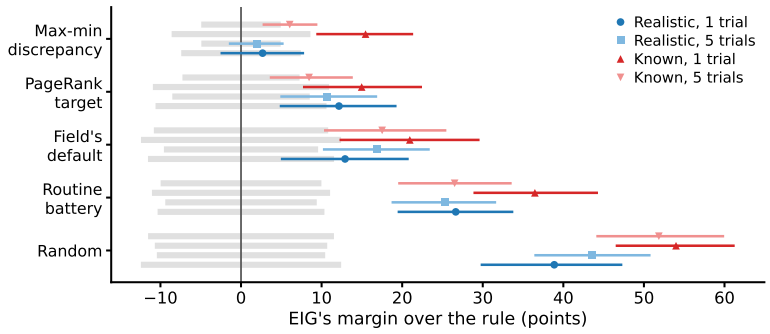

Figure 5: **What the 91-network benchmark resolves.** EIG's paired margin over each rule with its 95% bootstrap interval; grey marks the approximate effect detectable at 80% power before multiplicity adjustment. Every margin clears it except max-min discrepancy under an unknown transform, which an equivalence test places within 5 points at five trials and leaves undecided at one.

known, falls below random recordings, because the designs that drive every member hardest are often the ones that drive them alike. With the transform known, choosing the target is worth 15.0 points over a random accessible type and choosing the stimulus for it a further 11.0; under realistic uncertainty nearly all of the gain is in the stimulus (2.2 and 12.3).

Why a generic edge is the worse bet. The PageRank target recovers 61.4% led with a standard flash and 44.0% led with a standard moving edge, and the routine battery likewise does worse with edges first, although a moving edge is the best design for 61 of 65 types (Fig. 2). The two facts are consistent because only the type's *best* edge is best: for Tm5b, the highest-PageRank type, the catalog's 144 moving edges carry 1.20 to 3.24 nats while its 8 full-field flashes carry 1.84 to 2.98, so an edge drawn without a model is the worse bet even where the best edge wins. Choosing the stimulus for the type is what closes that gap (Table 5).

Resolving the comparison with the targeted baselines. The field's default experiment reaches 56.7% and 70.1% and the PageRank target 61.4 and 74.3% — gaps at the edge of what 50 generating networks resolve, since the paired test detects 12.1 and 12.7 points at 80% power, and the spread that limits it lies between networks rather than between noise draws, so increasing the number of generating networks is more effective than adding noise draws (Appendix D). We therefore repeated the benchmark with 91 same-recipe networks: the 50 published members plus two ensembles of 21 and 20 retrained on the measured connectome, each in turn the fly and all 91 in the hypothesis set. No difference in recovery by training origin was detected for any rule. The pooled evaluation includes the published ensemble and two independently retrained ensembles; the 50- and 91-network benchmarks share members and are not independent. EIG then leads the PageRank target by 8.4 to 15.0 points and the field's default by 12.9 to 20.9 across the four conditions, each significant after adjustment (Table 1). Two things changed at once and they do different work. Matching runs (the 91-network benchmark simulates a resting-level fly, where the 50-network margin over the PageRank target is 2.0 points, not the posterior-weighted 6.1), holding the flies at 50 and raising the hypothesis set to 91 moves that margin to 13.1 points, while adding the other 41 flies leaves it at 12.2 and makes it precise: a fixed connectome target degrades faster than a selected recording as the candidate set grows. Because both the hypothesis set and the number of generating networks increase, this comparison changes the discrimination task as well as its statistical precision (Appendix D).

What the benchmark can resolve. Contrasts use paired sign-flip tests over generating networks. For the per-network differences d, the estimated standard error is $\mathrm{SD}(d)/\sqrt{n}$. A normal approximation to the effect detectable at 80% power and an unadjusted two-sided 5% level is $(z_{.975} + z_{.80})\,\mathrm{SD}(d)/\sqrt{n} = 2.80\,\mathrm{SD}(d)/\sqrt{n}$. $\mathrm{SD}(d)$ is not a constant of the benchmark: two rules that pick similar recordings differ consistently from network to network, and two that pick differently do not. In the realistic single-trial column, with $n = 50$, it ranges from 13.2 points against max-min discrepancy to 38.2 against the routine battery, so the detectable effect ranges from 5.2 to 15.1 points here and from 5.0 to 15.2 across both realistic columns. Thus a 3.5-point difference from max-min and a 6.0-point difference from the PageRank target have different precision;

equivalence additionally requires a prespecified margin and the two one-sided tests reported below. Splitting $\mathrm{SD}(d)^2$ into a between-network and a within-draw part (EIG against the field's default, eight draws per network) gives between-network SD 32.3 and within-draw SD 43.0 points: with the draws already averaged, the detectable effect is 14.1 points, and would fall only to 12.8 with infinitely many draws. Averaging more noise draws therefore cannot resolve these gaps and more generating networks can, which is what the 91-network benchmark of Section 3 does; at $n = 91$ the same decomposition predicts 10.5 points and the realized detectable effects are 10.6 against the PageRank target and 11.5 against the field's default, both below the observed margins.

Hypothesis-set size against paired sample size. The 91-network benchmark raises both at once, so we ran the intermediate: the original 50 published members as flies against all 91 hypotheses (uncertain transform at inference, resting-level generating transform, single trial). These matched runs differ from the posterior-weighted transform average in Table 6. Recovery falls for every rule but by very different amounts — EIG 67.5 → 64.6%, the PageRank target 65.5 → 51.5, max-min discrepancy 73.8 → 62.3, the field's default 54.0 → 52.2 — so EIG's margin over the PageRank target grows from 2.0 to 13.1 points with the flies unchanged. Adding the remaining 41 flies moves that margin only to 12.2 while taking its p from 0.006 to 0.0018, and the margin over the field's default from 12.4 to 12.9 (p from 0.020 to 0.0029). The hypothesis set therefore sets the size of the effect and the number of flies its resolution.

Seeds and equivalence. Pooled over five selection seeds, EIG recovers 83.3% and 96.7% with the transform known, within one point of the first seed. The 90% intervals of EIG's paired difference from expected accuracy, expected MAP accuracy, tempered EIG, and max-min discrepancy lie within ±5 points at five trials under the resting-level transform (two one-sided tests); from a single trial they reach 8 points, and under the zero transform EIG leads expected accuracy, tempered EIG, and max-min discrepancy by intervals that exclude zero (up to $[12.0, 27.7]$).

Unknown transform. Assuming the other convention in selection and inference costs EIG 28.5 and 33.5 points (one and five trials) for a fly that rectifies at its resting level and 23.0 and 15.8 for one that rectifies at zero (Table 7). Crossing selection with inference separates the two. For the resting-level fly, the mixture likelihood costs 6.8 and 1.0 points when selection assumes the right transform (76.0% and 95.2%) and lifts wrong-convention selection to 69.5% and 89.5% (from 54.2% and 62.7%); selecting under the mixture as well gives 68.0% and 85.5%. For the zero fly, it lifts wrong-convention selection from 62.0% and 71.8% to 73.2% and 78.2%, and mixture selection reaches 70.0% and 87.0%. In the realistic benchmark, selecting under the mixture recovers 60.0% and 77.5% against 67.5% and 78.3% under the most probable transform; with a shortlist built under the mixture and 1024 draws the deficits are 5.2 and 5.5 points ($p = 0.16$ and 0.07). This comparison changes both screening and the Monte Carlo budget and does not isolate their contributions. With fly transforms outside the mixture (a power of 3) or a supralinear power of 2.5, mixture selection and inference recover 89.0% and 91.5% at five trials, ahead of routine practice and random recordings.

Indicator kinetics. When the fly's indicator decays twice as slowly as modeled, marginalizing the decay over 0.2, 0.4, and 0.8 s at inference and selection keeps EIG at 76.5% and 94.5% (one and five trials) against 54.8% and 71.2% with the nominal decay, within 5 points of max-min discrepancy (81.2% and 95.5%) and 34 points ahead of the routine battery ($p < 0.01$); when the kinetics are as modeled the same mixture costs 1.6 and 1.0 points (81.2% and 95.2% against 82.8% and 96.2%). Marginalizing only the transform recovers 39.8% at five trials, 25 points below the field's default ($p < 0.001$). On the published traces, decays of 0.1 to 0.4 s fit about equally well (median best-member R^2 0.67 to 0.72) and 0.8 s or slower fit worse (0.51 to 0.57).

Indicator saturation. Published responses reach a median 4.2 typical ranges and a 90th percentile of 9.7, and EIG's first recordings evoke a median member peak of 8.5, so the amplitudes it selects are within the range real recordings show. Saturating the realistic benchmark at that 90th percentile costs every rule a third to three-fifths of its recovery: against the resting-level run it matches, EIG falls 67.5 → 38.0% and max-min discrepancy 73.8 → 44.8% from a single trial. Those two keep the top places, while the targeted heuristics lose most (PageRank 65.5 → 27.5, the field's default 54.0 → 24.5) and fall behind the routine battery, which loses least (39.8 → 26.8). Saturation is repairable in the likelihood exactly as slow kinetics are. Adding three half-saturations (5, 10, 20

typical ranges) of the resting-level transform to the inference mixture lifts EIG from 38.0 to 52.5% and max-min discrepancy from 44.8 to 59.0% against a fly saturating at 10 (+14.5 and +14.2, $p < 0.0001$; every rule gains, the PageRank target most at +17.8), recovering about half of what saturation costs. Unlike the kinetics mixture it is free when the fly does not saturate: EIG moves $67.5 \to 68.0\%$ and no rule changes significantly ($p \geq 0.70$), so the half-saturation can be carried in the likelihood as a matter of course. With half-saturation at 1.5, 3, 5, and 10 typical ranges the best rule recovers 12.5, 22.0, 32.8, and 51.2% from a single trial and 26.5, 44.0, 56.8, and 73.8% from five, against 85.8 and 98.0% unsaturated.

Noise that grows with the signal. Within each published ROI the single-trial residual is $1.47\times$ larger above the ROI's median response amplitude than below it (interquartile range 1.02 to 2.07 over 269 ROIs), so noise grows with signal roughly as amplitude$^{0.46}$, close to shot noise. Simulating the fly with that power law, normalized to the measured σ at the median published response of 4.2 typical ranges, is the one perturbation that changes which rule wins: the rules that select large responses lose (EIG $67.5 \to 61.0\%$, max-min $73.8 \to 70.5$, the PageRank target $65.5 \to 58.5$) while those that do not gain (the routine battery $39.8 \to 44.5$, random $23.2 \to 31.0$). Max-min discrepancy, which profiles out the gain, then leads EIG by 9.5 points from a single trial and 5.5 from five (both significant), and EIG's leads over the PageRank target and the field's default fall to 2.5 and 5.0 points. Repeating this at 91 networks, where the comparison is better resolved, EIG recovers 57.5% and is level with max-min (+0.3, not significant) while its leads over the PageRank target and the field's default are 9.1 and 4.2 points, neither significant after adjustment ($p = 0.08$ and 0.68). The mismatch is one of selection rather than scoring. Giving the likelihood the same amplitude-dependent σ changes nothing ($57.5 \to 57.0\%$): the posterior was not being misled. Letting the objective *select* under it does, because a large predicted response is then no longer treated as a cheap one: EIG returns to 60.2%, its lead over the PageRank target to 12.8 points ($p = 0.004$ after adjustment) and over max-min to 3.1, while the field's default remains 6.6 points behind, within what the benchmark resolves. Amplitude-dependent noise therefore costs about 3.5 of the 6.2 points it takes from an unaware rule, and, like slow kinetics and saturation, is largely handled by carrying the right constant — here in the objective rather than in the likelihood. Noise that differs between cell types leaves EIG at 73.2%, ahead of the routine battery and random recordings, with max-min discrepancy and the PageRank target at 81.0% and 80.2% (not significantly different).

A larger hypothesis set. The published ensemble, two ensembles retrained on the measured connectome (21 and 20 members), and the three altered-wiring ensembles of Appendix G give 150 networks; the published members serve as flies. Five identity- or sign-shuffled networks exceed 100 times the across-member median voltage on at least 134 of the 598 stimuli (no other network on more than 11); we fixed that exclusion rule after a first run in which their stimuli captured the selections, and clipping their traces instead leaves every low-noise margin of the same sign. Table 7 gives the 145 remaining networks and, at the measured single-trial noise, all 150 clipped: EIG recovers 63.7%, ahead of voltage selection, the PageRank target, and the routine battery by 21 to 31 points (each $p < 0.01$); its lead over the field's default is 8.2 points and not significant.

E Feasible schedules and their transfer

A recording costs five trials of its stimulus plus intervals (32.5 s for a 1-Hz grating), a fly provides 30 minutes and one driver line, and cost-aware lazy greedy selection over the 40 types a single line isolates (Section 2) at $\sigma = 1.0$ stops when no candidate adds information. With three flies it chooses TmY4, Tm5b, and Mi10 (Table 8); with one fly it chooses TmY4. Ranked by information per second, both open with gratings, which are shorter than edge presentations and nearly as informative for TmY4 (2.91 nats for its best grating against 3.06 for its best edge at $\sigma = 1.0$); the one-fly schedule opens with a different grating, which transfers slightly better. Each schedule is applied unchanged to two ensembles retrained from scratch (21 and 20 members), each the hypothesis set at inference with its own normalization, with 32 noise draws per member; randomized controls average 25 draws.

Controls. Every control records a compact battery (two flashes, then the slowest ON and OFF edges in four directions) on one of: the designed targets in the same order, the three types whose calcium traces disagree most, the three with the highest PageRank, or three drawn at random —

Table 7: **First-recording recovery (%) under robustness conditions** (first selection seed; resting-level transform and single trial unless stated). Independent noise at $\sigma = 0.63$ gives 96.2% against 82.8% with the measured lag-one correlation, so the correlation alone costs about 13 points. Acc.: expected accuracy; Temp.: tempered EIG; Volt.: EIG on raw voltage; PR: PageRank target; Disagr.: observation disagreement; saturating at K: half-saturation at K typical ranges. * a rule EIG beats significantly, † one that beats EIG (paired permutation over the 50 generating members, Holm-adjusted within the row, $p < 0.05$); dashes, rules not run.

Condition	EIG	Acc.	Max-min	Temp.	Volt.	Field	PR	Battery	Disagr.	Rand.
Noise level										
Independent, $\sigma = 0.30$	99.0	–	**99.5**	–	85.2*	85.0*	92.5*	71.8*	81.2*	46.8*
$\sigma = 0.63$	**96.2**	–	95.5	–	74.0*	77.8*	80.2*	54.8*	63.0*	35.5*
$\sigma = 1.0$	86.5	–	**88.5**	–	63.7*	69.5*	69.2*	43.8*	50.0*	29.0*
$\sigma = 1.5$	74.2	–	**79.5**	–	55.2*	61.5	57.8*	36.8*	41.5*	24.2*
Lower quartile, 0.40, AR	90.5	93.5	**94.5**	86.8	78.2*	75.0*	88.0	62.3*	65.8*	38.8*
Upper quartile, 1.23, AR	**73.0**	71.8	66.5	69.2	60.5*	54.5*	59.8	33.8*	40.0*	23.8*
Selection among the 48 driver-accessible types										
1 trial	80.2	82.8	**85.8**	80.2	75.5	68.5	78.5	48.0*	54.8*	26.8*
5 trials	97.0	97.5	**98.0**	95.8	88.8*	79.5*	92.2	70.0*	78.2*	46.0*
Gain prior of width 10 at inference										
1 trial	76.0	77.2	**82.8**	76.5	57.2*	58.5*	75.2	37.2*	38.5*	18.5*
5 trials	90.2	89.8	**97.2**†	92.5	79.2	73.2*	89.8	63.2*	69.8*	32.8*
145 or 150 networks as hypotheses										
1 trial	**68.5**	66.2	53.5*	64.0	57.5	57.0	43.2*	42.0*	41.5*	24.2*
5 trials	**89.8**	86.8	73.5*	85.5	85.0	72.5*	69.0*	63.2*	62.0*	35.0*
150, clipped, 1 trial	**63.7**	–	–	–	32.8*	55.5	42.5*	40.2*	40.0*	23.8*
Gain prior of width 1 in the fly and at inference										
1 trial	72.2	72.8	**75.8**	68.2	63.0	58.2*	69.2	42.0*	45.2*	28.0*
5 trials	82.0	84.2	**86.5**	82.5	76.0	69.5	83.0	62.5*	65.8*	39.8*
Fly outside the measurement model (selection and inference nominal)										
Decay 0.4 s, 1 trial	54.8	54.8	**73.8**†	33.0*	43.5	59.5	30.0*	33.5*	30.2*	26.5*
Decay 0.4 s, 5 trials	71.2	76.8	**88.2**†	65.0	50.5*	69.8	31.8*	42.2*	41.2*	36.0*
Saturating at 3, 1 trial	21.0	20.8	**22.0**	12.2	18.5	20.2	14.5	12.5	13.5	10.0*
Saturating at 5, 1 trial	30.8	30.0	**32.8**	19.2	30.2	28.5	23.8	20.8	21.5	15.2*
Saturating at 10, 1 trial	49.2	47.0	**51.2**	29.2*	47.5	37.8	37.2	32.5	33.0	21.2*
Noise by type, 1 trial	73.2	71.5	**81.0**	55.2*	69.5	65.5	80.2	47.5*	46.0*	34.0*
Noise by type, 5 trials	95.2	95.5	**96.8**	94.2	83.8*	78.5*	93.0	63.5*	64.0*	42.5*
Power 2 at rest, 1 trial	55.8	56.0	**64.2**	43.0	53.0	51.7	59.2	28.7*	35.5	26.0*
Power 2 at rest, 5 trials	58.8	61.0	**72.8**	71.5	55.5	56.8	60.2	29.8*	39.0	34.5
Fly decay 0.4 s, decay marginalized over 0.2, 0.4, 0.8 s at inference										
1 trial	76.5	–	**81.2**	–	61.0*	67.8	65.5	42.2*	–	30.5*
5 trials	94.5	–	**95.5**	–	81.8*	78.2*	84.8*	62.3*	–	41.2*
Fly at zero, selection and inference at resting level										
1 trial	62.0	58.8	63.5	**69.5**	46.0	46.8	21.5*	26.8*	30.8*	16.5*
5 trials	71.8	73.0	72.5	**77.0**	56.8	59.8	39.2*	34.2*	44.2*	22.5*
Fly at resting level, selection and inference at zero										
1 trial	**54.2**	42.2	47.8	19.2*	37.5	47.8	33.2	34.5	16.8*	21.8*
5 trials	**62.7**	58.5	56.0	58.5	58.5	56.0	38.5	44.0	22.8*	27.8*

always from the schedule's own pool of 40. Table 9 pools the two retrained ensembles. At a matched 97.5 s of recording the lead over standard stimuli on the same targets (10.3 $[4.2, 16.9]$ points from single trials, 4.6 $[1.7, 8.0]$ from five) and on random targets (5.3 $[3.3, 7.3]$ and 1.5 $[0.6, 2.3]$) remains, while PageRank and disagreement targets catch up (intervals include zero). The one-fly schedule's first recording recovers 94% and 93% at five trials (82% and 83% from a single trial).

Table 8: **The frozen three-fly schedule** selected on the published ensemble under the resting-level transform; five trials per recording, times excluding preparation. Rows give the greedy selection order; in the lab each fly's recordings are made together, which changes the order but not the set.

Recording	Target	Line	Stimulus	Cumulative time (s)
1	TmY4	SS00785	grating 46.4°, 90°, 2.0 Hz	32.5
2	Tm5b	SS03729	grating 46.4°, 90°, 1.0 Hz	65.0
3	Mi10	SS38287	ON edge, 330°, 25°/s	101.6
4	TmY4	SS00785	grating 46.4°, 270°, 1.0 Hz	134.1
5	Tm5b	SS03729	OFF edge, 0°, 25°/s	170.7
6	Mi10	SS38287	ON edge, 0°, 25°/s	207.3
7	Tm5b	SS03729	OFF edge, 120°, 19°/s	244.4
8	Mi10	SS38287	ON edge, 30°, 25°/s	281.0
9	Tm5b	SS03729	OFF edge, 150°, 25°/s	317.6
10	TmY4	SS00785	grating 46.4°, 90°, 1.0 Hz	350.1
11	TmY4	SS00785	grating 46.4°, 225°, 1.0 Hz	382.6
12	Tm5b	SS03729	OFF edge, 30°, 25°/s	419.2

Table 9: **Transfer of the frozen schedule** to the two retrained ensembles (41 generating members pooled): first-recording recovery (%), the designed schedule's paired margin (bootstrap over members stratified by ensemble; intervals not adjusted for multiplicity), and scheduled seconds to 90% recovery (range over ensembles and noise levels).

	Single trial		Five trials		
Schedule	%	Margin [95% CI]	%	Margin [95% CI]	t_{90} (s)
Designed schedule	**76.4**	–	**89.6**	–	65
Standard, same targets	66.9	9.5 [2.9, 16.1]	83.4	6.2 [0.8, 11.8]	80–120
Standard, random targetable types	69.2	7.2 [1.1, 13.2]	82.7	6.9 [3.0, 10.8]	80
Standard, PageRank targets	58.6	17.8 [5.8, 29.6]	70.2	19.4 [9.6, 29.6]	80
Standard, largest-disagreement targets	66.4	10.0 [-1.2, 21.6]	83.7	5.9 [-1.5, 13.6]	80
Designed stimuli, PageRank targets	67.2	9.1 [0.3, 18.3]	85.6	4.0 [-1.4, 9.5]	65

F ATLAS CHECKS

A moving edge gives the best design for 61 of the 65 types (a flash for three, a naturalistic clip for one). At five-trial noise 56 of the 65 types come within 5% of the $\ln 50$ ceiling, which is why the atlas is scored at $\sigma = 1.0$. The ranking of cell types by their best design, computed from 10, 20, and 35 members, correlates with the 50-member ranking at Spearman 0.90, 0.94, and 0.97. Across types the best achievable EIG spans 1.6-fold between the 5th and 95th percentiles and 3.1-fold overall; across the catalog the median design carries 1.26 nats, standard flashes and edges 2.00, and standard stimuli on the routinely recorded types 2.08.

G WIRING-CLASS PROTOCOL

Four fixed graphs (measured connectivity, degree-preserving rewiring, shuffled cell-type identities, shuffled synapse signs) are each trained with several seeds under the published recipe (80 networks: 21, 20, 20, and 19 per class); normalized traces are clipped at the saturation level. Folds stratified by class keep every test network out of normalization, selection, and the inference mixture; 4,420 candidate recordings (65 types by 68 stimuli) are scored, each test network generates 64 noisy recordings, and the class likelihood marginalizes each class's training networks. A class-level objective maximizes $I(C; Y_d)$, the information about the wiring class C with members marginalized within each class (Box & Hill, 1967).

Under the transform the published traces favor and the measured noise, one class-level recording identifies the wiring class of 55.4% of held-out networks at five-trial noise and 52.4% from a single trial, averaged over five fold assignments, against 42.2% and 41.2% for the standard flash-and-edge battery and 49.7% and 47.7% for member-level design (chance 25%); the gain over the battery is

positive in every assignment (5.0 to 19.9 points at five trials, 4.5 to 19.0 from a single trial) and its interval excludes zero in three of five. Class-level accuracy does not grow with further recordings (55.4, 55.0, 54.7% over three), while the battery's does (42.2, 46.6, 49.4%), because a finite mixture of trained networks assigns a new network confidently to its nearest training network.

H PUBLISHED MEASUREMENTS

The corpus holds 110 records (27 tuning curves, 21 indices, 62 binary labels) from published datasets and deposits (Lappalainen et al., 2024; Maisak et al., 2013; Drews et al., 2020; Ketkar et al., 2022; Groschner et al., 2022), with 37 recording labels. Three holdouts remove every record of a target label, pool T4 and T5 subtypes, or remove all T4 and T5 records; rules spend budgets of 12 to 96 measured values on whole records ranked from model predictions alone. EIG per record peaks at 48 values in all three holdouts (0.90 to 0.92) and trails observation disagreement at 12 and 24; 48 is where the rules differ most, not a pre-registered choice. At that budget it beats the unweighted ensemble by 0.136 $[0.023, 0.306]$ in the cell-type holdout and by 0.034 $[-0.006, 0.086]$ and 0.018 $[-0.009, 0.068]$ in the two stricter ones, and does not beat observation disagreement (0.012). We quote effects and intervals rather than p-values here: the budget was chosen after seeing the curves, so a test at it would not have its nominal level.

I CALCIUM TIME SERIES

Contrast protocol. Drews et al. (2020) deposited single-trial $\Delta F/F$ traces (GCaMP6f) for 313 ROIs of 15 cell types under 42 contrast conditions of a 1-Hz grating, three trials each. We simulate the conditions for all members, pass the voltage through the observation model and normalization, convert each ROI to model units with one member-independent factor (the ensemble-median model peak-to-trough at full contrast, in the normalized units of Section 2.1, over the ROI's), and fit one gain and offset per ROI. The ensemble mean explains a median 31% of trial-mean variance and the best member 66%, against a trial-to-trial ceiling of 83%. The own-type prediction fits best for 9% of ROIs (chance 7%) and a same-polarity type for 76%. In the design test, EIG-chosen conditions beat random ones by 0.057 $[0.045, 0.068]$ R^2 from one condition and stay within 0.007 of a contrast battery and 0.004 of voltage EIG at every count. Each type's EIG at its own measured noise tracks the realized posterior concentration across the 15 types (Spearman 0.82, $p = 0.0002$) and its agreement with published tuning (0.62, $p = 0.014$). The noise level alone predicts them about as well in the other direction (-0.78 and -0.63), which is unsurprising because EIG here is evaluated at that noise (the two rank at -0.80). Controlling for it, EIG keeps a positive but marginal association with concentration (partial Spearman 0.50, $p = 0.055$) and none with agreement (0.25, $p = 0.37$). The atlas's best-design EIG, which is computed at a common noise level, predicts neither (0.06).

T5 gratings from a second laboratory. Ramos-Traslosheros & Silies (2021) recorded T5b terminals under 36 sine gratings moving back-to-front at full contrast (wavelengths 2 to 32 degrees, temporal frequencies 0.25 to 16 Hz, five trials; 21 ROIs from ten flies, each paired with a Tm9 measurement), deposited as Fourier amplitudes at the grating frequency. We simulate the gratings through a 5-degree Gaussian ommatidial acceptance function and take the same amplitude of the calcium-model trace. The test uses 50 random halves of the grid: greedy EIG, a standard sweep from the grid point nearest a 30-degree, 1-Hz grating, and 50 random orders choose gratings on one half, and the most probable member's map is scored on the other; intervals resample ten ROIs per bootstrap replicate. This is an ROI-level resampling procedure, not a bootstrap clustered by fly. Each deposited amplitude averages five trials over whole stimulus periods, so we use $\sigma = 0.1$; every conclusion below holds at 0.05 and 0.2. For T5, EIG-chosen gratings beat random ones at every count (0.06 to 0.19 R^2); the sweep leads at one grating (0.098 $[0.021, 0.159]$) and ties from two to six. For Tm9, EIG-chosen gratings beat random ones at every count and the sweep from two to six (0.065 to 0.111; 0.065 $[0.030, 0.101]$ at two), and EIG on voltage does worse than EIG on calcium (0.06 to 0.10 lower from two to six gratings). For T5 the ranking reverses: voltage-chosen gratings lead by 0.03 to 0.06 from two to four gratings and tie elsewhere. T5 and Tm9 were recorded in the same flies, so they are paired readouts rather than independent biological replications. The ensemble

median explains 81% of the T5 maps, while generic low-pass templates explain 79 to 87%; absolute map fit alone therefore does not establish specificity to the connectome-constrained ensemble.

J FLIES OUTSIDE THE HYPOTHESIS SET

Confidence and predictive fit. With the 41 retrained networks as generating models and the 50 published members as hypotheses, one recording places 0.95 of the posterior on a member ranking 22nd to 26th of 50 in fit to the generating network (chance rank 25.5). Identification confidence therefore need not imply accurate prediction for networks outside the hypothesis set. The continuous-response endpoint below holds out stimuli within the recorded cell type; the separate flash-sign endpoint evaluates other cell types.

Prediction of unseen responses. With the 41 retrained networks as flies and the 50 published members as hypotheses (the transform the published traces favor; prediction scored at $\sigma = 0.1$, posterior ranks at the measured noise), one recording of a driver-accessible type improves prediction of that type's responses to 24 unseen stimuli under every rule, measured against the unweighted ensemble (Table 10). EIG's recording helps less than a standard or random one at $\sigma = 0.1$, and about as much as a random one at $\sigma = 1.0$. A discrepancy term calibrated on the ensemble's own nearest-neighbour residuals lowers the mass on the top member from 0.95 to 0.82 but leaves the ordering unchanged; prediction-oriented acquisition (Bickford Smith et al., 2023) does no better, and ranking by EIG at tenfold noise does worse. The mechanism is visible in the recordings: EIG's separate the members about 4 and 40 times more than standard and random ones, while the retrained networks lie no farther from their nearest member on them (median 1.2 against 1.3 and 1.4). A qualitative endpoint behaves the same way: after one recording the posterior predicts the sign of other types' responses to full-field flashes, on the 66 (type, flash) pairs where at least ten members disagree, less accurately than the unweighted ensemble's majority — by 4.8 points $[2.9, 6.6]$ for EIG, 4.4 for the standard battery, and 2.2 for a random recording.

Table 10: **Prediction of unseen responses for flies outside the ensemble:** gain in mean correlation over the unweighted ensemble after one recording, at posterior temperature $\beta = 1/32$ or 1, with 95% bootstrap intervals over the 41 retrained networks.

	$\sigma = 0.1$		$\sigma = 1.0$	
Rule	$\beta = 1/32$	$\beta = 1$	$\beta = 1/32$	$\beta = 1$
EIG	0.046 $[0.025, 0.067]$	0.033 $[0.009, 0.057]$	0.026 $[0.018, 0.035]$	0.030 $[0.008, 0.051]$
EIG at tenfold noise	0.019 $[-0.003, 0.042]$	0.017 $[-0.009, 0.042]$	–	–
Prediction-oriented EIG	0.039 $[0.019, 0.060]$	0.027 $[0.003, 0.051]$	0.026 $[0.017, 0.034]$	0.026 $[0.006, 0.047]$
Standard battery	0.077 $[0.058, 0.096]$	0.061 $[0.038, 0.083]$	0.041 $[0.034, 0.049]$	0.055 $[0.036, 0.074]$
Random recording	0.075 $[0.060, 0.090]$	0.073 $[0.052, 0.094]$	0.030 $[0.024, 0.037]$	0.053 $[0.038, 0.069]$

K EXTENDED RELATED WORK

Choosing experiments to discriminate a finite set of mechanistic models is the oldest use of EIG (Box & Hill, 1967; Lindley, 1956), alongside classical discrimination criteria (Atkinson & Fedorov, 1975), and continues as adaptive design optimization (Cavagnaro et al., 2010); Daunizeau et al. (2011) already scored designs for comparing models of brain function through a hemodynamic observation model. For a finite ensemble the mutual-information objective coincides with BALD and is related to query-by-committee (Houlsby et al., 2011; Seung et al., 1992); our schedules select sets of recordings using a joint-information objective, related to BatchBALD (Kirsch et al., 2019), and prediction-oriented acquisition (Bickford Smith et al., 2023) targets the unseen-stimulus endpoint we test for networks outside the ensemble.

Stimulus design and observation models. Controversial stimuli (Golan et al., 2020) and related synthesis methods (DiMattina & Zhang, 2011; Walker et al., 2019; Burg et al., 2024; Kuo & Walker, 2026) separate models by maximizing their disagreement, a criterion we compare directly (max-min discrepancy). Amortized and implicit-likelihood design (Foster et al., 2019; 2021; Ivanova et al.,

2021; Rainforth et al., 2024) handles likelihoods without closed form, whereas ours is exact after nuisance marginalization, and design under misspecification (Overstall & McGree, 2022; Sloman et al., 2022) and the bias of active learning (Farquhar et al., 2021) motivate our tests with the fly under another transform or outside the ensemble.

Neuroscientific experimental design. In neuroscience, sequential design has estimated receptive fields, mapped synaptic connectivity, and chosen stimuli or photostimulation patterns for imaged populations (Lewi et al., 2009; Shababo et al., 2013; Cowley et al., 2017; Wagenmaker et al., 2024), and Beiran & Litwin-Kumar (2025) choose which neurons of a connectome-constrained model to record by how well their noiseless activity predicts the rest. What is new here is applying measurement-aware discrimination design to an ensemble of trained connectome-constrained networks at the scale of a whole visual system, under driver-line constraints and with a benchmark calibrated to published recordings. In this benchmark the calcium model's main effect is to keep the objective off its $\ln K$ ceiling, and tempering the voltage objective reproduces part of it (Appendix D).