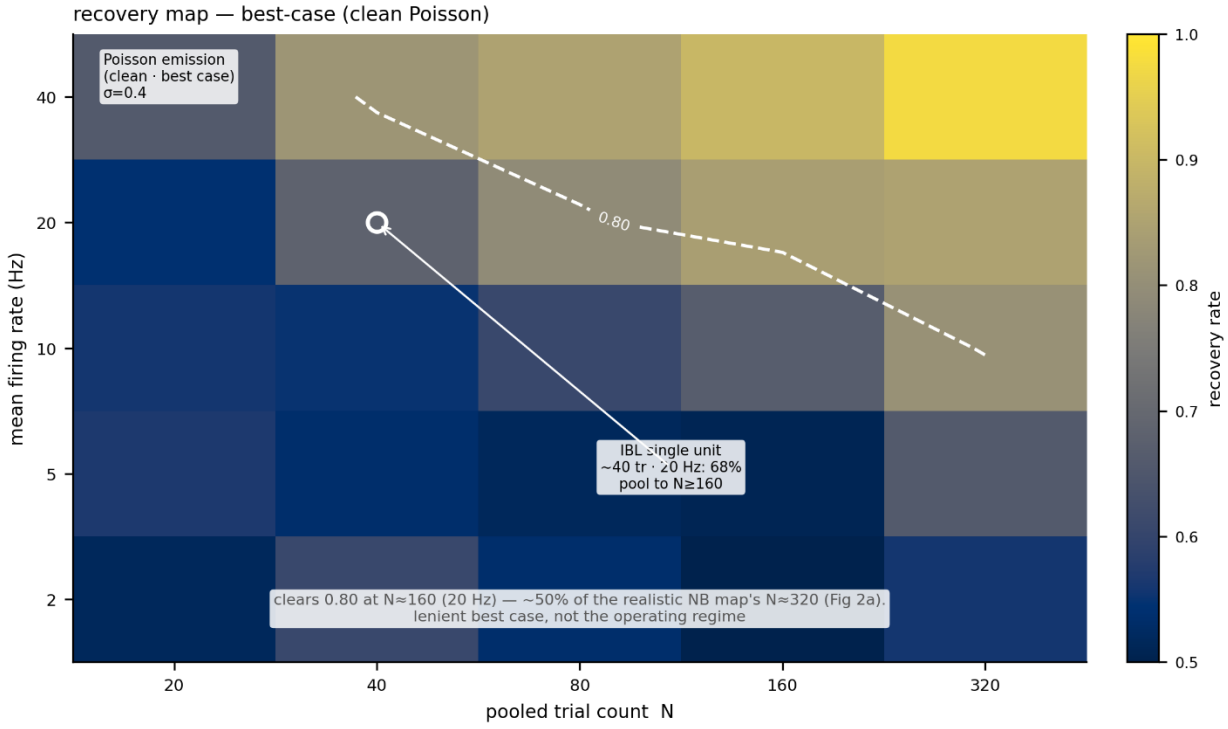
Movement controls bias decision-signal estimates across the brain: a calibrated, ground-truth test

Supplementary Information

This document contains 2 supplementary figures (S1, S2) and 5 supplementary tables (S1 to S5). All values are read from cached result files; the source file is named under each item. All text is ASCII (no unicode glyphs): “sigma” is spelled out, “>=” and “<=” replace the inequality symbols, “about” replaces the tilde, and “->” replaces arrows.

Contents: - Supplementary Figure 1. Best-case (clean Poisson) recovery map. - Supplementary Figure 2. Diffusion-sigma sensitivity and intermediacy detail. - Supplementary Table 1. Forward-likelihood validation of the model engines. - Supplementary Table 2. Movement-control calibration. - Supplementary Table 3. Confound cascade under the valid control with within-region FDR. - Supplementary Table 4. Robustness of the at-the-boundary decode to analysis choices. - Supplementary Table 5. Full-cohort simultaneity census.

Figure S1 · Poisson best-case recovery (does not apply to real, overdispersed spikes)

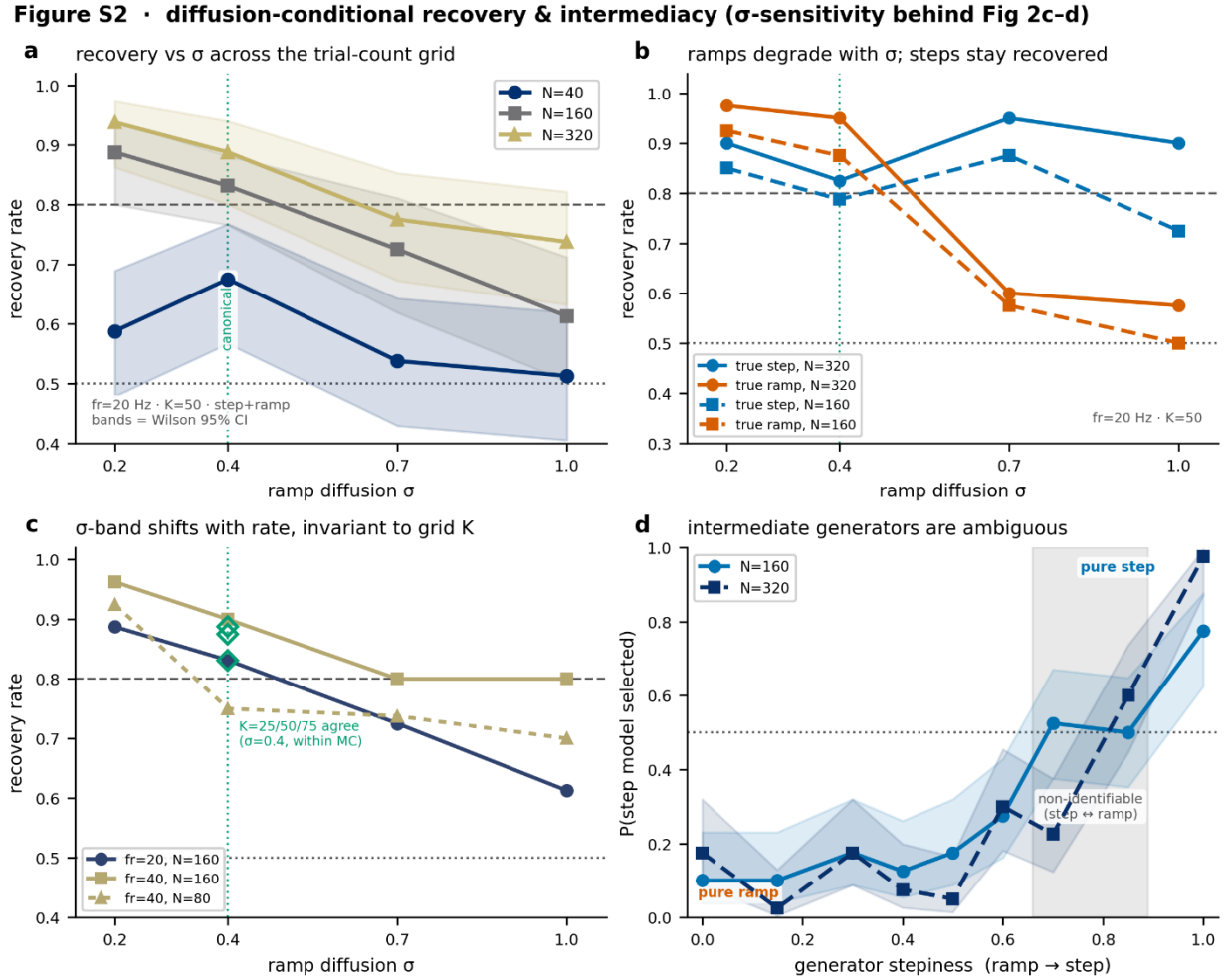


Supplementary Figure 1. Best-case (clean Poisson) recovery map.

Step-versus-ramp recovery under clean Poisson emissions, the lenient upper bound that does not apply to real, overdispersed spikes (Fig 2a uses the realistic negative-binomial map). Heatmap of recovery rate (cividis colour scale from 0.50 to 1.00) over mean firing rate (y axis, Hz) and pooled trial count N (x axis), with ramp diffusion sigma = 0.4. The white dashed contour marks the 0.80 recovery level. The circled point is the IBL single-unit operating point at the N = 40 grid column and 20 Hz, where best-case recovery is 0.68; reaching 0.80 needs pooling to N about 160, which is about 50 percent of the trial count the realistic negative-binomial map needs (N about 320, Fig 2a). Recovery rates are fractions of correctly recovered generators at each cell; no error bars.

Source: results/phase1_recovery_sweep.csv (clean Poisson sweep). Cross-reference: Fig 2a.

Note on the operating point. The recovery grid evaluates N in the set {20, 40, 80, 160, 320}. The IBL single unit is shown at N = 40, the grid column nearest the empirical median of about 53 decision-window trials per session (Fig 6c). The “about 40 trials” of this figure and the “about 53 trials” of Fig 6c are the same single-unit trial budget at two resolutions: 40 is the nearest grid bin, 53 is the empirical median.



Supplementary Figure 2. Diffusion-sigma sensitivity and intermediacy detail.

Expansion of main Fig 2c and 2d. (a) Recovery rate versus ramp diffusion sigma across the trial-count grid (N = 40, 160, 320) at 20 Hz with K = 50, with the canonical sigma = 0.4 marked; shaded bands are within-cell 95 percent confidence intervals. (b) Recovery split by true generator: true-step and true-ramp recovery versus sigma at N = 320 and N = 160, showing that step recovery stays high while ramp recovery falls as sigma grows. (c) The sigma sensitivity shifts with firing rate but is invariant to the latent grid resolution K (fr = 20 N = 160; fr = 40 N = 160; fr = 40 N = 80; K = 25, 50, 75 agree at sigma = 0.4). (d) Probability of selecting the step model versus generator stepiness at N = 160 and N = 320, with the non-identifiable band shaded and 95 percent confidence bands; the step-selection probability crosses 0.5 near stepiness 0.78.

Source: results/ramp_validation_robustness.csv (panels a to c) and results/phase1_recovery_stepiness.csv (panel d). Cross-reference: Fig 2c, 2d.

Supplementary Table 1. Forward-likelihood validation of the model engines.

Plain statement: the step-model and ramp-model likelihood engines agree with independent reference implementations to the limit of 64-bit floating-point precision, so the engine itself cannot change which model is selected.

Each engine's forward log-likelihood is compared against an independent reference: exhaustive enumeration of every latent path (small grid, exact) and the score from hmmlearn PoissonHMM at the production grid. The absolute differences are about 1e-15 of the about 1e3 absolute log-likelihoods, which is the 64-bit round-off floor.

comparison	absolute log-likelihood difference	relative error (difference / abs log-likelihood)
RAMP engine vs exhaustive path enumeration (K = 5)	2.13e-14	abs log-likelihood not stored
RAMP engine vs hmmlearn PoissonHMM score (K = 50)	3.64e-12	abs log-likelihood not stored
STEP engine vs hmmlearn PoissonHMM score	2.27e-13	2.61e-16

Source: results/ramp_validation_forward.json (the two RAMP rows) and results/phase1_recovery_validation.json (the STEP row). Cross-reference: Fig 1c.

Footnote. Model selection operates on per-trial cross-validated log-likelihood gaps of about 1e-6. The largest difference here (3.64e-12) is about 5 to 6 orders of magnitude smaller, so none of these differences can flip a step-versus-ramp decision.

Supplementary Table 2. Movement-control calibration.

Plain statement: before trusting any movement control, we test each candidate on synthetic data with a known answer. A valid control returns chance when there is no signal, keeps a movement-orthogonal signal, and removes a pure-movement signal. Only linear passes all three cleanly; the flexible controls fail in one direction or the other.

Each control is scored on three injection tests built from the real movement covariance. Values are decode AUC. "Preserve retained" is the fraction of the uncontrolled above-chance signal that survives (uncontrolled preserve AUC = 0.588). n = 28 co-recorded MRN sessions; each value is the mean over 14 synthetic repeats.

control	no-signal AUC (target about 0.50)	preserve AUC	preserve retained	remove AUC (target about 0.50)	verdict
none	0.483	0.588	1.00	0.671	uncontrolled reference (removes no movement)
linear	0.507	0.614	1.31	0.506	valid (recommended)
ridge	0.502	0.609	1.25	0.551	valid but removes less movement (under-removes)
pca	0.501	0.607	1.22	0.523	passes tests, under-removes on real data (see below)
expanded	0.500	0.545	0.51	0.477	over-corrects (fails preserve, eats orthogonal signal)
crossfit	0.466	0.479	-0.24	0.481	degenerate at this trial count (fails preserve)

Real-data companion (decisive evidence). On the real Brain-Wide Map data, of 301 choice-selective cells the number whose choice signal survives movement control is: expanded 66 (over-removes), linear 133 (valid), pca 193 (under-removes). The injection tests flag pca only mildly (remove AUC 0.523), but the real-data count shows pca leaves the most movement, confirming linear as the trustworthy middle.

Source: results/referee_response/proper_control/calibration.csv (the three tests) and results/referee_response/corrected_results/overcorrection_contrast.csv (the survivor counts). Cross-reference: Fig 3.

Supplementary Table 3. Confound cascade under the valid control with within-region FDR.

Plain statement: starting from choice-selective cells, we ask how many keep a choice signal after controlling for movement, after controlling for stimulus or prior, and after all three filters at

once. Under the valid linear control with within-region false discovery rate, a small set of superior colliculus cells survives all three.

Counts are cells passing each filter at within-region false discovery rate $q < 0.05$. The stages are parallel single-confound filters; triple-coded is their intersection with the leading (pre-movement) filter, not a nested funnel. The all-regions stage counts match Fig 5a exactly (choice-selective 102, movement-independent 5, stimulus-or-prior-independent 13, triple-coded 3, all in SCm).

IBL (movement control = wheel plus body pose):

region	high-FR cells	choice-selective	movement-independent	stim-or-prior-independent	triple-coded
MRN	777	47	0	0	0
SCm	427	27	5	11	3
SNr	47	14	0	2	0
GRN	184	6	0	0	0
IRN	185	8	0	0	0
all regions	1620	102	5	13	3

Steinmetz 2019 replication context (movement control = wheel plus face motion plus pupil, no body pose):

region	high-FR cells	choice-selective	movement-independent	stim-or-prior-independent	triple-coded
MRN	223	65	23	0	0
SC	168	45	16	0	0
SNr	79	9	8	0	0
all regions	470	119	47	0	0

Source: results/referee_response/corrected_results/corrected_cascade.csv (scheme = within-region BH; control = published(linear) for IBL, linear(recomputed) for Steinmetz). Cross-reference: Fig 5a (IBL), Fig 6b (Steinmetz is the replication context).

Footnote. The IBL movement control is the published wheel-plus-body-pose linear control. A separate recomputed 4-regressor linear control leaves more movement-independent cells (all-regions 17, triple 4) because it uses slightly different movement regressors; the figure and this table use the published wheel-plus-body-pose control (movement-independent 5, triple 3). Steinmetz has no body pose, which is why it retains more movement-independent cells (47), yet still yields 0 triple-coded because no cell is stimulus-or-prior-independent there.

Supplementary Table 4. Robustness of the at-the-boundary decode to analysis choices.

Plain statement: the MRN movement-controlled decode sits at the 0.57 recoverability boundary, and it stays there across every analysis choice we varied (firing-rate floor, analysis window, time bin).

Under the valid linear control, the MRN decode AUC and its 95 percent bootstrap confidence interval across settings. “vs bar” states whether the confidence interval clears, straddles, or falls below the 0.57 bar. The pooled per-cell effect (SD) with a cell-clustered 95 percent confidence interval is shown alongside. The peri-movement window is a positive control (a movement-locked window measures the wheel turn) and is not a decision result.

setting	decode AUC	95% CI	vs 0.57 bar	pooled effect (SD)	95% CI
firing-rate floor 10 Hz	0.581	[0.539, 0.622]	straddles	-0.013	[-0.051, 0.025]
firing-rate floor 15 Hz	0.581	[0.538, 0.627]	straddles	-0.013	[-0.051, 0.025]
firing-rate floor 25 Hz	0.562	[0.510, 0.613]	straddles	-0.013	[-0.051, 0.025]
window early	0.562	[0.521, 0.602]	straddles	-0.055	[-0.096, -0.016]
window deliberation	0.581	[0.539, 0.622]	straddles	-0.013	[-0.051, 0.025]
window peri (positive control)	0.640	[0.591, 0.691]	clears	0.110	[0.072, 0.149]
time bin 10 ms	0.579	[0.538, 0.620]	straddles	-0.008	[-0.045, 0.030]
time bin 20 ms	0.578	[0.535, 0.620]	straddles	-0.011	[-0.049, 0.027]
time bin 50 ms	0.573	[0.530, 0.614]	straddles	-0.019	[-0.057, 0.020]

Source: results/referee_response/corrected_results/corrected_robustness.csv (control = linear). Cross-reference: Fig 4.

Footnote. The triple-coded survivor count under the canonical setting (within-region FDR, valid control) is 3 cells, all in SCm (Supplementary Table 3). The corrected robustness sweep recomputed the decode AUC and the pooled effect at each setting but did not recompute the full cascade per setting, so a per-setting triple-coded count is not tabulated here; the canonical value is 3.

Supplementary Table 5. Full-cohort simultaneity census.

Plain statement: recovering single-trial step-versus-ramp dynamics at the population level needs about 120 simultaneously recorded cells. Across the full cohort, no session reaches that count.

Per region, the number of simultaneously recorded good-QC units (clusters with the IBL good-unit QC label, at any firing rate) summed across the probes of each session. The full cohort is 205 insertions across 176 sessions, with 259 region-sessions in total. None reaches the 120-cell requirement.

region	region-sessions	median simultaneous good-QC units	max simultaneous good-QC units	sessions reaching 120
MRN	118	18	91	0
SCm	68	17.5	90	0
GRN	17	28	79	0
IRN	33	18	73	0
SNr	23	3	36	0
all regions	259	–	91 (global max)	0 of 259

Source: results/referee_response/full_census.csv. Cross-reference: Fig 6a. The maximum count is the full-cohort good-QC simultaneous maximum (91 units); the recoverability requirement is about 120.