Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Rui Yamasaki

rui.yamasaki0724@gmail.com

Independent Researcher https://orcid.org/0009-0001-5124-2806

Article

Keywords:

Posted Date: June 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10037105/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

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

Rui Yamasaki*
Independent Researcher
*Correspondence: rui.yamasaki0724@gmail.com

Abstract

Across neuroscience, movement is routinely subtracted from large-scale recordings to isolate a purer cognitive signal, such as a forming decision. We ask whether that subtraction can be trusted. Using simulations with a known signal, we show that the movement controls in common use distort the signal they are meant to reveal: flexible nonlinear controls over-correct, erasing genuine decision signal when trials are few, while dimensionality-reduced controls under-correct, leaving low-variance movement behind. A linear control, validated against ground truth, removes movement without eating signal. We pair this with a recoverability framework that specifies, from firing rate, trial count, and spiking variability, when a gradual ramp can be told apart from a discrete step on single trials. Reanalyzing midbrain and hindbrain choice regions of a brain-wide map, and replicating in a second dataset, a properly controlled decision signal sits at the boundary of recoverability, not below it as an over-corrected analysis concludes, with a small but real movement-independent choice code in a subset of superior colliculus neurons. The simultaneity needed to recover single-trial dynamics is not reached at current yields. What can be inferred about single-trial decisions is set as much by how we analyze and record as by the brain.

Introduction

It has become routine to record from tens of thousands of neurons while an animal makes decisions, and then to ask what each region contributes to the choice. A central obstacle is that animals move, and movement drives neural activity almost everywhere in the brain. To recover a clean cognitive signal, the now-standard step is to regress movement out of the activity and treat what remains as the decision-related part. This single analysis choice underlies a large fraction of recent claims about where and how the brain computes a decision[1–3]. The practice grew from the now-well-established finding that uninstructed movement and behavioural state pervade neural activity across the brain[4–6].

We ask a question that is rarely tested directly: does that subtraction work? Removing movement is only safe if the control removes movement and nothing else. If a decision signal is itself correlated with the developing movement, as it often will be when an animal is about to act on its choice, an aggressive control can remove the decision signal along with the movement, and a researcher would wrongly conclude the region carries no decision. The opposite failure is also possible: a control that captures only the obvious, high-variance components of movement can

leave subtler movement behind and credit it to cognition. Neither failure is visible from the data alone, because the ground truth is unknown.

We study this in a concrete and contested case: whether the apparent ramp of decision-related activity in midbrain and hindbrain reflects a graded accumulation of evidence on single trials, or a discrete step that only looks like a ramp once averaged across trials[7]. The same recordings have historically supported opposite conclusions depending on how baseline firing, spiking noise, and cross-validation were handled[8–10], so the answer is known to be fragile. The distinction matters for how a downstream region should read the signal, and it cannot be settled from trial averages, because a discrete step with a variable onset time produces the same smooth average as a gradual ramp. Settling it requires single-trial latent-variable modeling, which in turn requires that the data carry enough information to tell the two dynamics apart.

We study this in the midbrain and hindbrain regions named as choice-encoding in the International Brain Laboratory Brain-Wide Map[11]. This paper makes three contributions, in order of breadth. First, using simulations in which we inject a decision signal of known size and known correlation with each session's real movement, we show that the movement controls in common use systematically distort the estimate, and we identify, by calibration against the ground truth, which control is trustworthy. Second, we provide a recoverability framework that states, before any control is applied, when single-trial stepping and ramping can be distinguished at all, given a region's firing rate, the number of trials, and the variability of spiking. Third, we apply the validated control and the framework to a brain-wide activity map and an independent dataset, and we report, at the strength the data support, where the decision signal actually lies. We package the calibration as an open diagnostic so that any laboratory can test its own controls against ground truth before trusting them.

Results

A trial-averaged ramp is consistent with two different single-trial dynamics

We first fixed the question and validated the machinery used to answer it. On single trials, decision-related firing could rise gradually toward a bound (a ramp, as in a drift-diffusion accumulator) or jump once at a variable time (a step). Averaged across trials, both produce the same smooth rise (Fig 1a), so the trial average cannot distinguish them. We implemented both as hidden Markov generators (Fig 1b), stepping as a jump process[12] and ramping as a discretised drift-diffusion-to-bound process[13], and verified the forward-likelihood engine against analytic and brute-force references to machine precision (Fig 1c), so that any later inability to tell the models apart reflects the data, not the fitting.

Whether the distinction is recoverable depends on firing rate and trial count

Before asking what a given dataset shows, we asked what any dataset could show. We simulated stepping and ramping spike trains across a grid of firing rates and trial counts under realistic, overdispersed (negative-binomial) spiking[14], then fit both models and measured how often the true generator was recovered (Fig 2a). Recovery improves with firing rate and trial count, as expected, and the 0.80-recovery contour defines a practical threshold: a per-cell choice decode of about 0.57 area under the ROC curve, equivalently a per-cell separation of about 0.24 in units of the spike-count standard deviation. Below this, the two dynamics are not reliably distinguishable on single trials regardless of method. The boundary is not a fixed line: dynamics that blend

stepping and ramping are unresolvable across the whole intermediate range, and a noisier (higher-diffusion) ramp is harder to separate from a step, to the point that a ramp that slams into the bound is statistically identical to a step (Fig 2c, d; Supplementary Fig 2). Two further results frame the rest of the paper. Assuming Poisson spiking, as many pipelines implicitly do, inflated apparent recovery and collapsed the distinction (Fig 2b; Supplementary Fig 1), so the spiking model matters. And recovering single-trial dynamics at the population level, rather than cell by cell, required on the order of 120 simultaneously recorded units in these regions (Fig 6a), a number we return to below.

Movement controls over-correct and under-correct, and a calibration tells them apart

We then tested the step every such analysis depends on. We took real activity from the choice regions, shuffled the choice labels to create a background with no genuine decision signal, and injected a decision signal of known size that was correlated with each session's real movement (wheel velocity and video-tracked body pose[15]) at a controlled level. Because the injected movement was drawn from each session's own movement covariance, the test reflects the variance structure of real confound rather than an idealized version of it. We then ran candidate movement controls and measured how much of the known signal survived (Fig 3a).

The result is a clear double failure. A flexible nonlinear control, of the kind increasingly used to be thorough about movement, over-corrected: at the roughly 30 trials per session typical of these surveys, it retained only 51 percent of a decision signal that was orthogonal to movement and therefore should have been left untouched, because fitting fourteen movement regressors on so few trials overfits and absorbs signal it has no business removing. A control that first reduced movement to its top principal components had the opposite failure: it under-removed, because the movement that carried the choice confound lived in low-variance dimensions that the reduction discarded. Critically, this under-removal was invisible to a calibration that injected high-variance synthetic movement, and only appeared when the injected movement matched the real confound's structure, which is why a calibration must use a laboratory's own movement data.

To choose objectively, we defined three ground-truth tests that any valid control must pass: it must return chance when no signal is present, preserve a signal that is orthogonal to movement, and remove a signal that is pure movement (Fig 3b). Of the controls we examined, a simple linear regression of movement out of activity, and only that, passed all three while removing the most real movement (Supplementary Table 2). We refer to this calibrated procedure throughout as the valid control. A valid control could register a genuine threshold-sized signal (a recovered decode of about 0.57 for an injected effect at the per-cell bar), whereas the over-correcting control was capped below the bar and could not register such a signal no matter how large it was (Fig 3c). We provide this calibration as an open, documented diagnostic (`mcdiag`) that takes a laboratory's firing rate, trial count, and movement regressors and reports which control is valid for its data, where each candidate fails, and whether the recording can resolve single-trial dynamics at all.

Properly controlled, the decision signal sits at the boundary, not below it

The choice of control changes the scientific conclusion. Under the over-correcting control, the movement-controlled decode of choice in the midbrain reticular nucleus fell to 0.528, below the

recoverability bar, which would license the conclusion that the region's choice-related activity is essentially action rather than deliberation. Under the valid control, the same data give a decode of 0.581 (95% confidence interval 0.539 to 0.622), which sits at the recoverability bar rather than below it (Fig 4a). The independent dataset agreed: midbrain reticular nucleus decode of 0.621 (0.542 to 0.722) under the valid control versus 0.541 under the over-correcting one. Across all regions examined, no decode cleared the bar, and the midbrain signal sat at it in both datasets, significant against chance but with a confidence interval that straddles the threshold, so we describe it as at the boundary and never above it. Secondary regions fell below the boundary (for example superior colliculus 0.543, gigantocellular reticular nucleus 0.443). The cost of over-correction was concrete: across the brain-wide dataset, the over-correcting control erased 67 of the 133 cells (about half) that a valid control identifies as carrying choice signal independent of movement, and pushed every region's decode below a bar the valid control places it at or near. The principal-component control failed in the opposite direction, inflating the movement-independent count to 193 cells by leaving low-variance movement in; the valid linear control sits between these two failures, at 133 cells (Fig 4b).

A small but genuine movement-independent choice code, clearest in superior colliculus

We next asked whether any individual cells carry a decision signal that survives a valid movement control and is also independent of stimulus and prior. Scoring per-cell choice selectivity by ROC[16] and applying Benjamini-Hochberg false-discovery-rate control[17] within each region, a small set of cells survived as movement-independent choice cells (Fig 5a; Supplementary Table 3), and the strictest subset, independent of movement, stimulus, and prior together, comprised three to four neurons, all in the superior colliculus. These are not marginal: their choice decode was about 0.79 to 0.81 before control and stayed at 0.74 to 0.75 under the valid linear control (0.787 to 0.749, 0.810 to 0.736, 0.794 to 0.736; Fig 5b), well above chance and the recoverability bar, so their choice signal is not explained by movement. More broadly, 39 cells (36 in IBL, 3 in Steinmetz) held a choice decode at or above 0.70 under both the linear and PCA controls. The picture across the population is a bulk of cells at chance (median decode near 0.50) with a real selective tail (7.6 percent above 0.65; Fig 5c). We are explicit about the limits: at the population level the superior colliculus effect is marginal and sensitive to the exact valid control used, so the robust claim is the existence of a small number of clean single-cell choice signals, not a strong population code.

The simultaneity to recover single-trial dynamics is not reached at current yields

Finally, even where a decision signal exists, recovering its single-trial dynamics is a harder requirement than detecting it. At the trial counts available per session, single-cell recovery stays below the 0.80 target (Fig 6c). A second obstacle blocks population recovery independently of effect size: single-trial latent methods such as LFADS need neurons recorded at the same time, on the same trials[18]. A brain-wide map earns its statistical power the opposite way, by pooling neurons recorded from different animals on different days into a pseudo-population, a composite with no joint single-trial activity to read. We counted, across the full brain-wide cohort (205 insertions, 176 sessions), how many good-quality units were recorded simultaneously in each region (Fig 6a; Supplementary Table 5). The maximum was 91 simultaneous units, in the midbrain reticular nucleus, against the roughly 120 the recoverability framework requires, and no region-session of 259 reached the threshold. Only by abandoning quality control and counting noise and multi-unit clusters did the count exceed 120, but those clusters carry no recoverable

single-cell rate signal. The wall therefore holds, and is strengthened by using the full cohort rather than a sample; the brain-wide map's companion prior-information analysis registers the same thinness, cautioning that the number of significant sessions per region limits robust regional claims[19]. The independent dataset reproduced the at-the-boundary decode picture (Fig 6b), so the conclusion is not specific to one survey.

Discussion

The central message is methodological and general: what a large-scale recording can tell you about a single-trial decision signal is set as much by the movement control and the recording yield as by the brain. We showed, with ground truth, that the movement controls in common use distort decision-signal estimates in opposite directions, that the distortion is large enough to flip a conclusion, and that a simple linear control calibrated against a laboratory's own movement is the defensible choice. Because regressing out movement now underlies a wide range of claims about cognitive signals in neural data, this matters beyond the specific case we studied.

The specific case illustrates the stakes. An over-correcting control made the midbrain choice signal look like pure action, below the threshold for any single-trial decision content. The same data, properly controlled, place the signal at the boundary of recoverability, with a small set of superior colliculus neurons carrying a genuinely movement-independent choice signal. We deliberately do not claim more. No region's decode clears the recoverability bar, the population-level superior colliculus effect is marginal, and the simultaneity required to recover single-trial dynamics, as opposed to detecting a signal, is not reached at present yields. The honest conclusion is that the question of stepping versus ramping in these regions sits right at the edge of what current brain-wide recordings can resolve, and that an earlier impression of a clear answer in either direction was an artifact of the control.

These results connect two literatures. Work showing that movement dominates large-scale neural activity motivated the now-standard practice of regressing it out; our results show that the practice, done with the aggressive tools that the same insight encourages, can erase the cognitive signal it was meant to expose. The recoverability half carries its own lesson: on realistically variable spike trains a Poisson-emission model, the field's default, collapses to chance and cannot tell the generators apart at any trial count, so a Poisson assumption silently destroys the distinction it is used to test. Quantifying when a single-trial coding question is answerable has precedent for other neural signals[20], and recoverability is logically prior to any account of the dynamics, including recent proposals that stepping and ramping are endpoints of a regime transition[21]: a continuum, a transition, and a discrete step are equally unsupportable in data that cannot tell them apart. The long debate over stepping versus ramping here meets a recording-yield limit: even with the right model and the right control, the populations are not yet large enough to recover single-trial dynamics in these regions.

The work has limits. The recoverability thresholds are derived under a specific family of generative models and spiking assumptions, and laboratories with different regimes should recompute them, which our diagnostic supports. The validated control is the best of those we tested, not a proof that no better control exists; the contribution is the calibration procedure that selects it, which others can apply to their own candidates. And the applied finding is a careful boundary result, not a discovery of new dynamics. One caveat on the replication: the second dataset's movement control omits body pose, so it removes less movement and its residual is, if

anything, an over-estimate of the genuine signal. The work is also constructive. It names the recording a dedicated experiment would need to settle the question, a per-cell choice separation of about 0.57 AUC in a sufficiently large simultaneous population, a target neither dataset here reaches but one that simultaneous recordings of hundreds of co-recorded neurons in cortex[22] and emerging brain-wide single-trial decoding[23] are beginning to meet. We see the most durable contribution as the diagnostic itself: a way to know, before trusting a movement-controlled estimate, whether the control is valid and whether the recording can answer the question at all.

Methods

Datasets

We analyzed the International Brain Laboratory Brain-Wide Map[11] and, as an independent replication, the Steinmetz et al. dataset of distributed coding across the mouse brain[24]. Both are openly available (see Data availability). We streamed spike times and clusters per insertion rather than downloading raw voltage, and computed all features on the fly.

Unit selection and trial definition

We restricted analysis to the midbrain and hindbrain choice regions identified in the brain-wide map. Units passed standard quality control (a firing-rate floor, interspike-interval violation limits, and stability), with a canonical 10 Hz floor and sensitivity analyses at 15 and 25 Hz. For each unit we formed per-trial spike counts in a deliberation window aligned to stimulus and movement onset, with early and peri-movement windows used as controls. To weaken the natural correlation between stimulus, choice, and movement, the primary decode analyses used decorrelated trials (error trials and zero-contrast trials), on which choice is not determined by stimulus.

Recoverability framework

We implemented stepping and ramping generators as hidden Markov models (a discrete latent jump with variable onset[12], and a drift-diffusion-to-bound process[13] on a latent grid), with Poisson or negative-binomial emissions[14], and validated the forward-likelihood computation against exhaustive path enumeration and an independent maintained HMM library (agreement to machine precision; Supplementary Table 1). We simulated spike trains across a grid of firing rates and trial counts under negative-binomial spiking, fit both models with cross-validation and a fully Bayesian information criterion, and recorded the fraction of simulations in which the true generator was recovered. The 0.80-recovery contour defined the per-cell threshold (about 0.57 choice AUC, about 0.24 SD). We repeated the grid under Poisson spiking to quantify the effect of model misspecification, and estimated the number of simultaneous units required for population-level recovery.

Movement controls and their calibration

Movement regressors comprised wheel velocity and tracked body-pose features. We considered candidate controls: no control; linear regression of movement out of activity; ridge-regularized and cross-fitted versions of an expanded nonlinear control; a control reducing movement to its leading principal components; and the full expanded nonlinear control. To calibrate, we constructed a choice-null background by shuffling choice labels within the real data, preserving

spiking statistics and the real movement design matrix, then injected a choice-dependent signal of controlled per-cell size (in SD units) and controlled correlation with each session's own movement subspace. We applied each candidate control and measured three quantities: the recovered effect when no signal was injected (must be chance), when a movement-orthogonal signal was injected (must be preserved), and when a pure-movement signal was injected (must be removed). A control passing all three with the greatest real-movement removal was selected as valid; linear regression met this criterion. On the brain-wide data the calibration recommended the linear control, flagged the expanded control as over-correcting (it failed the preservation test, retaining about half of a movement-orthogonal signal), and flagged the principal-component and ridge controls as under-removing (they passed the hard tests but removed less movement than linear and inflated the movement-independent count); the principal-component control fails the removal test outright when movement carries strong low-variance structure, as in the synthetic ground truth.

Decoding, per-cell effects, and the cascade

For each region we computed a cross-validated population decode of choice from movement-controlled activity, with confidence intervals by bootstrap and significance by permutation. Per-cell choice selectivity used ROC analysis[16] with a permutation null (at least 1000 shuffles). We separately computed a per-cell choice effect in SD units with a cell-clustered bootstrap[25] that resamples cells rather than trials. Movement regressors comprised wheel velocity and DeepLabCut body-pose features[15]. To identify cells with choice signal independent of movement, stimulus, and prior, we tested each filter per cell by permutation and applied Benjamini-Hochberg false-discovery-rate control, both dataset-wide and within each region, reporting both.

Simultaneity census

Across the full cohort we counted, per session and region, the number of simultaneously recorded units passing quality control, and report the per-region median, maximum, and the number reaching the population-recovery requirement, alongside the counts obtained without quality control for comparison.

Diagnostic tool

The calibration above is provided as an open, documented, installable diagnostic (`mcdiag`). Given a laboratory's per-trial activity, choice labels, and movement regressors, it runs the three ground-truth tests using that laboratory's own movement structure, recommends a valid control or reports that none of the candidates is valid for the data regime, and reports whether single-trial decision dynamics are recoverable given the firing rate, trial count, and simultaneity. It reproduces the control verdicts reported here on the brain-wide data as a built-in validation.

Statistics and reproducibility

All tests were two-sided unless otherwise noted. Exact P values are reported, including for non-significant results, except where a resampling test reaches its resolution floor, in which case P is given as a bound ($P < 0.001$ at 1000 resamples). Effect sizes are reported in standard-deviation units as point estimates with 95 percent bootstrap confidence intervals, computed with a cell-clustered bootstrap (resampling co-recorded cells, with n the number of cells per region and trial

counts as stated) to avoid the pseudoreplication that inflates trial-level tests. Population decoding reported out-of-sample AUC from cross-validated classifiers. Per-cell selectivity used a permutation null (at least 1000 shuffles per cell) with Benjamini-Hochberg false-discovery-rate control. The movement control was selected by ground-truth calibration as described, and the central results were tested as sensitivity analyses across firing-rate floor, analysis window, temporal bin size, false-discovery-rate scope, and the two calibration-passing controls; the corrected decode sits at the recoverability boundary across these settings (Supplementary Table 4), and where the two valid controls disagree (the superior colliculus population effect) this is reported openly rather than resolved in favour of the larger estimate. Analysis code, the diagnostic, and scripts that regenerate every figure are available (see Code availability).

Manuscript preparation

An AI assistant (Claude, Anthropic) was used to support drafting and language editing of this manuscript. All study design, analyses, interpretation, and conclusions are the author's own; the author reviewed and approved the final text and is solely responsible for its content.

Data availability

This study is a reanalysis of two publicly available datasets and generated no new experimental data. The International Brain Laboratory Brain-Wide Map data were accessed through the Open Neurophysiology Environment (ONE) API against the public IBL database and are openly available via the AWS Open Data Registry (https://registry.opendata.aws/ibl-brain-wide-map/), with access and download instructions at the documented 2025 data-release notebook[11]. The replication data are from Steinmetz et al.[24] and are openly available on figshare (https://doi.org/10.6084/m9.figshare.9598406; ONE-format files at figshare 9974357). The derived intermediate results (binned spike-count matrices, per-cell selectivity and confound-control outputs, the calibration injection grids, and recovery-simulation grids) needed to regenerate every figure are deposited with the analysis code (see Code availability).

Code availability

All custom code is openly available without restriction. This includes the identifiability-recovery framework, the forward-likelihood implementations and their validation fixtures, the movement-control calibration and decoding pipelines, the open diagnostic tool (mcdiag), the cross-dataset adapter, and the one-command figure-regeneration scripts, together with a pinned environment manifest and a README documenting installation, a runnable demonstration on bundled test data, and the test dataset itself. The repository is public at https://github.com/Rui-Yamasaki/mcdiag and is released under the BSD-3-Clause licence; a versioned snapshot is archived at Zenodo with the permanent DOI https://doi.org/10.5281/zenodo.20686415, which includes the separately installable diagnostic (mcdiag). The archived snapshot will be updated to the accepted version on acceptance. There are no restrictions on access or reuse.

References

1. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).

2. Roitman, J. D. & Shadlen, M. N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489 (2002).
3. Brunton, B. W., Botvinick, M. M. & Brody, C. D. Rats and humans can optimally accumulate evidence for decision-making. *Science* **340**, 95–98 (2013).
4. Stringer, C. et al. Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, eaav7893 (2019).
5. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci.* **22**, 1677–1686 (2019).
6. Wang, Z. A. et al. Brain-wide analysis reveals movement encoding structured across and within brain areas. *Nat. Neurosci.* **29**, 147–158 (2026).
7. Latimer, K. W., Yates, J. L., Meister, M. L. R., Huk, A. C. & Pillow, J. W. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* **349**, 184–187 (2015).
8. Shadlen, M. N. et al. Comment on "Single-trial spike trains in parietal cortex reveal discrete steps during decision-making". *Science* **351**, 1406 (2016).
9. Latimer, K. W. et al. Response to Comment on "Single-trial spike trains in parietal cortex reveal discrete steps during decision-making". *Science* **351**, 1406 (2016).
10. Zoltowski, D. M., Latimer, K. W., Yates, J. L., Huk, A. C. & Pillow, J. W. Discrete stepping and nonlinear ramping dynamics underlie spiking responses of LIP neurons during decision-making. *Neuron* **102**, 1249–1258.e10 (2019).
11. International Brain Laboratory et al. A brain-wide map of neural activity during complex behaviour. *Nature* **645**, 177–191 (2025).
12. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
13. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**, 59–108 (1978).
14. Goris, R. L. T., Movshon, J. A. & Simoncelli, E. P. Partitioning neuronal variability. *Nat. Neurosci.* **17**, 858–865 (2014).
15. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
16. Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis. Neurosci.* **13**, 87–100 (1996).
17. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
18. Pandarinath, C. et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
19. Findling, C. et al. Brain-wide representations of prior information in mouse decision-making. *Nature* **645**, 192–200 (2025).
20. Mohl, J. T., Caruso, V. C., Tokdar, S. T. & Groh, J. M. Sensitivity and specificity of a Bayesian single trial analysis for time varying neural signals. *Neurons Behav. Data Anal. Theory* **3**, 11880 (2020).

21. Luo, T. Z. et al. Transitions in dynamical regime and neural mode during perceptual decisions. *Nature* **646**, 1156–1166 (2025).
22. Steinemann, N. A. et al. Direct observation of the neural computations underlying a single decision. *eLife* **12**, RP90859 (2024).
23. Bondy, A. G. et al. Brain-wide coordination of internal signals during decision-making. Preprint at https://doi.org/10.1101/2024.08.21.609044 (2024).
24. Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576**, 266–273 (2019).
25. Field, C. A. & Welsh, A. H. Bootstrapping clustered data. *J. R. Stat. Soc. B* **69**, 369–390 (2007).

Acknowledgements

This work used openly released data, and the author thanks the International Brain Laboratory and the Steinmetz laboratory and their collaborators for collecting and sharing the datasets reanalysed here. All analyses were performed on commodity hardware, by streaming and caching only the trimmed data objects required. The author is also grateful to Mrs. Jennifer Johns, his former high school AP Biology teacher, whose encouragement first inspired him to pursue science. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

R.Y. conceived the study, designed the recoverability and movement-control calibration framework, implemented all analyses and software including the diagnostic tool, performed the reanalysis and validation, produced the figures, and wrote the manuscript.

Competing interests

The author declares no competing interests.

Materials & correspondence

Correspondence and requests for materials should be addressed to R.Y.

Figure legends

Figure 1. The ambiguity, the generative models, and the inference-engine validation. (a) Latent single-trial firing-rate trajectories from the stepping generator (top, blue) and the ramping generator (bottom, orange), at a representative 20 Hz operating point over a 1000 ms window (41 bins of 25 ms). Each sub-panel shows 10 faint single trials and the family mean as a thick line; the dashed black line is the common trial-average that both families produce by construction. The y axis is latent firing rate (Hz) and the x axis is time in the window (ms). n = 10 single trials per family; each family mean is computed from 20000 Monte-Carlo trials. No error bars are shown. (b) Schematics of the two generators. Stepping (blue) is a two-state model that jumps once from a baseline state to an elevated state. Ramping (orange) is a drift-plus-diffusion process on a gradual latent grid that advances to an absorbing bound. This panel is a schematic and contains no data. (c) Inference-engine validation. Absolute difference in forward log-likelihood

between independent implementations, on a logarithmic x axis, as horizontal stems ending in a circle. The three comparisons are ramp versus brute-force 2.1e-14 (orange), step versus hmmlearn 2.3e-13 (blue), and ramp versus hmmlearn 3.6e-12 (orange). The grey shaded band marks the machine-precision range (1e-16 to 1e-11) and the dashed green vertical line marks the inference-relevant scale (1e-6). The x axis is the absolute forward-log-likelihood difference. Each value is the maximum absolute difference over the compared trials (one paired comparison per row); no error bars are shown. Colour coding: blue marks the stepping model and orange marks the ramping model throughout; the dashed black line in (a) is the shared trial-average of both families; the green line in (c) is the inference-relevant reference. Print width about 180 mm (double column).

Figure 2. The step-versus-ramp identifiability map. Recovery rates are fractions of correctly recovered generators, and no error bars are shown (the corresponding confidence bands are in Supplementary Fig 2). Panels a and b aggregate R = 40 repeats per grid cell; panels c and d plot per-trial recovery fractions (about 80 simulated trials per point in c, 40 per stepiness and trial count in d). (a) Recovery map under realistic negative-binomial emissions (overdispersed, Fano >= 1.5; ramp diffusion sigma = 0.4). Heatmap of the step-versus-ramp 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). The white dashed contour marks the 0.80 recovery level. The circled point with a callout is the IBL single-unit operating point (the N = 40 grid column, nearest the empirical median of about 53 trials per session, at 20 Hz, recovery 0.56), which would require pooling to N of about 320 to reach 0.80. (b) Recovery rate versus pooled trial count N at 20 Hz on the same data, for two fitters: a misspecified Poisson fit (grey squares) and a matched negative-binomial fit (green circles). At N = 40, 160, 320 the Poisson values are 0.48, 0.53, 0.52 and the negative-binomial values are 0.56, 0.73, 0.80. The dotted line is chance (0.50) and the dashed line is the 0.80 target. (c) Recovery rate versus ramp diffusion sigma (firing rate 20 Hz, pooled K = 50), for three series: N = 320 pooling both families (grey circles), N = 160 both families (dark squares), and N = 160 true-ramp trials only (orange dashed triangles). The dotted green vertical line marks the canonical sigma = 0.4; the dotted line is chance (0.50) and the dashed line is the 0.80 target. (d) Probability that the step model is selected versus generator stepiness (0 is pure ramp, 1 is pure step). Faint blue lines are individual pooled-N curves and the bold blue line is their average. The grey band marks the non-identifiable region where step and ramp are not distinguishable; the step-selection probability crosses 0.5 near stepiness 0.78. Colour coding: cividis encodes the continuous recovery rate in (a); grey marks the misspecified Poisson fit and green the matched negative-binomial fit in (b); blue marks the step model and orange the ramping family throughout. Print width about 180 mm (double column).

Figure 3. Movement controls distort the decision-signal estimate. (a) Three control-validity tests applied to six candidate movement controls (none, linear, ridge, pca, expanded, crossfit) on synthetic data whose injected signals are drawn from each session's own movement covariance. For each control, three grouped bars give the cross-validated choice decode AUC under three injected conditions: no-signal, with no choice signal added (dotted grey); preserve, with a movement-orthogonal choice signal of 0.24 SD per cell added (solid blue); and remove, with a pure-movement signal of 0.24 SD per cell added (hatched vermillion). The dotted horizontal line is chance (AUC 0.50) and the dashed blue line is the uncontrolled (none) preserve level. The blue star marks the expanded preserve bar and the vermillion star marks the pca remove bar. n = 28 co-recorded midbrain (MRN) sessions; bar height is the mean decode AUC over 14 synthetic

repeats and error bars are 95 percent Monte-Carlo confidence intervals across repeats. (b) Movement-controlled choice decode AUC against injected per-cell effect size in SD units, one line per control, with the injected signal set to the measured choice-to-movement correlation (rho = 0.49). The dashed horizontal line is the 0.57 recoverability bar, the dotted horizontal line is chance, and the dotted vertical line is the 0.24 SD threshold injection. Colours and markers identify the control as in the legend; the linear control is drawn with a thicker line. n = 28 sessions; points are the mean over 14 synthetic repeats and shaded bands are 95 percent Monte-Carlo confidence intervals. (c) Number of choice-selective cells whose choice signal remains significant after movement control on the real Brain-Wide Map data, for three controls (expanded, linear, pca), out of 301 choice-selective cells pooled across regions. The dashed line marks the linear count. Bars are raw survivor counts (per-cell choice-AUC permutation test, $p < 0.05$, 2000 permutations); counts have no error bars. Colour coding: in (a) and (c) blue or green denotes the validated linear control context, vermillion denotes the over-correcting expanded control, and orange denotes the under-removing pca control; grey denotes the uncontrolled condition. Print width 180 mm (double column).

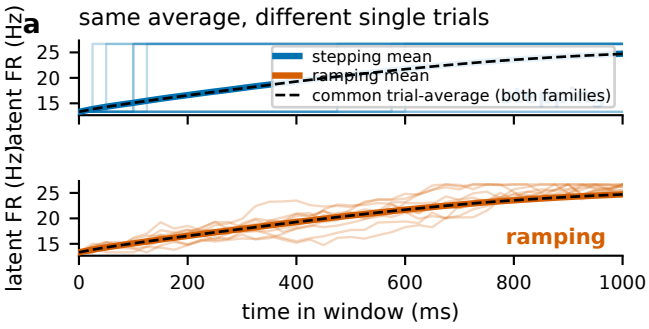
Figure 4. The decision signal at the boundary, not below it. (a) Movement-controlled choice decode AUC for each region under the valid linear control (filled blue circles) and under the over-correcting expanded control (open vermillion squares), for the IBL Brain-Wide Map (top four rows, bracket labelled IBL) and the Steinmetz 2019 dataset (bottom three rows, bracket labelled Steinmetz). Horizontal bars are 95 percent bootstrap confidence intervals resampled over sessions. The dotted vertical line is chance (AUC 0.50) and the dashed vertical line is the 0.57 recoverability bar. No region's confidence interval lies entirely above the 0.57 line. n is given as decodable sessions and pooled cells per region: IBL MRN 28 sessions, 356 cells; IBL SCm 18, 239; IBL IRN 9, 97; IBL GRN 6, 86; Steinmetz MRN 7, 223; Steinmetz SC 6, 168; Steinmetz SNr 4, 79. IBL SNr is omitted because only one session is decodable. (b) Number of choice-selective cells, of 301 pooled across IBL regions, whose choice signal remains significant after movement control, for three controls: expanded (over-removes), linear (valid middle), and pca (under-removes). The dashed line marks the linear count. Bars are raw survivor counts (per-cell choice-AUC permutation test, $p < 0.05$, 2000 permutations) and have no error bars. Colour coding: blue marks the valid linear control and vermillion marks the over-correcting expanded control in (a); in (b) vermillion marks expanded, green marks linear, and orange marks pca. Markers in (a) are distinguished by shape as well as colour (filled circle versus open square). Print width 180 mm (double column).

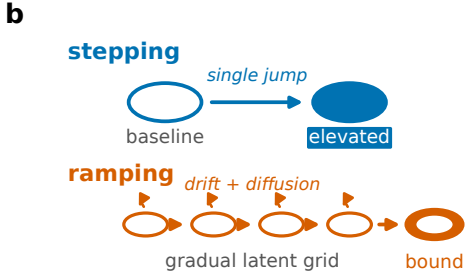
Figure 5. A small movement-independent choice code. (a) Survivor counts at four stages of the confound cascade applied under the published wheel-plus-body-pose linear movement control with within-region false discovery rate correction ($q < 0.05$), pooled across the Brain-Wide Map regions: choice-selective, movement-independent, stimulus or prior-independent, and triple-coded (leading and movement-independent and stimulus or prior-independent). Each horizontal bar gives the total count (grey) with the superior colliculus (SCm) contribution overlaid (blue); the number to the right of each bar is the total count, with the SCm count in parentheses. The movement-independent and stimulus or prior-independent stages are parallel single-confound filters and the triple-coded stage is their intersection with the leading filter, so the four counts are not a nested sequence. n = 1620 high-firing-rate cells entering the cascade; bars are counts and have no error bars. (b) Per-cell choice decode AUC before the control (raw deliberation-window rate) and after the recomputed 4-regressor linear movement control, for the

three superior colliculus triple-coded cells, shown as connected points (one line per cell). The dashed horizontal line is the 0.57 recoverability bar and the dotted horizontal line is chance (AUC 0.50). n = 3 cells; points are single-cell AUCs and have no error bars. (c) Distribution of per-cell choice decode AUC across the population after the recomputed 4-regressor linear movement control. Grey bars are cells near chance and blue bars are the selective tail (AUC above 0.65 or below 0.35). The dotted vertical line is chance (0.50) and the dashed vertical lines are 0.35 and 0.65. The text gives the median AUC and the fraction of cells above 0.65. n = 1620 cells; bars are cell counts. Colour coding: blue marks superior colliculus cells in (a) and (b) and the selective tail in (c); grey marks other regions in (a) and the near-chance bulk in (c). Panel a uses the published wheel-plus-body-pose linear control; panels b and c use the recomputed 4-regressor linear control, both calibration-passing linear controls (Fig 3, Supplementary Table 2). Print width 180 mm (double column).

Figure 6. Single-trial dynamics recovery is out of reach at current yields. (a) Maximum number of simultaneously recorded good-QC units per region across the full IBL Brain-Wide Map cohort, for the five decision-core regions ordered by their maximum. Bars are coloured by region. The dashed red line is the 120-cell requirement for population step-versus-ramp recovery. Good-QC units are clusters carrying the IBL good-unit QC label, at any firing rate. The full cohort is 205 insertions across 176 sessions, with 259 region-sessions in total (MRN 118, SCm 68, IRN 33, SNr 23, GRN 17), of which 0 reach 120 units. Bars are counts and have no error bars. (b) Movement-controlled choice decode AUC for MRN under the valid linear control, for the IBL dataset (purple) and the Steinmetz 2019 dataset (pink). Error bars are 95 percent bootstrap confidence intervals resampled over sessions. The dashed line is the 0.57 recoverability bar and the dotted line is chance (AUC 0.50). n = 28 sessions (356 cells) for IBL MRN and 7 sessions (223 cells) for Steinmetz MRN. (c) Single-cell step-versus-ramp recovery rate against pooled trial count N at 20 Hz firing rate under negative-binomial emissions (Fano 1.5 or greater), from the recovery grid. The dashed line is the 0.80 recovery target and the dotted line is chance (0.50). The green star is the IBL single-unit operating point (median about 53 trials per session, per-cell choice AUC 0.53). Points are mean recovery rates over the grid at each N. Colour coding: in (a) each region has a fixed qualitative colour (MRN green, SCm orange, GRN grey, IRN purple, SNr blue); in (b) purple marks IBL and pink marks Steinmetz; in (c) black marks the recovery curve and green marks the IBL operating point. Print width 180 mm (double column).

Figure 1 · the ambiguity, the models, and engine validation





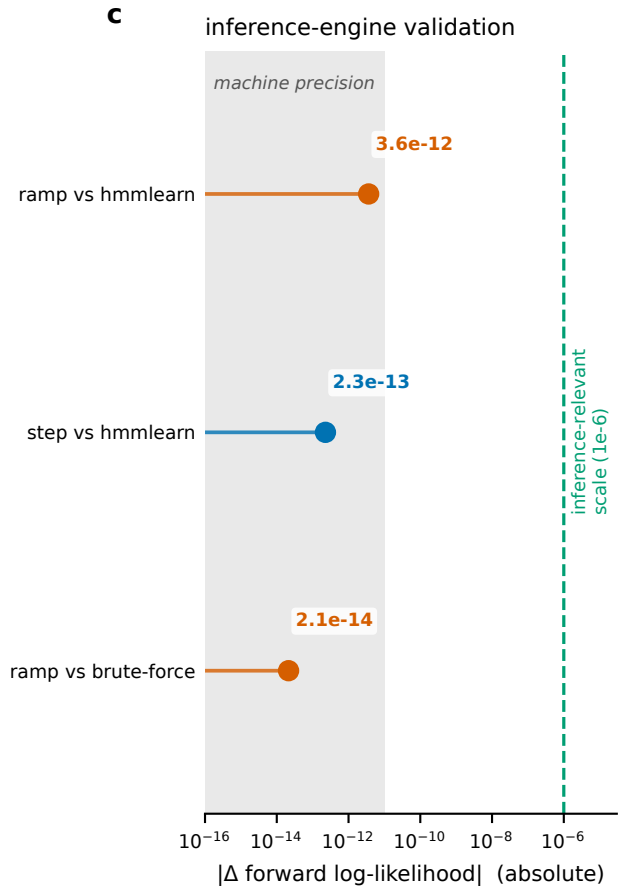


Figure 2 · the step-vs-ramp identifiability map

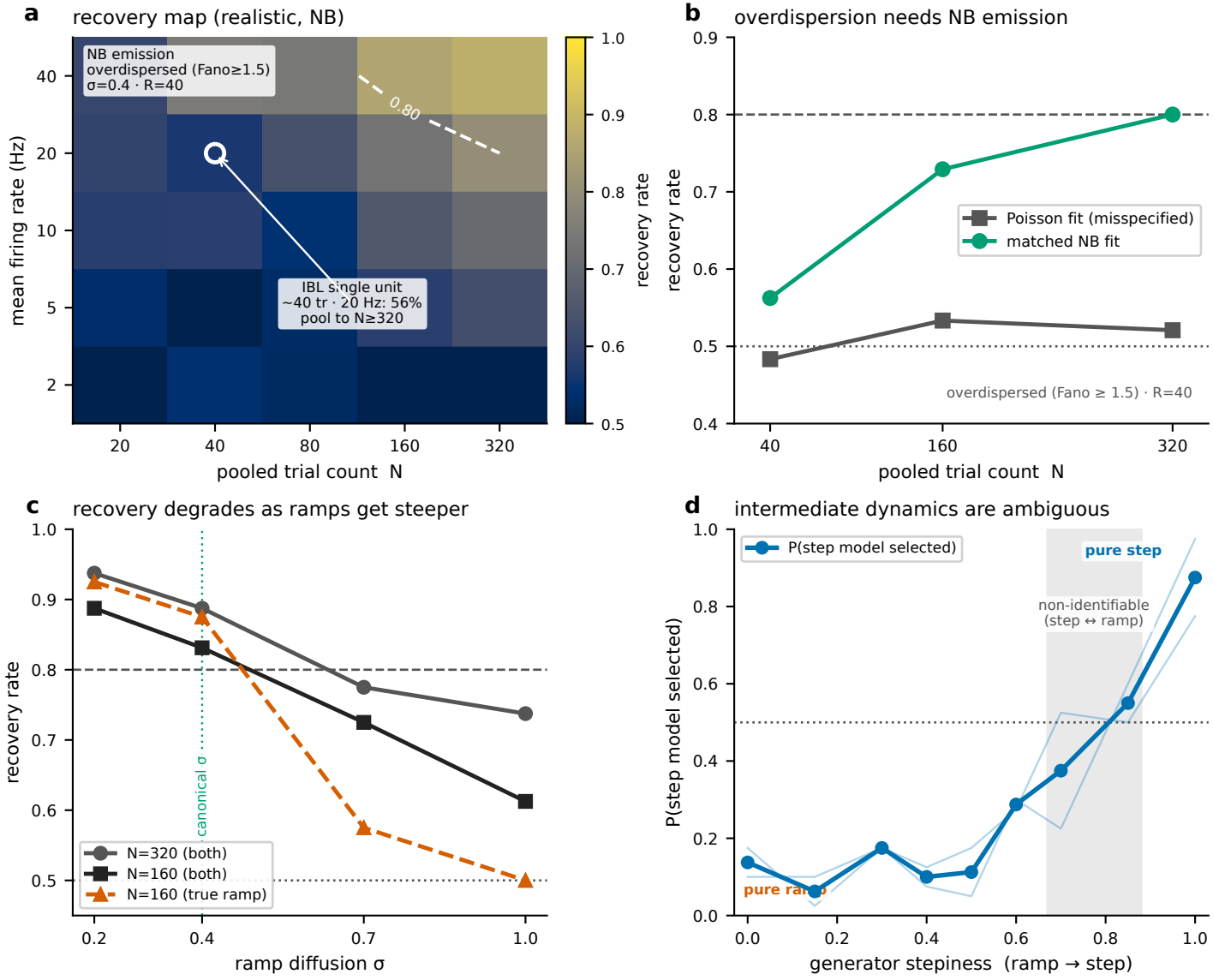


Figure 3 · movement controls distort the decision-signal estimate

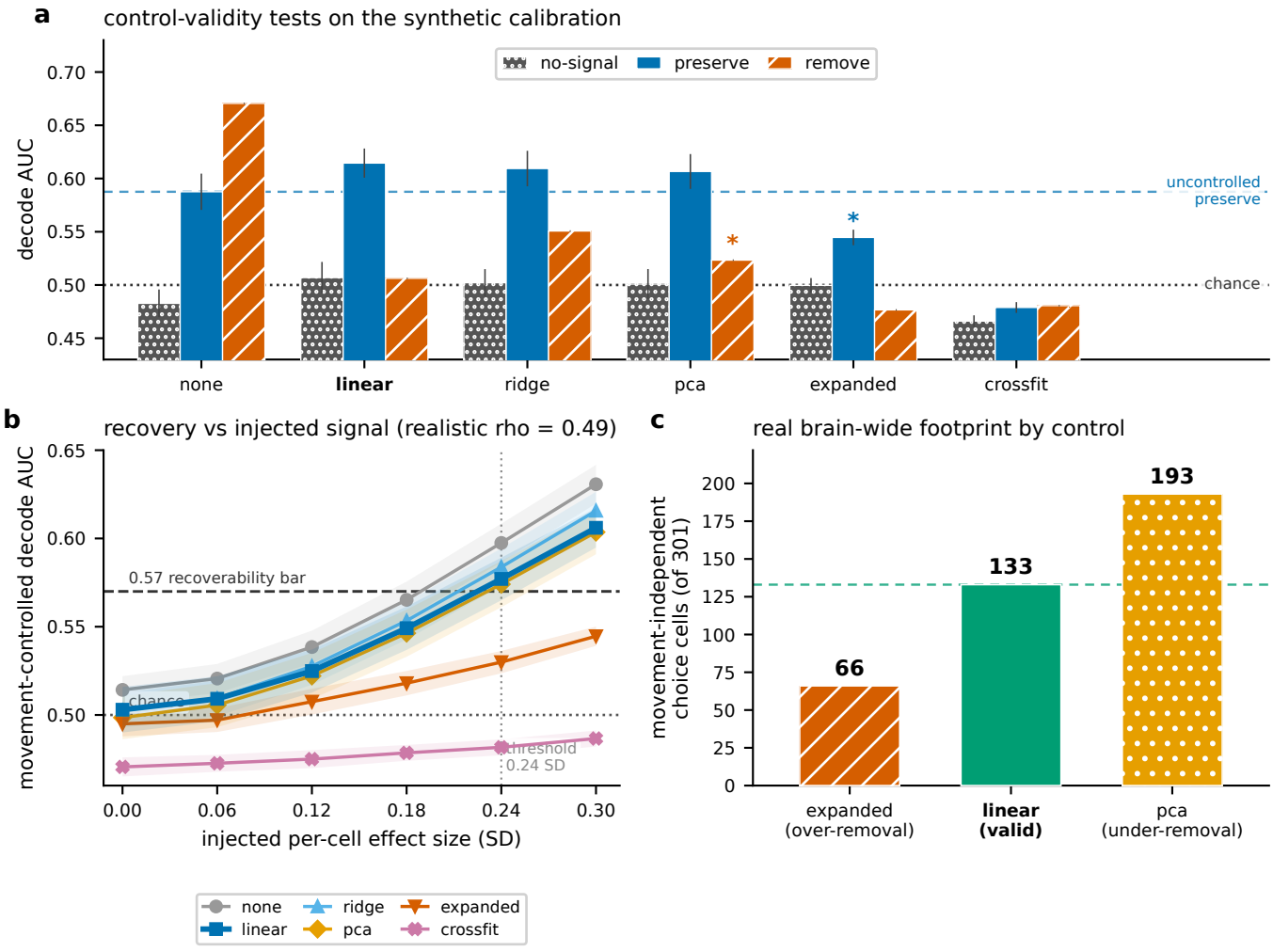


Figure 4 · the decision signal at the boundary, not below it

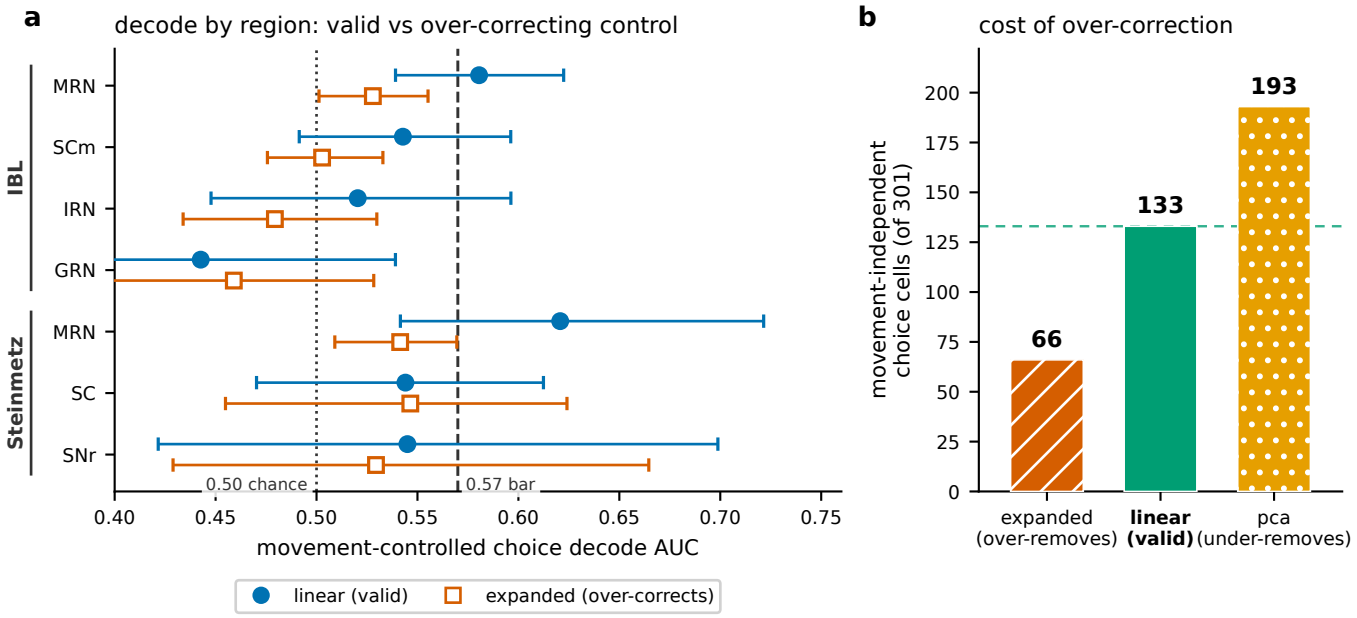


Figure 5 · a small movement-independent choice code

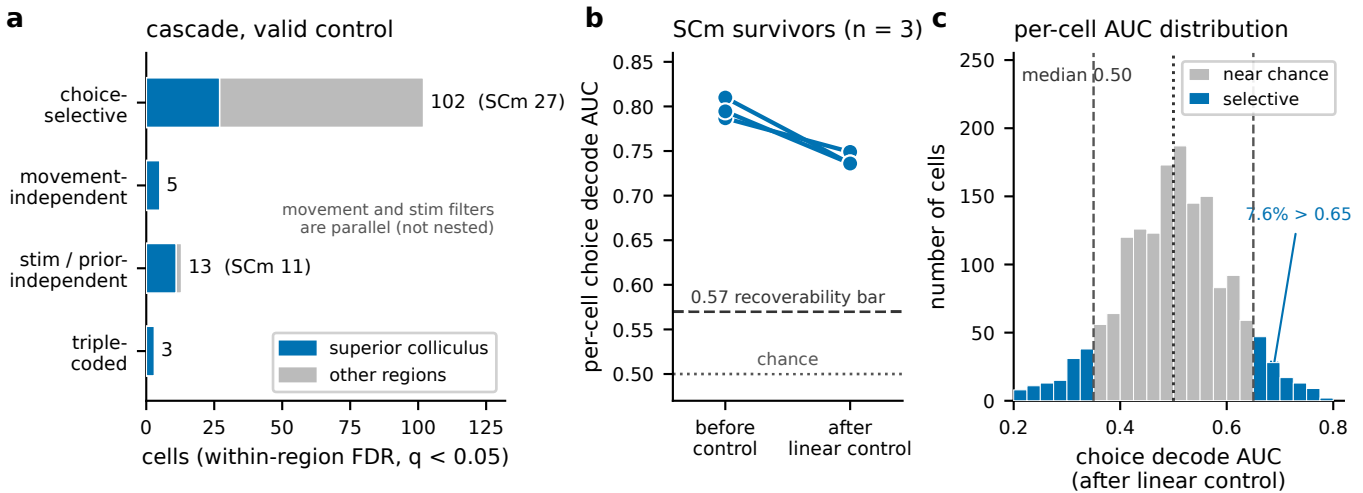
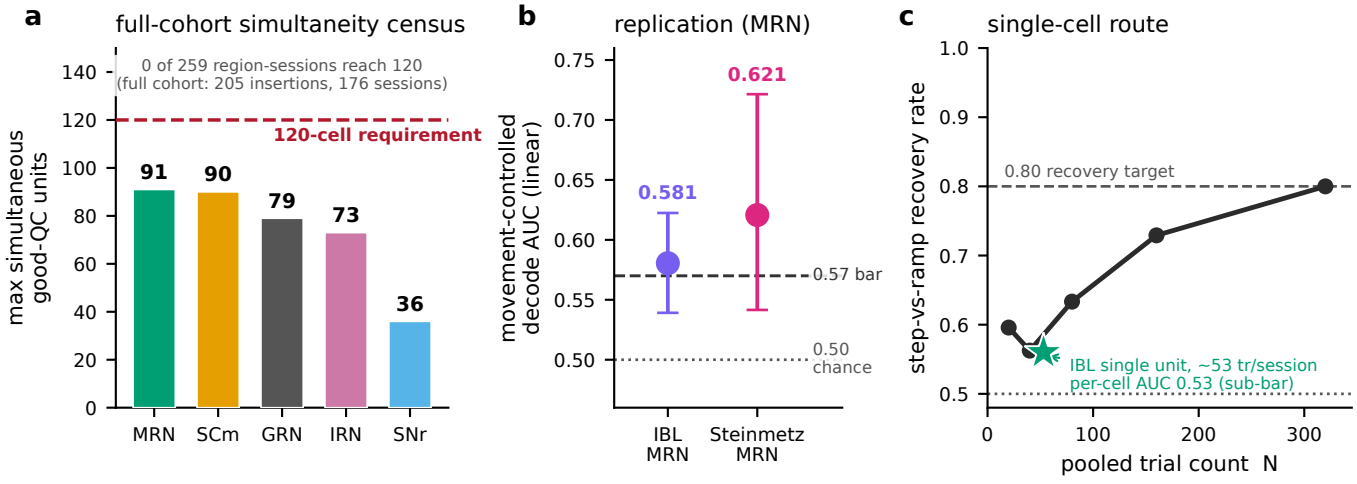


Figure 6 · single-trial dynamics recovery is out of reach at current yields



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- nrreportingsummary.pdf
- supplementaryinformation.pdf