WHAT DOES A CHANNEL NAME IDENTIFY? TASK TRANSFER VERSUS CHANNEL IDENTITY ACROSS NEURAL RECORDING SESSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Successful decoding does not establish correct channel identity, and within-session matching need not remain reliable across recording sessions. We separate channel order, acquisition or registration identifiers, functional similarity, and decoder performance. In LINK, matching disjoint samples within a session recovers 88.4% of individual channels after an imposed permutation, but only 1.8% of complete 96-channel maps. These imperfect maps incur a mean R^2 loss of 0.0018 for one fixed decoder. Across chronological targets from two other macaques, behavior-conditioned recovery of electrode-channel identities falls from 44.8% in paired within-session controls to 4.7% across sessions. In an exploratory oracle analysis in one macaque, label-fitted reindexing raises mean held-out R^2 from 0.06 to 0.38. Across the three macaques, maps fitted on disjoint calibration halves agree on only 5–14% of channels. In 24 mice from the Allen Brain Observatory, repeated-stimulus profiles recover 20.7% of registered cells against a 1.2% random reference. A separate, prospectively frozen four-animal extension tests a locally fitted matcher across sessions. Mean false-match risk against stitched-sorting proxies is 25.66% at 62.55% source coverage, failing the prespecified requirement of at most 10% risk. Together, these experiments show that small decoder losses can coexist with incorrect channel assignments, while local matching competence does not ensure reliable cross-session matching. Identifier recovery, decoder utility, and cross-session matching risk require separate evaluation; acquisition and registered-cell identifiers do not establish biological continuity.

Table 1: A result requires an authority appropriate to its claim. An authority establishes a reference, not necessarily biological truth.

Claim	Reference or authority	What passing the test establishes
Transport	Known storage action	Values and their declared metadata are routed consistently.
Binding	Acquisition or registration record	Assignments agree with that record's objects.
Function	Current input–output relation	A specified conditional law or decoder remains appropriate.
Task	Held-out endpoint labels	A fixed model attains the reported performance on that endpoint.

1 TASK SUCCESS IS NOT CHANNEL IDENTIFICATION

A learned model can retain its predictions while assigning a measurement to the wrong channel. Conversely, a recording can retain every channel identifier while its decoder fails. These are different failures, but both are often discussed using the word "robustness." The ambiguity matters whenever a representation binds a tensor position to a measured source: an electrode, a sensor, an assay feature, or a model-specific unit embedding. We study the distinction in neural recordings, where channel identity and prediction have separate scientific uses.

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.

Existing neural models already address natural shift. SPINT evaluates held-out sessions using unlabeled calibration; NoMAD aligns new neural activity without target behavior labels; NDT2 compares several adaptation budgets (Le et al., 2025; Karpowicz et al., 2025; Ye et al., 2023). Their task results do not, by themselves, establish that the same measurement source has been identified. We ask which evidence permits moving from one claim to the other.

We separate three interface contracts and a task endpoint (Table 1). This is a measurement distinction, not a new invariance theorem (Jacobs & Wallach, 2021). Its empirical content is a set of comparisons that could have agreed but do not: individual versus complete-map recovery, map correctness versus decoder loss, local versus chronological matching, and within-session calibration versus cross-session risk. New exploratory extensions test stronger matchers, finite-sample effects, and the consequences of fitting a permutation directly for prediction. We preserve the original failed risk gate and negative sensitivity results.

2 Contracts and matching limits

Let X contain measurements, M their metadata, and f a fixed pipeline. Transport invariance under a storage action P means $f(XP, MP) = f(X, M)$, with the action on metadata defined consistently. Binding is relative to an authority $\mathcal{A}$: a map b preserves binding between sessions s, t if $\mathcal{A}_s(M_{si}) = \mathcal{A}_t(M_{t,b(i)})$. The authority may name an electrode, a registered cell, or a neuron. Those objects are not interchangeable. Function concerns a declared conditional law or Bayes decoder under a specified loss. Channelwise conditional stability need not imply stability of the joint decoder. Task performance concerns the risk of one fixed decoder on one endpoint.

Co-permuting input values and the corresponding weights preserves a linear predictor exactly. Exchangeability can also make a retrained pipeline's score distribution unchanged under a fixed permutation. Neither fact establishes invariance of a realized index-bound decoder when only its input is moved. Nor does either establish stable biological sources (Gallego et al., 2020; Rule et al., 2020). We use this algebra as a scope control, not as the paper's main contribution.

Why individual recovery does not determine a complete map. Write squared profile-matching cost as $C(\pi) = \sum_i \|x_i - z_{\pi(i)}\|^2$, with authoritative identity relabeled as the identity map. Swapping just i, j changes its cost by

$$C((ij)) - C(\mathrm{id}) = 2\langle x_i - x_j, z_i - z_j \rangle. \tag{1}$$

A negative value rules out exact recovery by minimum-cost assignment. Nonnegative values for every pair do not guarantee recovery: a longer cycle can still improve the assignment. The complete-map event depends on joint error structure, not on multiplying a marginal channel accuracy as if errors were independent. We therefore retain assignment cycles and count exact maps directly.

A model can explain a ceiling without establishing one empirically. For profiles $Y_i^{s,a,m} = \mu_i + \delta_i^s + \epsilon_i^{s,a,m}$, assume independent Gaussian channel/session offsets with variance $\tau^2 > 0$, and independent noise with variance σ^2/m. Within-session halves share δ_i^s; their assignment error tends to zero as m increases. Across independent sessions, the probability of the negative swap in Eq. 1 remains positive. This gives a matcher-specific ceiling under those assumptions (Appendix A). Shared translations or sufficiently small bounded offsets are counterexamples to an unconditional ceiling. We do not infer a biological asymptote from four or five sample sizes. Classical feature-matching theory addresses related separation requirements under common latent means (Collier & Dalalyan, 2016).

3 Evidence, information budgets, and analysis status

Frozen evidence. LINK records finger kinematics and intracortical activity (Temmar et al., 2025b;a). Features are $\log(1 + x)$ of mean spiking-band power over five 20 ms samples, giving 100 ms bins; the endpoints are two finger velocities. Six early sessions train the original decoder and four select hyperparameters. The original identity assay estimates eight behavior-conditioned profiles from disjoint, behavior-balanced trial samples in each of 20 later sessions. The fixed final 150 trials evaluate decoder effects. The 500 resulting maps are nested within sessions; they are not 500 independent

Table 2: Data and reference identities. Repeated sessions are not independent animal replicates. New analyses reuse these datasets and are exploratory.

Resource	Scope used here	Identity reference and principal boundary
LINK	One macaque; original 20 evaluation sessions	Known within-session permutation; declared electrode identity across sessions.
O'Doherty	Two macaques; 45 eligible sessions, 43 later targets	Published M1 channel identifiers; not neuron continuity.
Allen Visual Coding	24 mice; one container per mouse	Registered cell specimen IDs; reference itself is estimated.
UnitMatch examples	One development mouse and four frozen test animals	Shared stitched-sorting cluster IDs; sorting proxies.

biological observations. Relabelings of one fixed cost matrix are not new identification experiments (Lazic, 2010).

The O'Doherty archive contains 47 candidate files (O'Doherty et al., 2020). A frozen occupancy rule retains 45 sessions, including the earliest reference for each macaque, and excludes two. Neural event times are pooled within declared M1 channels, binned at 100 ms, transformed with $\log(1 + x)$, and paired with velocity at a 100 ms lag. Original matching uses first/last 1,500-bin controls and the first 1,500 bins of each reference/target pair. Published identifiers are withheld from all learned costs. The behavioral profile uses eight direction means; the neural-only profile uses seven marginal summaries.

Allen Visual Coding supplies ten repeats of natural movie one in both sessions of each selected container (de Vries et al., 2020; Allen Institute for Brain Science, 2018; n.d.). Responses are normalized to 900 time points per repeat. Repeated-stimulus profiles average presentations; dynamics profiles use moments, quantiles, autocorrelation and spectral bands. All target-session candidates remain present, including cells without a counterpart. Registered IDs enter scoring, not profiles or assignment. The mouse is the unit for bootstrap intervals. The original cohort and amendments are retained, including the decision to keep mice with fewer than 30 shared cells.

Exploratory extensions. Every new experiment has a dated local specification and code/data hashes written before its new outcomes. These are author-maintained records, not external registrations; the datasets and previous outcomes were already known. Matcher settings and decoder history are selected only on the four designated LINK tuning sessions, or the excluded Allen development mouse. New LINK chronological comparisons use 19 matched-style targets: center-out source 20200127 and random-target source 20200313. The latter was previously an evaluation session and is explicitly reused as a reference. This is not an independent replication or the complete 312-session LINK archive.

Graph matching aligns channel correlation matrices with FAQ (Vogelstein et al., 2015). Barycenter, marginal/behavior profile assignment, and 20 randomized starts compete by graph objective, never by evaluated identity. Tuning selects raw or behavior-residual correlation; removing the first principal component is a prespecified sensitivity. Behavior-residual graphs use behavior labels, although not identity labels. Fused Gromov–Wasserstein (FGW) combines a profile cost with relational geometry (Vayer et al., 2019), with its mixing weight selected on tuning units. Rectangular candidate sets are preserved. Graph solutions provide hard maps, not top-five rankings. Solver failures, occupancy exclusions and ties remain in the ledgers.

4 Local identification and chronological matching separate

Strong local identification is not reliable complete binding. The original LINK assay recovers 88.4% of individual channels at top one and 99.3% at top five. Only 1.8% of the 500 complete maps are exact, with exact maps appearing in three sessions. A map contains 11.1 errors on average. Despite those errors, the mean fixed-decoder R^2 loss is only 0.0018. Session-level Hamming error has Spearman correlation 0.736 with mean absolute prediction discrepancy; decoder-weighted error has correlation 0.973. The latter trace incorporates the decoder and observed feature covariance by construction. The result demonstrates decoder-specific insensitivity to some assignment errors, not that identity is unimportant for every task.

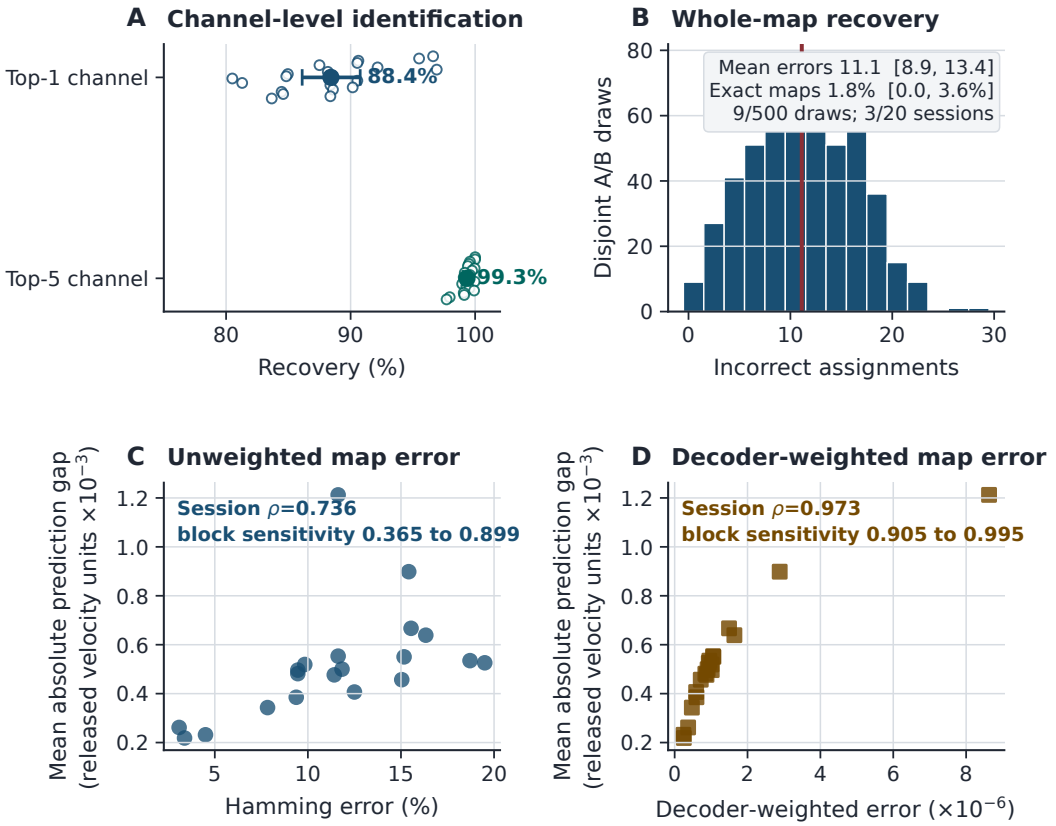

Figure 1: **Identity and task are different estimands.** Original LINK evidence. Individual identification is high, complete maps are rare, and the fixed-decoder consequence is small. Maps are nested within 20 sessions. The decoder-weighted trace is an algebraic re-expression of squared prediction discrepancy, not an independent validation target.

Stronger matchers do not remove the observed natural-session gap. In the frozen O'Doherty analysis, behavior-conditioned top-one recovery falls from 44.8% in paired local controls to 4.7% across 43 later targets; the neural-only values are 20.8% and 3.1%. The new randomized-storage ladder gives approximately 5.1% cross-session recovery for its behavior baseline and 3.3% for marginal summaries. This is a distinct analysis: tie resolution can change a selected optimum under storage relabeling, without changing its cost. We retain both records rather than replacing the frozen estimates.

The stronger cross-session values are 2.3% for behavior-residual graph matching, 2.5% after first-PC removal, and 5.2% for FGW. Across all 45 local controls, the respective means are 16.6%, 49.5%, and 51.3%. The first-PC and FGW controls are therefore informative: these procedures can recover many channels locally but recover few under chronological change. The residual-graph control is weaker, so its poor transfer cannot alone locate the failure in drift. All values describe two macaques, not an animal population.

A post-outcome activity check retains every candidate and fitted map, then scores channels with nonconstant source activity and a nonconstant authoritative target counterpart. This stratum covers 91.0% of cross-session channels. Behavior and FGW accuracy remain 4.7% and 5.3%, respectively; silent channels therefore do not explain most failures of these tested matchers.

The new LINK chronological sample-size comparison supports the same finite-range distinction. For marginal neural summaries, increasing each profile from 15 to 75 trials raises local assignment accuracy from 62.6% to 84.4%; matched-style cross-session accuracy changes from 5.5% to 5.8%. The original eight-cluster fingerprint, restored in an explicitly dated additional analysis, gives 86.3% local and 5.9% chronological accuracy at 75 trials. These chronological pools differ from the original

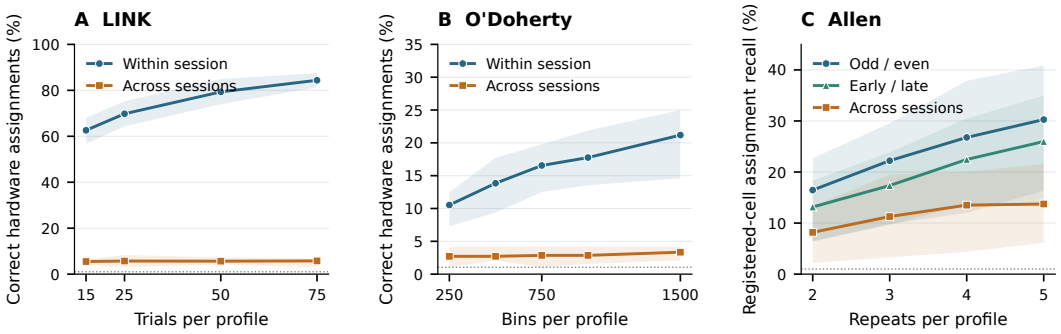

Figure 2: **Finite-sample improvement does not erase the measured chronological gap.** LINK uses 20 local and 19 cross-session comparisons in one macaque; O'Doherty uses 45 local and 43 cross-session comparisons in two macaques. Both panels use marginal neural profiles. Allen uses movie profiles, with directions or local session controls averaged within each of 24 mice. Lines connect measured means; shading is the unitwise interquartile range, not a confidence interval. Gray dotted lines give random-assignment references. The panels have different vertical scales and denominators: hardware accuracy in A/B, registered-common-cell recall in C. Candidate sets are never restricted to known matches. These are fixed-subset curves, not fitted asymptotes.

behavior-balanced identity assay, so 86.3% is not a revised estimate of 88.4%. More observations help locally over the measured range; these curves do not identify an infinite-data ceiling.

The O'Doherty neural profiles improve from 10.5% to 21.2% locally as the budget increases from 250 to 1,500 bins; chronological accuracy changes from 2.7% to 3.3% (Fig. 2). The target-paired gap grows from 7.9 to 17.8 percentage points. Behavior profiles fail the frozen occupancy rule at both smallest budgets, so we do not join averages over their changing eligible subsets into a comparable trajectory.

A positive boundary in registered visual cells. The original Allen repeated-stimulus profile recovers 20.7% of registered cells, with mouse bootstrap interval [16.6,24.9]%, against a 1.2% random reference. Fixed fusion reaches 22.5% in secondary retrieval. Retaining the top 20% of fused assignment edges by criticality raises precision from 15.5% to 34.6%, while retaining 10.3% of registered cells. Simpler post-outcome cost and margin rankings are not clearly worse. Criticality is optimization sensitivity, not calibrated identity probability. Known repeated input supplies structure without behavioral or identity labels; it does not establish complete matching or biological continuity.

The new Allen five-repeat ladder measures assignment recall, not the original ten-repeat row-ranked retrieval. Movie profiles recover 30.3% of registered cells in odd/even controls, 26.0% in early/late controls, and 13.7% across sessions. FGW reaches 14.4% across sessions; raw-correlation and first-PC-removed graphs reach 2.2% and 3.3%. No exact registered-common map is recovered. The two-to-five-repeat curves still improve locally, so the available range does not identify an estimation-noise floor or an irreducible drift component.

5 Can reindexing explain natural decoder loss?

A recovered identity map and a prediction-optimal map need not coincide. We test that distinction with an explicitly oracle analysis. A fixed five-bin ridge decoder is trained on source data. Calibration labels fit four transformations, compared with static hardware routing (C0): per-channel affine correction preserving binding (C1), a shared permutation at every lag (C2), a full linear input correction (C3), and a newly fitted target ridge decoder (C4). C1 and C3 penalize deviations from identity and select regularization by blocked calibration-only cross-validation. Every class is scored on a separate target half. Source calibration/test partitions provide an overfitting control using the same trained decoder. Calibration occupies a source quarter versus a target half; absolute sample counts depend on session length and are not matched.

Table 3: **Prediction-optimal routing is not reference-identity recovery.** Mean held-out R^2 after labeled calibration; each row describes one animal. C0: static hardware route; C1: channel affine; C2: selected shared-lag permutation; C3: full linear input map; C4: target ridge. No score is clipped. Penalized fits need not be ordered by expressivity.

Animal	Targets	C0	C1	C2	C3	C4
LINK N	19	-2.817	0.050	0.097	0.189	0.221
Indy	35	0.059	0.229	0.376	0.364	0.488
Loco	8	-1.774	-0.203	-0.147	-0.326	-0.008

For C2, a convex relaxation optimizes a doubly stochastic map with Frank–Wolfe (Jaggi, 2013). The iterate's loss f and gap g give $\max(0, f - g) \leq \min_P f(P)$ over permutations. Here f is $1-$uniform-average calibration R^2. The bound concerns calibration only, not unseen trials. We round the relaxation and run pair-swap descent from identity, the tuning-selected label-free map, and 20 random starts. The selected hard map minimizes observed calibration loss among those candidates. It is not a proven global optimum; caps and remaining gaps are reported.

We report held-out R^2 for every class, hardware disagreement, decoder-weighted disagreement, and stability between fits on separate calibration halves. The recovered-gap fraction $(R^2_k - R^2_{C0})/(R^2_{C4} - R^2_{C0})$ is defined only where the denominator is at least 0.05; all raw outcomes remain visible. Calibration ceilings use calibration denominators separately. A large fraction in a negative-R^2 regime is not successful decoding. A large hardware disagreement shared by the within-session control is not evidence of uniquely cross-session identity loss.

Prediction improves, but the fitted binding is unstable. In Indy, reindexing raises mean held-out R^2 from 0.059 to 0.376, exceeding binding-preserving affine correction (0.229) but falling short of target recalibration (0.488). C2 also exceeds C1 on average in LINK and Loco. However, LINK's mean C2 score is only 0.097, and Loco's is negative (-0.147; median 0.111). Mean recovered fractions are 0.909, 0.696 and 0.894 in LINK, Indy and Loco, respectively; the gap gate retains all 19 LINK targets, 34 of 35 Indy targets, and all eight Loco targets. Large fractions partly reflect poor static transfer, not near-ceiling absolute prediction.

The hard maps disagree with hardware on 94.6–96.5% of channels, but the four source controls already disagree on 58.3–94.8%. Half-calibration map agreement is only 4.8–14.1% across the three animals. The unequal calibration budgets further limit the control comparison. Neither prespecified headline criterion is met; all animals take the mixed, descriptive branch. The evidence supports prediction-improving transformations, not stable alternative biological identities or a conclusion that natural shift is a permutation. Frank–Wolfe mean gaps are 0.3215, 0.0017 and 0.0206; none reaches the 10^{-6} tolerance within 500 iterations. The checked numerical bounds therefore retain their calibration-only scope and, especially in LINK, substantial slack.

6 Calibration transfer can fail despite competent local controls

A frozen four-animal risk failure. Using the public UnitMatch examples (van Beest et al., 2025), we fit a selective scorer from within-recording split-half positives and high-scoring off-diagonal negatives. It sees waveform-derived similarities, the UnitMatch probability/total score, and summary features. It never sees cross-session reference IDs. Mutual-best predictions are retained; unmatched units abstain. On the single development mouse, the chosen rule attains 8.33% proxy false-match risk at 72% source coverage. The rule, thresholds and comparison pipeline were then frozen.

Across four test animals, equal-animal mean proxy risk is **25.66%**, above the prespecified 10% ceiling, at 62.55% coverage. The fixed simple baseline gives 27.02% risk at 57.13% coverage. More coverage is not successful risk control. One animal reaches 60.58% proposed risk. The reference consists of shared stitched-sorting cluster IDs and can contain errors; this experiment tests agreement with that proxy, not the original UnitMatch paper's biological validation or all waveform matchers. We make no probability-calibration or certified partial-map claim from this failure.

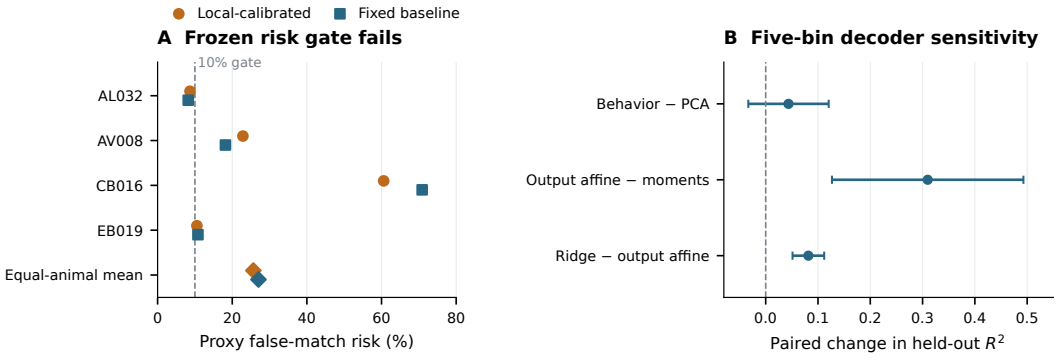

Figure 3: **Two checks constrain claims about repair.** A: the frozen four-animal test exceeds its 10% equal-animal mean proxy-risk gate. Individual animals and the equal-weight mean are shown; the baseline is the fixed pipeline in this experiment, not all of UnitMatch's validation. B: the five-bin decoder rerun retains two positive repair contrasts but not the behavior-over-PCA contrast. Bars are simultaneous 95% intervals for the three declared contrasts, using circular blocks of four among 20 LINK sessions. Animals, sessions and calibration settings differ between panels.

Decoder competence narrows an apparent repair advantage. The original LINK repair study compares moment normalization, CORAL, PCA alignment, behavior-conditioned alignment, output affine calibration, and target ridge (Sun et al., 2016; Jude et al., 2022). The last three use current behavior or output labels. In the new competence check, five bins of causal history are selected on four tuning sessions only. Mean within-session R^2 over the original 20 evaluation sessions rises from 0.170 to 0.219 relative to a one-bin ridge. A Wiener filter with identical linear lag features is the same model family, not an independent baseline.

We then rerun the three original repair contrasts using the stronger decoder and the original 75-trial calibration/final-150-trial evaluation pools. Historical latent dimensions are held fixed rather than retuned on evaluation sessions. The behavior-minus-PCA gain is 0.044 with simultaneous 95% interval $[-0.033, 0.120]$: it is no longer clear. Its absolute behavior-alignment mean is 0.002, with a marginal interval spanning zero. The affine-minus-moments gain is 0.310 $[0.127, 0.493]$, and ridge-minus-affine is 0.082 $[0.051, 0.112]$. Circular session-block resampling preserves the original familywise construction (Politis & Romano, 1992; Romano & Wolf, 2005). These sensitivity results narrow the repair claim; they do not erase the original, separately frozen experiment. The later 40-session cohort was not rerun with the new history model.

7 Relation to neural representation learning

The audit concerns what an evaluation establishes, not which published decoder wins. FALCON supplies standardized natural-shift decoding tasks (Karpowicz et al., 2024); we do not report a new FALCON benchmark. SPINT infers activity-dependent unit embeddings from unlabeled calibration and studies natural held-out sessions (Le et al., 2025). CRRL learns reference-day rearrangement and reconstruction, evaluates natural day shift, and distinguishes fixed future-day inference from supervised few-shot variants (Tan et al., 2025). NoMAD aligns latent dynamics without ordinary target behavior labels (Karpowicz et al., 2025). These are compatible with our distinction between task transfer and reference-identity recovery.

POYO fits new unit/session embeddings with labeled target data and already includes a within-session embedding re-identification probe (Azabou et al., 2023). A genuinely cross-session hardware probe would ask a different question; it is not a result of this paper. NDT2's context tokens, BrainBERT's shared single-electrode encoder, and LaBraM's declared channel embeddings represent different binding choices (Ye et al., 2023; Wang et al., 2023; Jiang et al., 2024). Appendix B records those choices against primary-source locations. The table is purposive and interpretive, not a systematic prevalence audit, a decoder leaderboard, or a criticism that the cited works tested only artificial permutations.

8 LIMITATIONS AND CONCLUSION

The observed units are three macaques, 24 selected Allen mice, and four animals in a separate sorting-proxy test. Sessions, pairs, maps and channels do not enlarge those animal counts. Hardware identifiers do not establish neuron continuity; Allen IDs and stitched-sorting IDs are estimated references. Task style and time remain partly confounded in historical LINK results. Matching failures depend on the tested features, solvers, calibration budgets and reference authority; they do not prove that identity is unrecoverable. The theory's Gaussian ceiling is conditional and has not been fitted and validated as a biological model. Our new analyses are exploratory extensions of known datasets. No independent electron-microscopy identity assay, large decoder zoo, deployment guarantee, or state-of-the-art decoder claim is made.

The practical conclusion is narrower and testable. Specify what names a channel, which action changes it, and which information a repair sees. Measure individual identification, complete binding, task performance and selective risk separately. Preserve a positive competence control and report failures on the same unit of analysis as successes. A correct interface contract prevents useful task transfer from being mistaken for source identification, and prevents reliable local matching from being mistaken for validated calibration transfer.

REPRODUCIBILITY STATEMENT

The supplement preserves protocols, amendments, input hashes, append-only new ledgers, failed gates, solver tests, independent metric validators and figure code. Raw neural and behavioral recordings are not redistributed. Provider URLs and archived acquisition records support reacquisition. Local freezing is distinguished from external registration, and numerical solver bounds from formal floating-point certificates. Details and experiment coverage are in the appendix.

ETHICS STATEMENT

This secondary analysis introduces no animal intervention or clinical decision. Public identifiers are used only within their documented scope. No biological-continuity or deployment-safety claim is made. Dataset terms and attribution requirements are recorded in the supplement.

AI USE STATEMENT

AI-assisted tools supported literature retrieval, code drafting, experimental implementation, mathematical derivations and proof checks, numerical checks, figure construction and manuscript editing. Their proposed interpretations were subject to source and ledger checks. The author remains responsible for the claims, provenance and submission; automated tests are not independent scientific replication.

REFERENCES

Allen Institute for Brain Science. Allen brain observatory: Visual coding overview. Technical Report Version 4, Allen Institute for Brain Science, October 2018. URL https://brain-map.org/support/documentation/data-visual-coding.

Allen Institute for Brain Science. Brain observatory. AllenSDK documentation, version 1.3.2, n.d. URL https://allensdk.readthedocs.io/en/v1.3.2/brain_observatory.html. Accessed August 31, 2026.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44937–44956, 2023.

Olivier Collier and Arnak S. Dalalyan. Minimax rates in permutation estimation for feature matching. *Journal of Machine Learning Research*, 17(6):1–31, 2016. URL https://www.jmlr.org/papers/v17/collier16a.html.

Saskia E. J. de Vries, Jerome A. Lecoq, Michael A. Buice, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23: 138–151, 2020. doi: 10.1038/s41593-019-0550-9.

Juan A. Gallego, Matthew G. Perich, Raeed H. Chowdhury, Sara A. Solla, and Lee E. Miller. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, 23:260–270, 2020. doi: 10.1038/s41593-019-0555-4.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Justin Jude, Matthew Perich, Lee Miller, and Matthias Hennig. Robust alignment of cross-session recordings of neural population activity by behaviour via unsupervised domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10462–10475, 2022.

Brianna M. Karpowicz, Joel Ye, Chaofei Fan, Pablo Tostado-Marcos, Fabio Rizzoglio, Clay Washington, Thiago Scodeler, Diogo de Lucena, Samuel R. Nason-Tomaszewski, Matthew J. Mender, Xuan Ma, Ezequiel Matias Arneodo, Leigh R. Hochberg, Cynthia A. Chestek, Jaimie M. Henderson, Timothy Q. Gentner, Vikash Gilja, Lee E. Miller, Adam G. Rouse, Robert A. Gaunt, Jennifer L. Collinger, and Chethan Pandarinath. Few-shot algorithms for consistent neural decoding (FALCON) benchmark. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-2439.

Brianna M. Karpowicz, Yahia H. Ali, Lahiru N. Wimalasena, Andrew R. Sedler, Mohammad Reza Keshtkaran, Kevin Bodkin, Xuan Ma, Daniel B. Rubin, Ziv M. Williams, Sydney S. Cash, Leigh R. Hochberg, Lee E. Miller, and Chethan Pandarinath. Stabilizing brain-computer interfaces through alignment of latent dynamics. *Nature Communications*, 16:4662, 2025. doi: 10.1038/s41467-025-59652-y.

Stanley E. Lazic. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neuroscience*, 11:5, 2010. doi: 10.1186/1471-2202-11-5.

Trung Le, Hao Fang, Jingyuan Li, Tung Nguyen, Lu Mi, Amy Orsborn, Uygar Sümbül, and Eli Shlizerman. SPINT: Spatial permutation-invariant neural transformer for consistent intracortical motor decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0870.

Joseph E. O'Doherty, Mariana M. B. Cardoso, Joseph G. Makin, and Philip N. Sabes. Nonhuman primate reaching with multichannel sensorimotor cortex electrophysiology, May 2020. URL https://doi.org/10.5281/zenodo.3854034.

Dimitris N. Politis and Joseph P. Romano. A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation. *The Annals of Statistics*, 20(4):1985–2007, 1992. doi: 10.1214/aos/1176348899.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005. doi: 10.1111/j.1468-0262.2005.00615.x.

Michael E. Rule, Adrianna R. Loback, Dhruva V. Raman, Laura N. Driscoll, Christopher D. Harvey, and Timothy O'Leary. Stable task information from an unstable neural population. *eLife*, 9:e51121, 2020. doi: 10.7554/eLife.51121.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. doi: 10.1609/aaai.v30i1.10306.

Xianhan Tan, Binli Luo, Yu Qi, and Yueming Wang. CRRL: Learning channel-invariant neural representations for high-performance cross-day decoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 95961–95993. Curran Associates, Inc., 2025. doi: 10.52202/085713-2889.

Hisham Temmar, Nina Gill, Yixuan Wang, Nicholas B. Mellon, Chang Liu, Luis Hernan Cubillos, Rio I. Parsons, Joseph T. Costello, Matteo Ceradini, Madison M. Kelberman, Matthew J. Mender, Aren I. Hite, Dylan M. Wallace, Samuel R. Nason-Tomaszewski, Matthew S. Willsey, Parag G. Patil, Anne W. Draelos, and Cynthia A. Chestek. LINK: Long-term intracortical neural activity and kinematics. DANDI Archive, Dandiset 001201, version 0.251023.2336, 2025a. URL https://doi.org/10.48324/dandi.001201/0.251023.2336.

Hisham Temmar, Yixuan Wang, Nina Gill, Nicholas Mellon, Chang Liu, Luis Cubillos, Rio Parsons, Joseph Costello, Matteo Ceradini, Madison Kelberman, Matthew J. Mender, Aren Hite, Dylan M. Wallace, Samuel Nason-Tomaszewski, Parag Patil, Matt Willsey, Anne Draelos, and Cynthia A. Chestek. Long-term intracortical neural activity and kinematics (LINK): An intracortical neural dataset for chronic brain-machine interfaces, neuroscience, and machine learning. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 38, 2025b. doi: 10.52202/085713-4642.

Enny H. van Beest, Célian Bimbard, Julie M. J. Fabre, Sam W. Dodgson, Flóra Takács, Philip Coen, Anna Lebedeva, Kenneth D. Harris, and Matteo Carandini. Tracking neurons across days with high-density probes. *Nature Methods*, 22:778–787, 2025. doi: 10.1038/s41592-024-02440-1.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284, 2019. URL https://proceedings.mlr.press/v97/titouan19a.html.

Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Louis J. Podrazik, Steven G. Kratzer, Eric T. Harley, Donniell E. Fishkind, R. Jacob Vogelstein, and Carey E. Priebe. Fast approximate quadratic programming for graph matching. *PLOS ONE*, 10(4):e0121002, 2015. doi: 10.1371/journal.pone.0121002.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xmcYx_reUn6.

Joel Ye, Jennifer Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: Multi-context pretraining for neural spiking activity. In *Advances in Neural Information Processing Systems*, volume 36, pp. 80352–80374, 2023.

A THEORY AND NUMERICAL CERTIFICATES

A.1 CONTRACTS AND THE LIMITS OF PROFILE MATCHING

Transport concerns the pipeline: a common reindexing of values and metadata must preserve its endpoint output. Binding concerns an authority: a map b preserves binding when $\mathcal{A}_s(M_{si}) = \mathcal{A}_t(M_{t,b(i)})$. The authority may identify an electrode, a registration label, or a neuron; these are not interchangeable. Function concerns a specified conditional law, or a Bayes decoder under a specified loss. Channelwise conditional stability does not imply stability of the joint conditional law. Task performance is the risk of a fixed decoder on a declared endpoint and population. None of these contracts, by itself, establishes the others.

For analysis, relabel the authoritative map as the identity and write $C(\pi) = \sum_i \|x_i - z_{\pi(i)}\|^2$. Exchanging only i and j gives

$$C((ij)) - C(\mathrm{id}) = 2\langle x_i - x_j, z_i - z_j\rangle. \tag{2}$$

A negative inner product rules out exact recovery by minimum-cost assignment. Its absence for every pair is necessary, but not sufficient, for the identity to be optimal: an improving cycle of length three or more may remain. Since every permutation decomposes into cycles, identity is uniquely optimal if and only if every nontrivial cycle has positive cost increment. A union bound over all nonpositive-cost cycles, including ties, bounds failure probability without assuming independent channel errors. Strictly improving cycles suffice only when cost ties have probability zero.

A conditional drift ceiling. Consider fixed $n \geq 2$ and $d \geq 1$, with profiles

$$Y_i^{(s,a,m)} = \mu_i + \delta_i^s + \epsilon_i^{(s,a,m)}, \qquad \delta_i^s \sim \mathcal{N}(0, \tau^2 I_d), \quad \epsilon_i^{(s,a,m)} \sim \mathcal{N}(0, \sigma^2 I_d/m). \tag{3}$$

Assume $\tau > 0$, independent offsets across channels and sessions, and independent errors. Within-session halves share the same offset; m measures independent information. Least-squares assignment then recovers the within-session map with probability tending to one as $m \to \infty$. For independent sessions, define

$$q_{ij} = \Pr(\langle U, V\rangle < 0), \qquad U, V \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mu_i - \mu_j, 2\tau^2 I_d). \tag{4}$$

Gaussian full support gives $q_{ij} > 0$, and

$$\lim_{m\to\infty} \Pr(\widehat{\pi} = \mathrm{id}) \leq 1 - q_{ij}, \qquad \lim_{m\to\infty} \mathbb{E}[A_{\mathrm{assign}}] \leq 1 - \frac{2q_{ij}}{n}. \tag{5}$$

Here A_{assign} is the correct fraction of a complete assignment, not independent top-1 retrieval. For the latter the same argument gives the weaker upper bound $1 - q_{ij}/n$.

Proof and scope. Within a session, both halves converge to the same distinct realized profiles, so every wrong assignment has positive limiting cost. Across sessions, the pair differences converge to U, V. Their opposing open neighborhoods have positive probability, making a swap strictly better by Eq. (2). Any wrong permutation moves at least two channels. For independent retrieval, at least one off-diagonal distance must be better than its diagonal distance. Continuous profiles give zero probability of assignment-cost ties, which permits taking limits. These statements concern this matcher and random-session model, not all identification methods. Positive drift variance alone is insufficient: a common translation of all channels leaves least-squares assignment intact, and sufficiently small bounded drift also permits exact recovery. We have not validated Eq. (3) as a quantitative model of the recordings.

Classical Gaussian feature-matching theory establishes separation requirements under common latent means, not an unconditional ceiling for biological drift Collier & Dalalyan (2016). In Eq. (3), equally sized independent profiles satisfy $\mathbb{E}[D_{\mathrm{cross}} - D_{\mathrm{within}}] = 2d\tau^2$. Thus the discrepancy divided by d estimates increment variance, not per-session variance. Estimating this quantity with authoritative pairs is an identity-informed explanatory analysis and must be declared separately from label-free matching.

Model	Channel reference	New-session information	Selected evaluation	Supported scope
NDT2 (2023)	Ordered patches; context tokens	Zero-shot, unlabeled tuning, or labeled tuning	New sessions, subjects and tasks	Task transfer under the stated calibration
POYO (2023)	Learned unit and session embeddings	Labels fit new embeddings; optional full tuning	New sessions and animals	Task transfer with learned new bindings
SPINT (2025)	Activity-conditioned unit embeddings	Unlabeled calibration; no gradient updates	Natural FALCON session shift	Transport symmetry and empirical task transfer
CRRL (2025)	Learned reference-day rearrangement	Fixed future-day inference after labeled early-day training[a]	Synthetic perturbations; natural day shift	Task robustness after rearrangement and reconstruction
BrainBERT (2023)	Shared single-electrode encoder	Labeled classifier; optional encoder tuning	New sessions, subjects and electrode locations	Single-electrode representation transfer
LaBraM (2024)	Declared EEG names select spatial embeddings	Labeled downstream tuning	New datasets and EEG configurations	Task transfer with declared channel binding
NoMAD (2025)	Learned population-to-latent mapping	Unlabeled alignment; initial decoder uses labels	Natural recording drift	Task preservation through latent alignment

Table 4: Channel-reference mechanisms and adaptation settings in selected neural models. The final column is our contract interpretation, not a claim made in those terms by the cited authors. Task transfer does not by itself establish physical identity; this is not a decoder ranking. [a]CRRL also reports separate supervised few-shot variants.

What the permutation certificate certifies. Let $f(D)$ be a fixed decoder's convex squared calibration loss, with the same doubly stochastic matrix D applied at every lag. Its Frank–Wolfe gap is $g(D) = \max_{Q \in \mathcal{B}} \langle \nabla f(D), D - Q \rangle$. Convexity gives Jaggi (2013)

$$\max\{0, f(D) - g(D)\} \leq \min_{Q \in \mathcal{B}} f(Q) \leq \min_{P \in \mathcal{P}} f(P). \tag{6}$$

This bounds achievable calibration performance only. It does not bound test performance, and every gap-recovery fraction must use the same split and metric in its numerator and denominator. For uniform-average multi-output R^2, use $f(D) = p^{-1} \sum_j \mathrm{SSE}_j(D)/\mathrm{SST}_j$ with positive, fixed SST_j, rather than a pooled SSE/SST ratio.

Finally, if the data transformation and learning procedure are equivariant to a common reindexing of training and test inputs, equal performance after retraining is compatible with index-bound predictors. It does not establish invariance of a fixed decoder.

B Binding mechanisms in selected neural models

The evidence audit in the supplement records primary PDF page and section locations for every cell. Contract classifications are our interpretation of the stated experiment, not terminology attributed to the cited authors. No independent human inter-coder agreement statistic was collected. The table does not imply that a model lacks an ability merely because a specific ability was not evaluated in the cited setting.

C Implementation and analysis boundaries

Original identities and uncertainty. LINK's 88.4% result uses disjoint behavior-balanced profile samples; the new finite-sample curves use chronological pools. Original same-session O'Doherty controls reverse the target order, while the new ladder randomizes storage order with a fixed seed. Assignment optima can tie; all associated estimates retain their own provenance. Original Allen

top-one retrieval and one-to-one assignment precision/recall have different denominators. Graph matching has no top-five output. Complete assignment cycles are recorded only for equal, complete identity sets. The post-outcome O'Doherty activity sensitivity defines nonconstant activity by variance greater than 10^{-12} in each profile block. Identity labels define scoring strata only; candidates are not filtered and maps are not refitted. Its within and cross strata cover different sets, so the conditional comparison is descriptive, not a new causal drift estimate.

Map classes. C1 uses one gain and offset per channel shared across lag blocks. C3 uses one full linear correction shared across lag blocks. Corrections are parameterized in source-standardized channel coordinates; the fixed decoder and source normalization remain unchanged. Full-map directions that cannot affect the decoder are removed through an exact minimum-norm factorization. C1/C3 penalty candidates are $0.01, 0.1, 1, 10, 100, 1000$, selected within calibration. The source ridge and C4 use five causal 100 ms bins and ridge penalty 1000, selected on LINK's four tuning sessions without O'Doherty retuning. C4 is explicitly a newly fitted target decoder, so its fitting statistics come from calibration. C2 acts before fixed source normalization. Its stochastic relaxation and hard permutation are different fitted objects and have separate losses. Split-half map agreement is a stability diagnostic, not identity truth.

The half-calibration stability diagnostic changes both the observations and random solver starts. Output rows are disjoint, but adjacent O'Doherty halves can share up to four raw history bins. It measures fit-to-fit stability, not independent-sample identity reliability or a drift-only effect. C2 improves calibration yet reduces held-out R^2 in all four source controls, by 0.015–0.060 relative to C0; the controls therefore reveal ordinary fitting error as well as assignment ambiguity.

Risk definition. For an animal with accepted pairs S, proxy false-match risk is the incorrect fraction in S, source coverage is $|S|/n_s$, and proxy recall is correct accepted pairs divided by the number of shared proxy IDs. The primary endpoint averages animal risks equally, not by pair count. Within-recording controls fit a depth-two histogram gradient-boosted classifier with 100 iterations, learning rate 0.05, L2 penalty 1 and seed 0. The fixed decision accepts mutual maxima without an additional positive score or margin threshold. The four-animal gate additionally requires improved coverage and no higher mean risk than the baseline. It fails because mean risk exceeds 0.10.

Data and software terms. The archived LINK and O'Doherty records list CC BY 4.0. Allen use follows its research and citation terms. UnitMatch example-data provenance and runtime versions are recorded with the new branch. The package excludes raw responses and third-party article PDFs. Figure regeneration from recorded ledgers is not new data collection. All new experiments run on CPU; no GPU benchmark or complete retrospective compute budget is claimed.

D EXTENDED NUMERICAL RESULTS

The locally frozen E1 low-ceiling branch required mean calibration-only $F_{\mathrm{UB}} < 0.25$. The task-reindexing branch required mean eligible test $F_{C2} \geq 0.25$ and mean paired excess hardware disagreement of at least 0.50. Neither branch is met in any animal. The high within-control disagreement itself prevents the excess-disagreement threshold from being reached for these sources. Thresholds were not lowered after observing this outcome. The within controls are four source sessions, not an independent estimate of a population floor. Full per-output correlations and both calibration/test sufficient statistics remain in the supplement.

Table 5: Cross-session E1 detail. Fractions are reported only where the held-out C4–C0 gap is at least 0.05; they are not clipped to $[0, 1]$.

Animal	Class	n	Mean R^2	Median R^2	Eligible	Mean F	Median F
LINK N	C0	19	-2.817	-1.796	19	0.000	0.000
	C1	19	0.050	0.039	19	0.860	0.917
	C2	19	0.097	0.081	19	0.909	0.934
	C3	19	0.189	0.193	19	0.981	0.993
	C4	19	0.221	0.205	19	1.000	1.000
Indy	C0	35	0.059	0.101	34	0.000	0.000
	C1	35	0.229	0.225	34	0.349	0.303
	C2	35	0.376	0.381	34	0.696	0.677
	C3	35	0.364	0.383	34	0.683	0.708
	C4	35	0.488	0.506	34	1.000	1.000
Loco	C0	8	-1.774	-0.585	8	0.000	0.000
	C1	8	-0.203	0.028	8	0.755	0.753
	C2	8	-0.147	0.111	8	0.894	0.897
	C3	8	-0.326	0.061	8	0.612	0.875
	C4	8	-0.008	0.218	8	1.000	1.000

Table 6: C2 map and optimization diagnostics. H_w is decoder-weighted disagreement; H_{within} is paired by source with unequal calibration budgets. Stability is agreement between two half-calibration maps. FW gaps are in one-minus-uniform-R^2 calibration units.

Animal	H	Median H	H_w	H_{within}	Stability	Mean gap	Max gap
LINK N	0.9457	0.9583	0.9337	0.7325	0.0477	0.3215	0.6364
Indy	0.9512	0.9583	0.9350	0.7812	0.1414	0.0017	0.0061
Loco	0.9648	0.9740	0.9379	0.9479	0.0703	0.0206	0.0324

Table 7: Calibration-only bounds and convergence. Eligible calibration fractions use a calibration C4–C0 gap of at least 0.05.

Animal	Eligible	Mean F_{UB}	Median F_{UB}	Median gap	Converged	Local caps
LINK N	19	1.062	1.050	0.2967	0/19	0
Indy	35	0.676	0.658	0.0014	0/35	0
Loco	8	0.987	0.988	0.0194	0/8	0

Table 8: Allen assignment recall (%) at five repeats per profile. Brackets are marginal 95% percentile bootstrap intervals from 10,000 mouse resamples, not simultaneous intervals. Each of 24 mice receives equal weight after averaging its two directions or two session controls. All candidates remain available; the denominator is the number of registered-common cells. These values differ from the original ten-repeat row-ranked retrieval.

Matcher	Odd/even	Early/late	Across A/B
Movie profile	30.3 [23.8, 36.9]	26.0 [20.3, 31.7]	13.7 [10.0, 17.7]
Neural dynamics	19.8 [16.2, 23.7]	17.1 [13.6, 21.0]	5.4 [3.8, 7.2]
Correlation FAQ	14.1 [9.5, 19.5]	10.6 [6.9, 14.7]	2.2 [1.4, 3.1]
FAQ, PC1 removed	13.2 [9.3, 17.6]	11.3 [7.5, 15.6]	3.3 [2.1, 4.7]
FGW (α=0.25)	30.3 [23.8, 36.9]	25.9 [20.2, 31.6]	14.4 [10.9, 18.1]

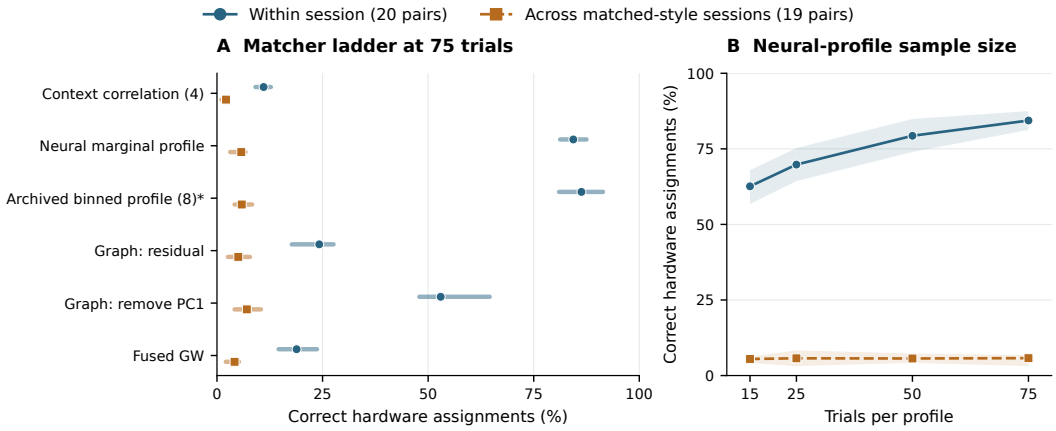

Figure 4: LINK matcher ladder and finite-sample detail. Segments and shading show pairwise interquartile ranges, not confidence intervals. There are 20 local and 19 chronological comparisons within one macaque, with shared style-specific references. The archived eight-bin profile was restored in a separately frozen post-ladder amendment. It is not the initial four-component context-correlation profile. All methods use 75 trials per profile in the left panel.

E EVERY ORACLE-ANALYSIS TARGET

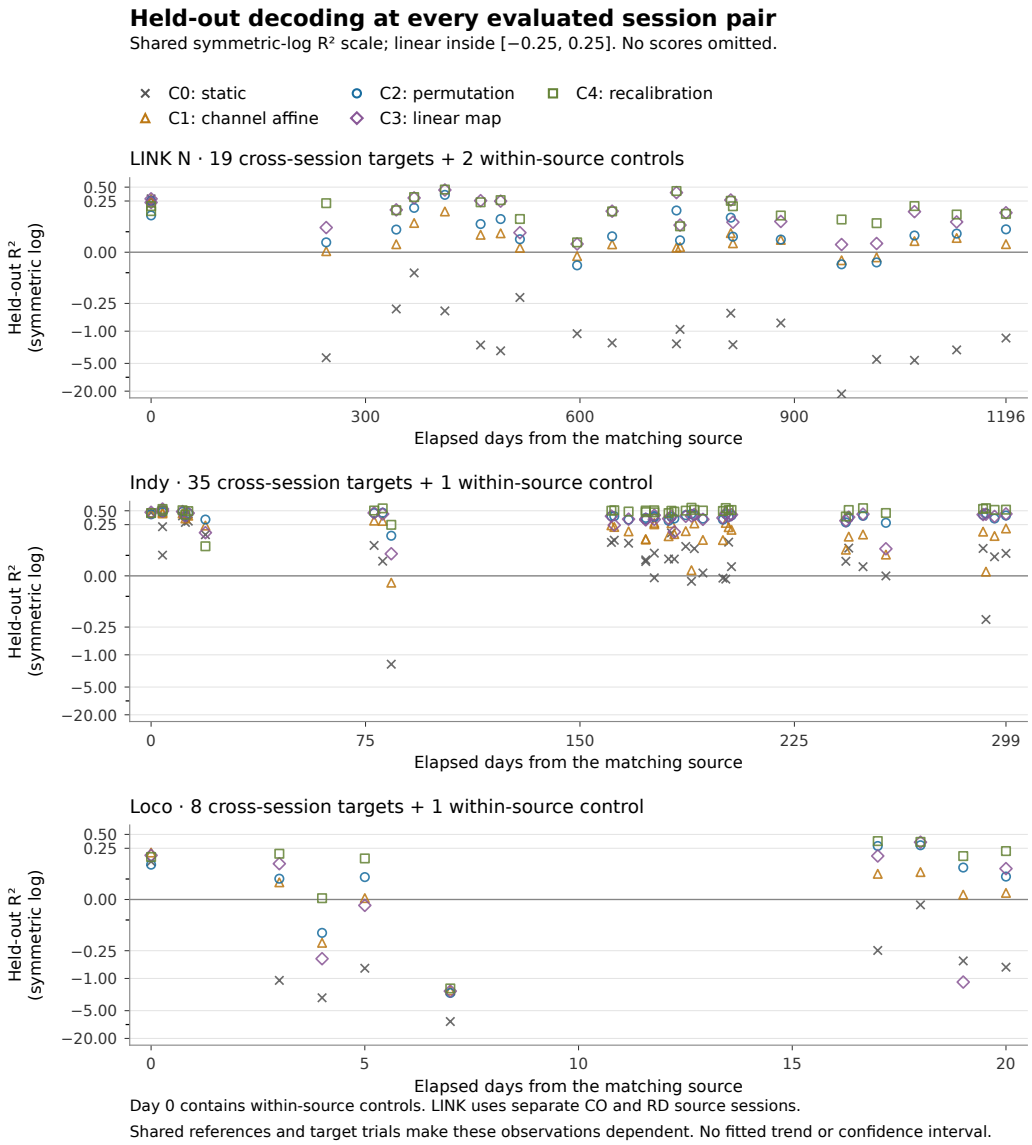

Figure 5: All five classes for every completed E1 unit. The common symmetric-log R^2 scale is linear between -0.25 and 0.25 and retains every negative score. Day zero contains source controls. LINK uses distinct center-out and random-target sources, so elapsed days are measured from each target's matching source. Shared references and evaluation trials induce dependence. No trend or confidence interval is fitted; C1–C4 use endpoint labels.

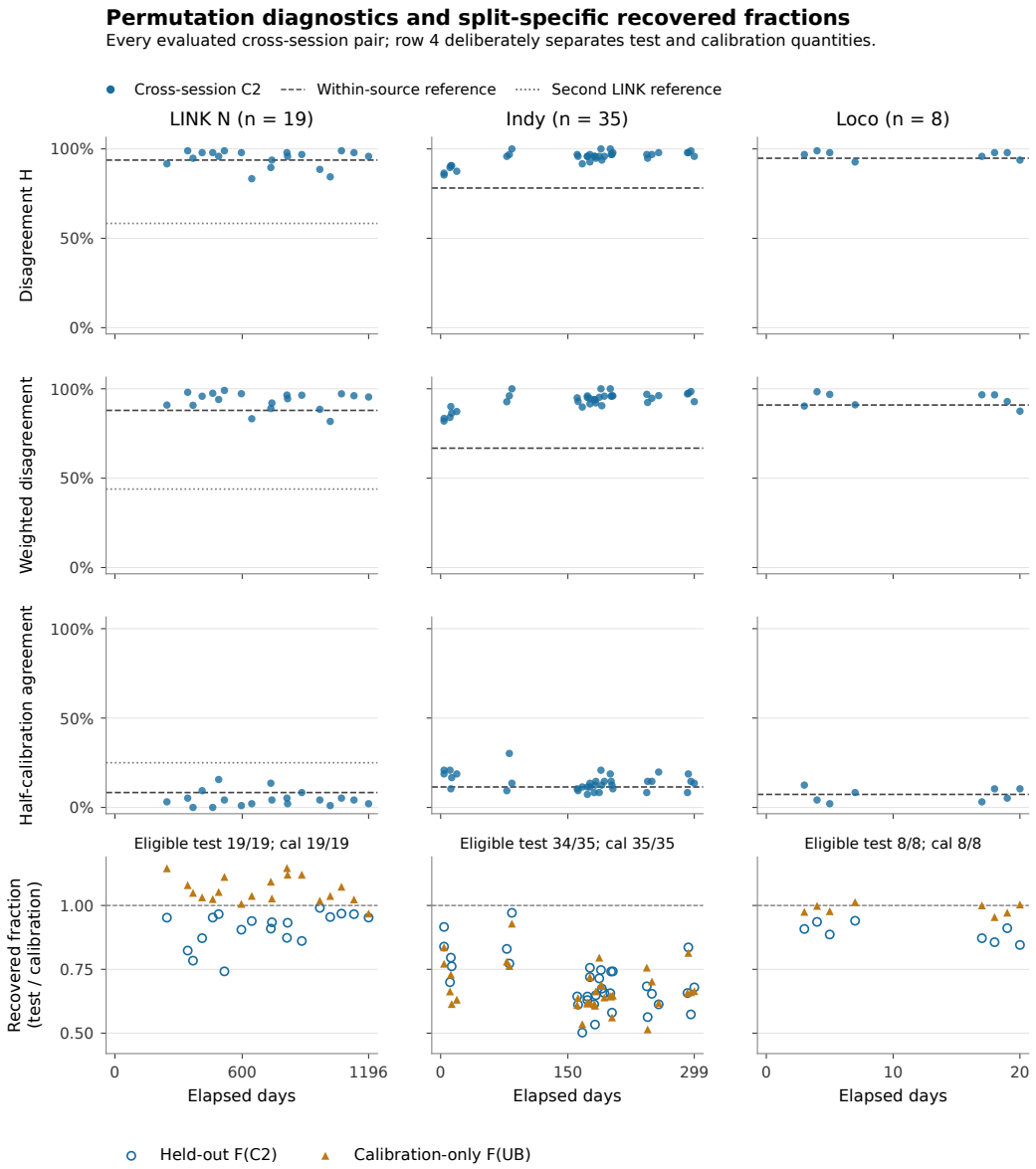

Figure 6: Oracle-map diagnostics for every cross-session target. Within-source reference lines do not have matched calibration budgets. Row four separates test recovered fractions from calibration-only relaxation bounds, each using its own split's gap of at least 0.05. One Indy test unit is ineligible; its raw scores remain in Fig. 5. Fractions are not clipped, and calibration bounds are not held-out ceilings. All animals retain the mixed branch.