SELF-TRAINING NEURAL DECODERS UNDER DRIFT INDUCES SPECTRAL SHRINKAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term neural decoders require recalibration as recording conditions and neural representations drift, but repeated collection of supervised labels limits unattended deployment. We study a fully label-free baseline that refits a decoder to its own predictions. For ridge decoders, the update has the exact form $W_{k+1} = S_k W_k$, where S_k is a data-dependent shrinkage operator. Each update attenuates weight directions by $\sigma_i^2/(\sigma_i^2 + \lambda)$ while leaving the intercept unchanged; under a fixed operator, lower-variance modes are removed fastest and the decoder converges to a constant output. More generally, any update whose targets are computed only from current neural activity, the current decoder, and independent randomness introduces no intent information beyond those variables. On 44 chronic sessions from four subjects in the FALCON benchmark, self-training improves variance-weighted R^2 by 0.087 relative to a frozen decoder while reducing scale-invariant behavioural correlation by 0.137. A frozen decoder rescaled to the same gain achieves higher R^2 without losing correlation, showing that the apparent improvement is explained by attenuation rather than improved alignment with behaviour. This analysis suggests replacing pseudo-label mixing with an explicit proximal prior. With labels from 5% of each session, the resulting update comes within 0.007 R^2 of full daily supervision while retaining substantially more decoder gain, and the same prior also improves fully supervised refits. These results separate the roles of drift detection, intent information, and regularization in long-term decoder maintenance.

1 INTRODUCTION

Intracortical brain–computer interfaces (iBCIs) map neural activity to intended behaviour. The mapping changes over days and months as recording channels fail, spike waveforms change, and neural representations reorganize (Perge et al., 2013; Downey et al., 2018; Sponheim et al., 2022; Colachis et al., 2021; Rule et al., 2019; Driscoll et al., 2017; Chestek et al., 2011). Fixed-decoder performance therefore degrades over long deployments (Pun et al., 2024). The usual remedy is supervised recalibration, which requires a labelled task and often clinical support. For systems intended for independent home use, reducing that supervision burden is a central engineering objective (Simeral et al., 2021; Nuyujukian et al., 2018).

Several successful recalibration methods avoid direct daily labels by obtaining intent information from elsewhere. Retrospective target inference uses the known reach goal (Gilja et al., 2012); self-calibrating cursor control exploits task structure (Jarosiewicz et al., 2015); CORP scores decoded text with a language model before retraining (Fan et al., 2023); and hidden-state models infer discrete targets over long deployments (Wilson et al., 2025). A different family aligns the current neural distribution to the calibration distribution (Degenhart et al., 2020; Farshchian et al., 2019; Ma et al., 2023; Silversmith et al., 2020). These approaches differ in what information is available to the update.

We isolate the limiting case in which the recalibration target is only the decoder's own output. This is the direct analogue of pseudo-label self-training (Lee, 2013; Arazo et al., 2019) and is an attractive fallback when no external task signal is available. For ridge regression, however, the update can be written exactly: refitting to the current decoder's predictions left-multiplies its weights by a shrinkage operator determined by that day's neural design matrix. The operator contains no term that moves the decoder toward the unknown supervised optimum. Its spectral action is strongest on low-variance neural modes.

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.

That spectral ordering is important because neural variance and behavioural relevance need not coincide. Low-variance components can carry substantial task information, and decoders constructed from a small fraction of population variance can outperform those restricted to dominant components (Li et al., 2024; Gallego et al., 2017). A variance-ordered contraction can therefore remove behaviourally useful directions before high-variance directions.

The same information restriction also affects monitoring. A score computed only from the neural activity and the current decoder can reflect changes in those quantities, but without additional assumptions it cannot identify how the current predictions relate to an unobserved intent variable. We make this non-identifiability explicit and compare output-side and input-side monitors under both self-recalibrating and frozen decoders.

We evaluate the mechanism on chronic recordings from the FALCON benchmark (Karpowicz et al., 2024). The central empirical result is a metric dissociation: self-training improves the benchmark's variance-weighted R^2 while reducing scale-invariant behavioural correlation. A control that simply rescales the frozen decoder reproduces and exceeds the R^2 improvement without losing correlation, indicating that attenuation can be rewarded by the reported metric even when the behavioural relationship deteriorates.

Contributions.

- We derive the exact ridge self-training update and characterize its spectral shrinkage (Propositions 1–3).
- We give a model-independent conditional-information statement for targets computed from neural activity and the current decoder (Proposition 4), clarifying the distinction between pseudo-label reuse, input-distribution adaptation, and external intent information.
- We show that label-free self-consistency cannot identify behavioural accuracy in general (Proposition 5) and compare eleven practical monitors on chronic recordings.
- Across 44 chronic sessions from four FALCON subjects, we identify attenuation as the source of the apparent R^2 improvement under self-training using frozen, rescaled, and constant-output controls.
- We replace pseudo-label mixing with an explicit proximal prior and show that a small labelled subset can approach daily supervised refitting while avoiding the same degree of gain collapse.

2 Self-training as a spectral update

Let $X_k \in \mathbb{R}^{T_k \times C}$ denote binned neural activity in session k, $Z_k \in \mathbb{R}^{T_k \times d}$ the corresponding intent, and $\widetilde{X}_k = [X_k \ \mathbf{1}]$ the design matrix. A ridge decoder (Hoerl and Kennard, 1970) $W_k \in \mathbb{R}^{(C+1)\times d}$ predicts $\widetilde{X}_k W_k$. Label-free self-training uses those predictions as targets:

$$T_k = \widetilde{X}_k W_k, \qquad W_{k+1} = \arg\min_W \|\widetilde{X}_k W - T_k\|_F^2 + \lambda \|W_{1:C}\|_F^2, \tag{1}$$

where the offset is unpenalized. With $P = \mathrm{diag}(1, \ldots, 1, 0)$, define $S_k = (\widetilde{X}_k^\top \widetilde{X}_k + \lambda P)^{-1} \widetilde{X}_k^\top \widetilde{X}_k$.

Proposition 1 (Exact recursion)**.** *The update in* (1) *satisfies* $W_{k+1} = S_k W_k$.

The result follows directly from the normal equations. Numerically, we recover the recursion to a maximum relative error of 6.2×10^{-13} over random draws of (T, C, λ) with $\lambda \in [10^{-2}, 10^4]$.

Proposition 2 (Per-session spectral contraction)**.** *After centering the neural features, let* $\sigma_1^2 \geq \cdots \geq \sigma_C^2 \geq 0$ *denote the eigenvalues of the session-k Gram matrix. The weight block of* S_k *has eigenvalues* $\sigma_i^2/(\sigma_i^2 + \lambda)$*, while the intercept direction has eigenvalue* 1*. Thus every identifiable weight direction is contracted for* $\lambda > 0$*. Under repeated application of a fixed operator, the weights converge to zero while the intercept is preserved, yielding a constant-output decoder.*

The update depends on X_k and W_k, but not on the current intent Z_k or on the supervised optimum $W_k^\star = \arg\min_W \|\widetilde{X}_k W - Z_k\|^2 + \lambda \|W_{1:C}\|^2$. Changes in the recording distribution alter S_k and therefore alter the direction and rate of shrinkage, but they do not create a supervised correction term toward $W_k^\star$.

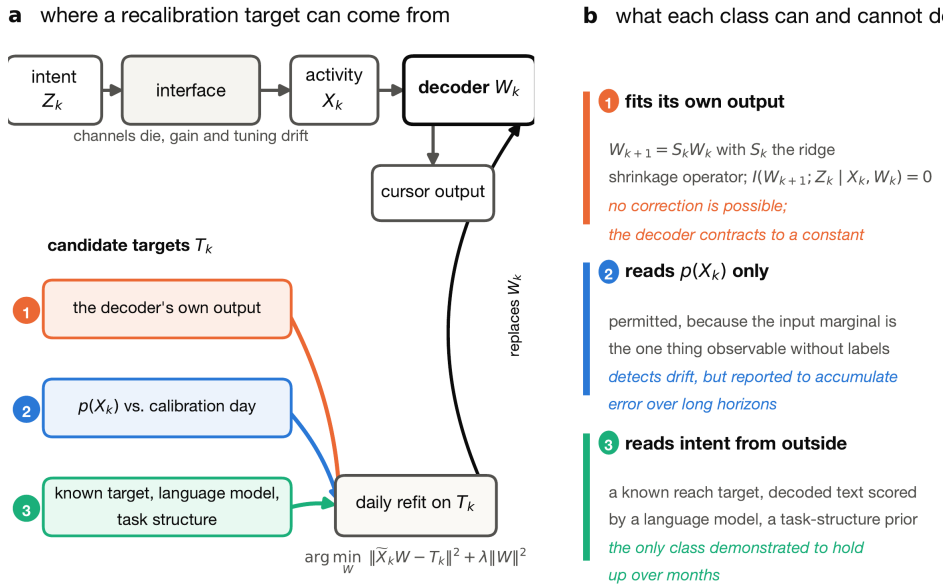


Figure 1: **Sources of information for recalibration.** Class 1 reuses the decoder's own output; Class 2 uses changes in the neural-input distribution; Class 3 introduces an external signal related to intent. The present analysis focuses on Class 1 and on the information distinction among the three classes.

Proposition 3 (Mode-dependent retention)**.** *For a fixed operator with weight-mode eigenvalue* $a_i = \sigma_i^2/(\sigma_i^2 + \lambda)$*, repeated application* m *times retains mode* i *by the factor* a_i^m*. Its half-life is* $\log(1/2)/\log a_i$*, which increases with* σ_i^2 *whenever* $0 < a_i < 1$.

The contraction is therefore not uniform across neural modes. On the 60-mode spectrum used in Fig. 2 with $\lambda = 50$, the strongest mode has a half-life of 56.1 sessions and the weakest 0.14, a factor of 391. Because behavioural information need not be concentrated in high-variance modes (Li et al., 2024), this ordering can preferentially remove behaviourally relevant directions.

2.1 What unlabeled updates can and cannot add

The previous propositions use ridge regression. The following information statement does not.

Proposition 4 (No new intent information from self-generated targets)**.** *Let* $T_k = f(X_k, W_k, \xi_k)$ *for any measurable* f*, where the internal randomness satisfies* $\xi_k \perp Z_k \mid (X_k, W_k)$*, and let* $W_{k+1} = g(X_k, T_k)$ *for any fitting rule* g*. Then*

$$W_{k+1} \perp Z_k \mid (X_k, W_k), \qquad I(W_{k+1}; Z_k \mid X_k, W_k) = 0.$$

The updated decoder is measurable with respect to (X_k, W_k, ξ_k), so the result follows from conditional independence and data processing. The statement does not rule out methods that exploit structural assumptions about distribution shift; it states that a self-generated target cannot introduce intent information that was not already available through (X_k, W_k).

This distinction separates three common recalibration settings (Fig. 1). *Decoder-output targets* reuse only information already represented by the current decoder. *Input-distribution methods* use changes in $p(X_k)$ and can be effective when those changes are linked to a correctable transformation (Degenhart et al., 2020; Farshchian et al., 2019; Ma et al., 2023). *External-intent methods* add information through known goals, task structure, language models, or latent target models (Gilja et al., 2012; Jarosiewicz et al., 2015; Fan et al., 2023; Wilson et al., 2025). The latter two classes are not excluded by Proposition 4; they use information unavailable to pure pseudo-label reuse.

Proposition 5 (Non-identifiability of label-free accuracy)**.** *For the update in* (1)*, self-consistency* $R^2(T_k, \widetilde{X}_k W_{k+1})$ *is a deterministic function of* (X_k, W_k, λ) *and is unchanged if* Z_k *is changed while*

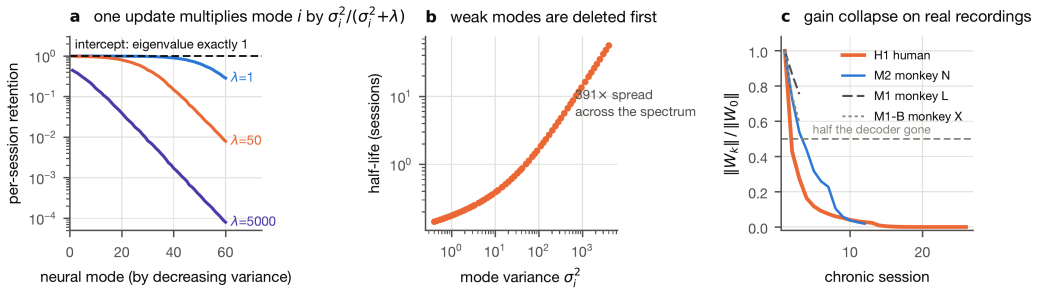


Figure 2: **Self-training acts as a spectral filter. (a)** A ridge self-training update multiplies mode i by $\sigma_i^2/(\sigma_i^2+\lambda)$; the unpenalized intercept is unchanged. **(b)** For a fixed spectrum at $\lambda = 50$, mode half-lives span a factor of 391. **(c)** Decoder gain during self-training on the FALCON chronic recordings.

those variables are fixed. More generally, any monitor $M = h(X_k, W_k)$ cannot distinguish two environments with the same distribution of (X_k, W_k) but different conditional relationships between Z_k and X_k.

Holding X_k and W_k fixed while constructing forty ground truths spanning true $R^2 \in [-224, -0.09]$, self-consistency is identical to machine precision (Fig. 4a); iGCV (Wu et al., 2026) is also invariant in this construction. This is a concrete instance of the general difficulty of label-free performance prediction under distribution shift (Garg et al., 2022).

3 Chronic Evaluation Protocol

Session 0 is the supervised calibration session and supplies the common initial decoder. Each later session is divided contiguously into an *adapt* half and an *evaluation* half. We first score the incoming decoder on the evaluation half, which no arm has fit, and then update on the adapt half. Smoothing, z-scoring, projection onto calibration-session principal components, and causal lag embedding are fixed using session-0 statistics. Appendix A gives full preprocessing and statistical details.

Update arms. FROZEN never updates. SELF-TRAIN applies (1). ANCHORn replaces $n\%$ of the pseudo-label target with true labels; PROBEn fits only those labels. PROXn uses the explicit prior introduced in Sec. 5. ALIGN applies a Procrustes alignment of the leading current neural subspace to the calibration subspace. ORACLE refits on all labelled adapt data, and ORACLE+PRIOR adds the same proximal prior. Two controls isolate output attenuation: RESCALED multiplies the session-0 decoder by the gain reached by self-training, and OFF emits the calibration-session mean.

Metrics. We report the benchmark's variance-weighted R^2 (Karpowicz et al., 2024) and the variance-weighted mean per-dimension correlation ρ. Correlation is invariant to affine output scaling, so attenuation can change R^2 without changing ρ. Under optimal per-dimension affine recalibration, $R_j^2 = \rho_j^2$.

4 Results

4.1 Simulated Chronic Degradation

We simulate 60 sessions of a 128-channel iBCI with Poisson spiking and four concurrent degradation processes: channel death, gain drift, tuning rotation, and baseline drift, over 24 seeds. Relative to a frozen decoder, self-training is worse ($\Delta R^2 = -0.020 \pm 0.003$, $p = 4.8 \times 10^{-12}$, paired). Input-side alignment recovers 24.6% of the frozen-to-oracle gap ($+0.101 \pm 0.028$, $p = 2.6 \times 10^{-7}$), while a 5% labelled anchor recovers 84.7%.

Table 1: Results pooled over 44 chronic sessions and four subjects. R^2 is the benchmark variance-weighted score; ρ is scale-invariant; gain is $\|W_k\|/\|W_0\|$.

arm	median R^2	median ρ	mean gain
frozen	-0.167	0.303	1.000
self-training	-0.084	0.153	0.253
frozen, matched rescaling	$+0.007$	0.303	1.000
constant output	-0.002	0.000	—
self-training, gain restored	-22.16	0.153	1.000
input alignment	-0.356	0.178	1.000
5% labels + pseudo-labels (ANCHOR)	-0.028	0.231	0.278
5% labels only (PROBE)	-0.061	0.192	0.845
5% labels + prior (PROX)	-0.001	0.230	0.719
10% labels + prior	$+0.014$	0.258	0.661
full daily labels (ORACLE)	-0.018	0.254	0.667
full daily labels + prior	$\mathbf{+0.041}$	**0.291**	0.597

4.2 REAL CHRONIC RECORDINGS

We use four FALCON subjects (Karpowicz et al., 2024): one human participant performing 7-DoF reach and grasp over forty days (27 sessions, 176 units), one monkey performing finger movements over thirty-six days (13 sessions, 96 units), and two monkeys performing reach-to-grasp with EMG targets (4 sessions each, 64 units). The evaluation therefore contains 44 chronic sessions across two species and three task families.

Self-training improves the reported R^2 relative to freezing ($\Delta R^2 = +0.087$, $p = 7.1 \times 10^{-6}$) but reduces behavioural correlation ($\Delta\rho = -0.137$, $p = 3.9 \times 10^{-8}$). Tests are paired over the 44 chronic sessions after averaging random seeds within session. All four subjects show the same direction in R^2, and the subjects with substantial gain contraction also lose correlation.

The matched-rescaling control isolates the main source of the R^2 change. Multiplying the frozen session-0 decoder by the scalar gain attained by self-training, without refitting to the current session, produces a higher R^2 than self-training ($\Delta = +0.088$, $p = 3.6 \times 10^{-8}$) while preserving $\rho = 0.303$. Thus the improvement in R^2 does not require better alignment with behaviour; attenuation alone reproduces it. Self-training also scores below the constant-output control on R^2 ($\Delta = -0.082$, $p = 5.9 \times 10^{-6}$), illustrating how an unbounded-below metric can reward reduced output variance when the incoming decoder is poorly calibrated.

The contraction is substantial. Mean gain falls to 0.25 of its calibration value, and for the human participant it falls below 0.5 by the second chronic session (Fig. 2c). Restoring the self-trained decoder to the norm of W_0 leaves ρ unchanged but reduces R^2 to -22.2, indicating that the lost performance is not explained by a scalar gain change alone. Mode-retention measurements also follow the predicted ordering: over the first eight sessions, the decoder retains more weight on the strongest sixth of the neural spectrum than on the weakest sixth (mean difference $+0.063$, $p = 6.9 \times 10^{-10}$, paired within subject, session, and seed).

Pooling multiple early sessions for calibration does not remove the effect. For H1 and M2, calibration on the first four and three sessions respectively yields calibration R^2 of 0.214 and 0.161. On the remaining 33 chronic sessions, self-training changes R^2 by $+0.103$ and ρ by -0.115, with mean gain falling to 0.20; matched rescaling again exceeds self-training by 0.054 R^2.

Input alignment. The Procrustes alignment arm improves the simulated setting but underperforms freezing on the real recordings ($\Delta R^2 = -0.072$, $\Delta\rho = -0.120$, $p = 2.8 \times 10^{-12}$). Sweeping the aligned subspace dimension reduces the damage as the transformation approaches the identity; even an oracle that selects the dimension using held-out performance remains below freezing in correlation ($\Delta\rho = -0.028$, $p = 0.014$). This result is specific to the transformation family evaluated here and is consistent with prior reports that distribution matching can accumulate long-horizon error (Wilson et al., 2025). It does not contradict methods that exploit stronger structure in the input shift.

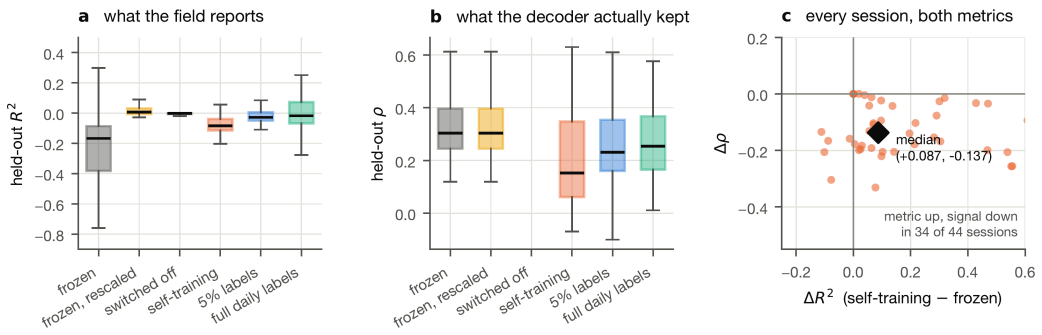


Figure 3: R^2 **and scale-invariant correlation move in opposite directions under self-training.** **(a)** R^2 by arm. **(b)** ρ by arm; matched rescaling coincides with the frozen decoder because ρ is scale-invariant. **(c)** Session-level changes under self-training relative to freezing.

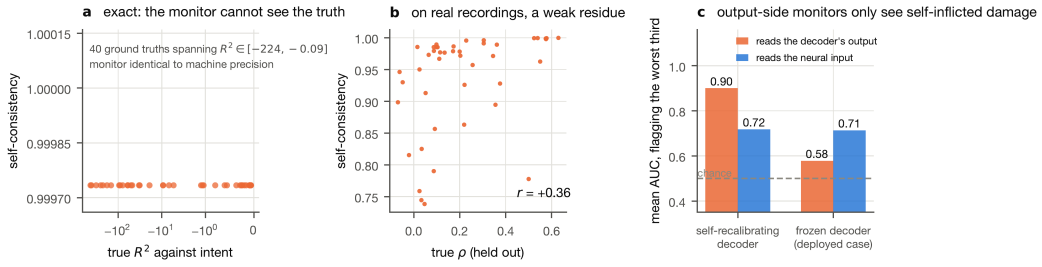


Figure 4: **Limits of label-free monitoring.** **(a)** For fixed (X, W), forty alternative ground truths spanning $R^2 \in [-224, -0.09]$ produce identical self-consistency. **(b)** Association between self-consistency and performance on real recordings. **(c)** Mean AUC by monitor family under self-recalibrating and frozen decoders.

4.3 Label-free monitoring

We evaluate eleven monitors against held-out ρ: seven read the decoder output or parameters (self-consistency, iGCV, output scale, decoder gain, ensemble disagreement, pseudo-label residual quantile, and output roughness) and four use only neural inputs (channel-marginal KL, latent-distribution KL, principal-subspace angle, and rate-drop fraction). Each is scored by AUC for detecting the worst third of sessions.

With a self-recalibrating decoder, output-side monitors achieve mean AUC 0.900, with decoder gain at 0.938. However, this performance partly reflects the update dynamics encoded in W_k itself. Repeating the audit with the decoder frozen reduces the output-side mean AUC to 0.578, with four of seven monitors below chance, whereas the input-side family is nearly unchanged ($0.718 \to 0.713$; Fig. 4c). Self-consistency is weakly associated with held-out performance during self-training (within-subject $\rho_s = 0.17$); under a frozen decoder it becomes a statistic of the changing input distribution and reaches AUC 0.844. These results support treating decoder-state and input-drift monitors as different signals rather than interchangeable measures of behavioural accuracy.

5 Replacing pseudo-labels with an explicit prior

For ridge self-training, pseudo-labels do not provide additional intent information, but mixing them with a small labelled subset changes the optimization objective. In particular, the pseudo-label term acts as a data-dependent pull toward the current decoder. This suggests separating that regularization effect from the pseudo-labels themselves.

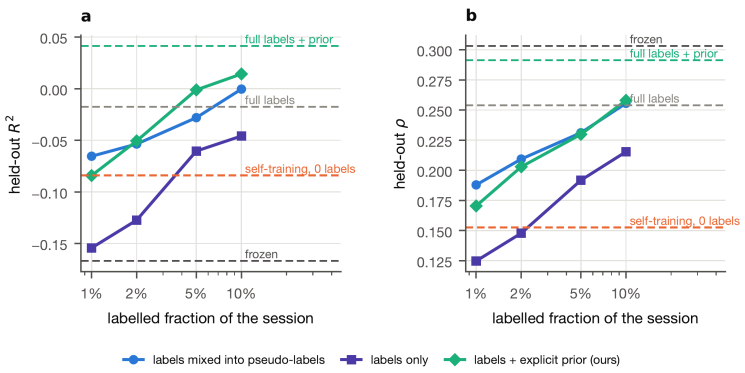


Figure 5: **Label-budget comparison.** Explicit proximal recalibration in (2) is compared with pseudo-label mixing and labels-only fitting across supervision budgets.

We therefore fit, on the labelled subset ℓ,

$$W_{k+1} = \arg\min_{W} \ \|\widetilde{X}_k^{(\ell)} W - Z_k^{(\ell)}\|_F^2 + \lambda \|W_{1:C}\|_F^2 + \mu \|W - W_k\|_F^2. \tag{2}$$

At a 5% label budget, labels-only fitting underperforms the pseudo-label mixture in correlation ($\Delta\rho = -0.027$, $p = 2.9 \times 10^{-2}$), consistent with a useful regularization effect. Making that effect explicit improves over labels-only fitting in every subject ($\Delta\rho = +0.036$, $p = 3.0 \times 10^{-7}$; $\Delta R^2 = +0.047$, $p = 5.7 \times 10^{-7}$) and also improves correlation over the pseudo-label mixture ($+0.009$, $p = 5.4 \times 10^{-3}$). The explicit prior retains mean gain 0.72 compared with 0.28 for the pseudo-label mixture. With 5% of a session labelled, PROX is within 0.007 R^2 and 0.004 ρ of full daily refitting.

The same prior also benefits full supervision: full daily labels plus the proximal term reach $R^2 = 0.041$ and $\rho = 0.291$, compared with -0.018 and 0.254 for the unregularized oracle. This comparison indicates that the useful component of pseudo-label mixing is better represented as an explicit prior whose strength can be chosen independently of the fraction of pseudo-labels. The connection is analogous to parameter-stability penalties used in sequential learning (Kirkpatrick et al., 2017), here applied to repeated regression updates.

6 Related work

Self-training and recursive training. Pseudo-label methods can exhibit confirmation bias (Arazo et al., 2019; Lee, 2013), and recursively training on generated data can degrade learned distributions (Shumailov et al., 2024). Wu et al. (2026) derive asymptotic risk recursions for self-training in overparameterized linear regression with independently resampled designs, separating denoising from signal forgetting. Entropy minimization and test-time adaptation provide related unlabeled objectives (Grandvalet and Bengio, 2004; Wang et al., 2021). Our analysis gives a finite-sample operator form for ridge self-training and connects that operator to chronic decoder maintenance, metric behaviour, and monitoring.

iBCI recalibration. Closed-loop decoder adaptation has a substantial literature (Dangi et al., 2013; 2014; Shanechi et al., 2016; Orsborn et al., 2014). Target-inference methods obtain supervisory information from known goals or task structure (Gilja et al., 2012; Jarosiewicz et al., 2015; Fan et al., 2023; Wilson et al., 2025), while subspace stabilization and adversarial alignment operate on the input distribution (Degenhart et al., 2020; Farshchian et al., 2019; Ma et al., 2023; Silversmith et al., 2020). These two families correspond to the information sources separated by Proposition 4. FALCON (Karpowicz et al., 2024), the Neural Latents Benchmark (Pei et al., 2021), and cross-session neural foundation models (Ye et al., 2023; Azabou et al., 2023; Pandarinath et al., 2018) address complementary questions of few-shot and cross-session neural decoding.

Monitoring. MINDFUL measures divergence between a reference neural distribution and a moving window (Pun et al., 2024). More generally, predicting performance under shift without labels is an active problem (Garg et al., 2022). Common uncertainty estimators such as dropout, ensembles, calibration scores, OOD scores, and conformal intervals (Gal and Ghahramani, 2015; Lakshminarayanan et al., 2016; Guo et al., 2017; Hendrycks and Gimpel, 2016; Angelopoulos and Bates, 2021) measure properties of model predictions or input distributions; Proposition 5 emphasizes that such quantities require additional assumptions before they can identify behavioural accuracy.

7 Scope and Limitations

The spectral formulas are exact for ridge regression. For a changing sequence of session operators, Propositions 2–3 characterize each update locally; global convergence of an arbitrary operator sequence requires additional conditions. The conditional-information result in Proposition 4 is model-independent. Appendix B evaluates MLP and GRU decoders trained by gradient descent and finds the same qualitative direction, with contraction strength determined by optimization and regularization rather than a closed-form ridge factor.

The evaluation is open-loop, which isolates decoder changes from user adaptation. Proposition 4 constrains the information available to the update rule in either setting, but the behavioural consequences of closed-loop user compensation require direct measurement. The real-data study contains four subjects across two species and three task families; session-level paired tests quantify within-record effects, while the small number of subjects remains the main limit on population-level generalization.

8 Conclusion

Refitting a neural decoder to its own predictions does not provide new information about current intent. For ridge regression, the update is exactly a data-dependent spectral shrinkage operator, and low-variance directions are attenuated most strongly. On chronic FALCON recordings, self-training raises variance-weighted R^2 while reducing scale-invariant behavioural correlation; matched rescaling shows that attenuation alone explains the apparent metric improvement. Replacing pseudo-label mixing with a small amount of labelled data and an explicit proximal prior preserves the regularization benefit without tying it to repeated self-distillation. More broadly, long-term recalibration should distinguish three separate resources: information about input drift, information about intent, and regularization of the update.

AI use statement

GPT was used solely to polish the wording of this manuscript.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. doi: 10.48550/arxiv.2107.07511.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *arXiv preprint arXiv:1908.02983*, 2019.

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

Cynthia A Chestek, Vikash Gilja, Paul Nuyujukian, Justin D Foster, Joline M Fan, Matthew T Kaufman, Mark M Churchland, Zuley Rivera-Alvidrez, John P Cunningham, Stephen I Ryu, and Krishna V Shenoy. Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex. *Journal of Neural Engineering*, 8:045005, 2011. doi: 10.1088/1741-2560/8/4/045005.

Samuel C. Colachis, Collin Dunlap, Nicholas V. Annetta, Sanjay Tamrakar, Marcia Bockbrader, and David A. Friedenberg. Long-term intracortical microelectrode array performance in a human: a 5 year retrospective analysis. *Journal of Neural Engineering*, 2021. doi: 10.1088/1741-2552/ac1add.

Siddharth Dangi, Amy L. Orsborn, Helene G. Moorman, and Jose M. Carmena. Design and analysis of closed-loop decoder adaptation algorithms for brain–machine interfaces. *Neural Computation*, 25(7):1693–1731, 2013.

Siddharth Dangi, Suraj Gowda, Helene G. Moorman, Amy L. Orsborn, Kelvin So, Maryam Shanechi, and Jose M. Carmena. Continuous closed-loop decoder adaptation with a recursive maximum likelihood algorithm allows for rapid performance acquisition in brain–machine interfaces. *Neural Computation*, 26(9):1811–1839, 2014.

Alan D. Degenhart, William E. Bishop, Emily R. Oby, Elizabeth C. Tyler-Kabara, Steven M. Chase, Aaron P. Batista, and Byron M. Yu. Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nature Biomedical Engineering*, 4:672–685, 2020. doi: 10.1038/s41551-020-0542-9.

John E Downey, Nathaniel Schwed, Steven M Chase, Andrew B Schwartz, and Jennifer L Collinger. Intracortical recording stability in human brain–computer interface users. *Journal of Neural Engineering*, 15:046016, 2018. doi: 10.1088/1741-2552/aab7a0.

Laura N. Driscoll, Noah L. Pettit, Matthias Minderer, Selmaan N. Chettih, and Christopher D. Harvey. Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell*, 170:986–999.e16, 2017. doi: 10.1016/j.cell.2017.07.021.

Chaofei Fan, Nick Hahn, Foram Kamdar, Donald Avansino, Guy Wilson, Leigh Hochberg, Krishna V Shenoy, Jaimie Henderson, and Francis Willett. Plug-and-play stability for intracortical brain–computer interfaces: A one-year demonstration of seamless brain-to-text communication. In *Advances in Neural Information Processing Systems 36*, 2023.

Ali Farshchian, Juan A. Gallego, Joseph P. Cohen, Yoshua Bengio, Lee E. Miller, and Sara A. Solla. Adversarial domain adaptation for stable brain–machine interfaces. In *International Conference on Learning Representations*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2015. doi: 10.48550/arxiv.1506.02142.

Juan A. Gallego, Matthew G. Perich, Lee E. Miller, and Sara A. Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022.

Vikash Gilja, Paul Nuyujukian, Cindy A Chestek, John P Cunningham, Byron M Yu, Joline M Fan, Mark M Churchland, Matthew T Kaufman, Jonathan C Kao, Stephen I Ryu, and Krishna V Shenoy. A high-performance neural prosthesis enabled by control algorithm design. *Nature Neuroscience*, 15:1752–1757, 2012. doi: 10.1038/nn.3265.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems 17*, volume 17, pages 281—-296, 2004.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. doi: 10.48550/arxiv.1706.04599.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2016. doi: 10.48550/arxiv.1610.02136.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Beata Jarosiewicz, Anish A. Sarma, Daniel Bacher, Nicolas Y. Masse, John D. Simeral, Brittany Sorice, Erin M. Oakley, Christine Blabe, Chethan Pandarinath, Vikash Gilja, Sydney S. Cash, Emad N. Eskandar, Gerhard Friehs, Jaimie M. Henderson, Krishna V. Shenoy, John P. Donoghue, and Leigh R. Hochberg. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. *Science Translational Medicine*, 7, 2015. doi: 10.1126/scitranslmed.aac7328.

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 37, Datasets and Benchmarks Track*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521–3526, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, 2016. doi: 10.48550/arxiv.1612.01474.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Yangang Li, Xinyun Zhu, Yu Qi, and Yueming Wang. Revealing unexpected complex encoding but simple decoding mechanisms in motor cortex via separating behaviorally relevant neural signals. *eLife*, 13:e87881, 2024. doi: 10.7554/eLife.87881.

Xuan Ma, Fabio Rizzoglio, Eric J. Perreault, Lee E. Miller, and Ann Kennedy. Using adversarial networks to extend brain computer interface decoding accuracy over time. *eLife*, 12:e84296, 2023.

Paul Nuyujukian, Jose Albites Sanabria, Jad Saab, Chethan Pandarinath, Beata Jarosiewicz, Christine H. Blabe, Brian Franco, Stephen T. Mernoff, Emad N. Eskandar, John D. Simeral, Leigh R. Hochberg, Krishna V. Shenoy, and Jaimie M. Henderson. Cortical control of a tablet computer by people with paralysis. *PLOS ONE*, 13:e0204566, 2018. doi: 10.1371/journal.pone.0204566.

Amy L. Orsborn, Helene G. Moorman, Simon A. Overduin, Maryam M. Shanechi, Dragan F. Dimitrov, and Jose M. Carmena. Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control. *Neuron*, 82:1380–1393, 2014. doi: 10.1016/j.neuron.2014.04.048.

Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, L. F. Abbott, and David Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15:805–815, 2018. doi: 10.1038/s41592-018-0109-9.

Felix Pei, Joel Ye, David L. Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark M. Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan W. Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity. *arXiv preprint arXiv:2109.04463*, 2021.

János A Perge, Mark L Homer, Wasim Q Malik, Sydney Cash, Emad Eskandar, Gerhard Friehs, John P Donoghue, and Leigh R Hochberg. Intra-day signal instabilities affect decoding performance in an intracortical neural interface system. *Journal of Neural Engineering*, 10:036004, 2013. doi: 10.1088/1741-2560/10/3/036004.

Tsam Kiu Pun, Mona Khoshnevis, Tommy Hosman, Guy H. Wilson, Anastasia Kapitonava, Foram Kamdar, Jaimie M. Henderson, John D. Simeral, Carlos E. Vargas-Irwin, Matthew T. Harrison,

and Leigh R. Hochberg. Measuring instability in chronic human intracortical neural recordings towards stable, long-term brain-computer interfaces. *Communications Biology*, 7, 2024. doi: 10.1038/s42003-024-06784-4.

Michael E Rule, Timothy O'Leary, and Christopher D Harvey. Causes and consequences of representational drift. *Current Opinion in Neurobiology*, 58:141–147, 2019. doi: 10.1016/j.conb.2019.08.005.

Maryam M. Shanechi, Amy L. Orsborn, and Jose M. Carmena. Robust brain–machine interface design using optimal feedback control modeling and adaptive point process filtering. *PLOS Computational Biology*, 12(4), 2016. doi: 10.1371/journal.pcbi.1004730.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Daniel B. Silversmith, Reza Abiri, Nicholas F. Hardy, Nikhilesh Natraj, Adelyn Tu-Chan, Edward F. Chang, and Karunesh Ganguly. Plug-and-play control of a brain–computer interface through neural map stabilization. *Nature Biotechnology*, 39:326–335, 2020. doi: 10.1038/s41587-020-0662-5.

John D. Simeral, Thomas Hosman, Jad Saab, Sharlene N. Flesher, Marco Vilela, Brian Franco, Jessica N. Kelemen, David M. Brandman, John G. Ciancibello, Paymon G. Rezaii, Emad N. Eskandar, David M. Rosler, Krishna V. Shenoy, Jaimie M. Henderson, Arto V. Nurmikko, and Leigh R. Hochberg. Home use of a percutaneous wireless intracortical brain-computer interface by individuals with tetraplegia. *IEEE Transactions on Biomedical Engineering*, 68:2313–2325, 2021. doi: 10.1109/tbme.2021.3069119.

Caleb Sponheim, Vasileios Papadourakis, Jennifer L. Collinger, John E. Downey, Jeffrey M. Weiss, Lida Pentousi, Kaisa Elliott, and Nicholas G. Hatsopoulos. Longevity and reliability of chronic unit recordings using the utah, intracortical multi-electrode arrays. *Journal of Neural Engineering*, 2022. doi: 10.1088/1741-2552/ac3eaf.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Guy H. Wilson, Francis R. Willett, Elias A. Stein, Foram Kamdar, Donald T. Avansino, Leigh R. Hochberg, Krishna V. Shenoy, Shaul Druckmann, and Jaimie M. Henderson. Long-term unsupervised recalibration of cursor-based intracortical brain–computer interfaces using a hidden Markov model. *Nature Biomedical Engineering*, 2025. doi: 10.1038/s41551-025-01536-z.

Mingqi Wu, Archer Y. Yang, and Qiang Sun. Why self-training helps and hurts: Denoising vs. signal forgetting. *arXiv preprint arXiv:2602.14029*, 2026.

Joel Ye, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. Neural Data Transformer 2: Multi-context pretraining for neural spiking activity. In *Advances in Neural Information Processing Systems 36*, 2023.

A Protocol details

Preprocessing. Spike counts are binned at the dataset's native resolution, square-root transformed, smoothed with a Gaussian kernel ($\sigma = 2.5$ bins), z-scored with session-0 statistics, projected onto the top 48 session-0 principal components, and lag-embedded with causal lags $\{0, 3, 6, 10\}$ bins. The projection keeps the feature count (192) below the bin count of the shortest released session. All transforms are fixed from session 0. Only bins passing the dataset's `eval_mask` are scored; the mask is applied after smoothing and lag construction so temporal features are not constructed across excluded intervals.

Regularization. λ is selected once per subject by a contiguous 80/20 split of the calibration session over 25 values from 10^0 to 10^6 and is then held fixed for every arm and session. Calibration-session held-out R^2 is 0.197 (H1), 0.319 (M2), 0.617 (M1), and 0.657 (M1-B). H1 uses a single-session calibration for a 7-DoF task, unlike pooled multi-session baselines reported elsewhere for the benchmark. In (2), μ is set to the probe block's ridge scale, so no additional hyperparameter is tuned.

Metrics. Variance-weighted R^2 is $1 - \sum_j \mathrm{SS}_j^{\mathrm{res}} / \sum_j \mathrm{SS}_j^{\mathrm{tot}}$, matching the FALCON metric (Karpowicz et al., 2024) and `sklearn`'s `variance_weighted` implementation to machine precision. ρ is the variance-weighted mean of per-dimension Pearson correlations.

Statistics. Random seeds are averaged within each (subject, session, arm) combination before testing. Unless stated otherwise, contrasts are paired over the 44 chronic sessions and evaluated with two-sided Wilcoxon signed-rank tests. Seeds vary only the selected labelled indices and are treated as repeated randomizations rather than independent observations. We also report directional consistency across the four subjects for headline comparisons.

B NONLINEAR DECODERS

Proposition 4 does not depend on model class, although the dynamics of self-training do. We repeat the simulated degradation experiment with a 2×256 GELU MLP and a 128-unit GRU, both trained with AdamW and dropout, using six seeds and 30 sessions. Final-session R^2 (mean $\pm$ 95% CI) is:

	frozen	self-training	5% anchor	full labels
MLP	0.751 ± 0.027	0.677 ± 0.037	0.808 ± 0.026	0.887 ± 0.015
GRU	0.745 ± 0.032	0.667 ± 0.092	0.813 ± 0.051	0.892 ± 0.011

Self-training underperforms freezing for both models (-0.074, $p = 0.002$ for the MLP; -0.078, $p = 0.10$ for the GRU), while a 5% labelled anchor improves on freezing ($p = 0.001$ and $p = 0.014$). The nonlinear decoders contract less strongly than ridge because the effective update is set by finite-step optimization, weight decay, and network geometry rather than a closed-form ridge operator.

Self-consistency remains high throughout (0.997 for the MLP and 0.991 for the GRU) and correlates weakly with true accuracy (-0.07 and $+0.37$, respectively). MC-dropout predictive variance (Gal and Ghahramani, 2015) correlates at $+0.81$ with accuracy for the MLP: as the decoder output contracts, predictive variance can decrease together with performance. Thus lower output uncertainty need not indicate improved behavioural decoding under self-training.