Supplementary Information

Assistive sensory-motor perturbations influence learned neural representations

Pavithra Rajeswaran[1], **Alexandre Payeur**[2,3], **Guillaume Lajoie**[2,3], **and Amy L. Orsborn**[1,4,5,*]

[1]University of Washington, Bioengineering, Seattle, 98115, USA
[2]Université de Montréal, Department of Mathematics and Statistics, Montréal (QC), Canada, H3C 3J7
[3]Mila - Québec Artificial Intelligence Institute, Montréal (QC), Canada, H2S 3H1
[4]University of Washington, Electrical and Computer Engineering, Seattle, 98115, USA
[5]Washington National Primate Research Center, Seattle, Washington, 98115, USA
[*]aorsborn@uw.edu

Supplementary Methods

Neural tuning curves

Neural tuning curves were defined as the mean firing rate of the neuron in a time window (0-300ms after the go cue) as a function of the eight target directions. We first identified the mean firing rate of each unit per target direction and then modeled the relationship between neural activity and movement direction via a cosine tuning model [1]:

$$f = B_1 \cos\theta + B_2 \sin\theta + B_3$$

where θ represents the target direction and B_1, B_2 and B_3 are model coefficients. The coefficients were estimated via linear regression and then used to compute each unit's modulation depth (MD) and preferred direction (PD):

$$MD = \sqrt{B_1^2 + B_2^2}$$
$$PD = \arctan2(B_2/B_1)$$

We calculated the change in tuning properties within each learning series by comparing MDs on early and late training days: $\Delta\text{MD} = \text{MD}_{\text{late}} - \text{MD}_{\text{early}}$. To compare these tuning changes to our model, we calculated the change in model coefficients between early and late training day: $\Delta\mathbf{w} = \mathbf{w}_{\text{late}} - \mathbf{w}_{\text{early}}$, where $\mathbf{w}_{\text{late,early}} \in \mathbb{R}^{n_{\text{units}}}$ were calculated from the model coefficients by taking the mean across time bins and then, for each unit, selecting the mean weight for the most contributing target direction. Figure S2B shows the correlation between ΔMD and $\Delta\mathbf{w}$.

Comparison of offline velocity estimates

We used the same velocity Kalman filter decoders from the experiments to estimate cursor velocities offline. We compared velocity estimates obtained using all readouts and using only designated subsets of readouts (see below and Fig. S4G-I for detail). To quantify the similarity between the estimated velocities, we calculated the following time-average matching:

$$M = 1 - \frac{1}{T}\sum_{t=1}^{T} \begin{cases} 0 & \text{if } \|\mathbf{V}_t^{(\text{all})}\| = \|\mathbf{V}_t^{(\text{subset})}\| = 0 \\ \frac{\|\mathbf{V}_t^{(\text{all})} - \mathbf{V}_t^{(\text{subset})}\|^2}{\|\mathbf{V}_t^{(\text{all})}\|^2 + \|\mathbf{V}_t^{(\text{subset})}\|^2} & \text{otherwise} \end{cases}$$

where $\mathbf{V}_t^{(\text{all})}$ and $\mathbf{V}_t^{(\text{subset})}$ represent the cursor velocities estimated using all readouts and the subset, respectively. We defined the summand as zero when both velocities are zero. This ensures that $0 \leq M \leq 1$. We compared the velocities estimated using

all readouts and the top N_c^{late} units—from the target-encoding analysis—on late day (Fig. S4G). To assess how the addition of units improved the accuracy of cursor velocity estimates, we computed M as we progressively added more units ranked according to the neuron adding curves obtained from the target-encoding analysis on early and late days (Fig. S4H). Finally, we compared the matching score M for the two training phases when using a number of units equal to N_c^{late} on both the early and late day (Fig. S4I).

Supplementary Tables

Table 1. Information about the number of readouts and nonreadouts, the number of readout units replaced during ensemble change events, the number of weight change events and the number of ensemble change events for every series analyzed in the main text. The '+' and '-' indicate the number of units newly added or removed, respectively. The '-' symbol in the ensemble change events column indicates no ensemble changes for that series. The range in the nonreadouts column denotes the total number of nonreadouts in the early and late stages.

Subject	# of readouts (early)	# of readouts (late)	# of readout units replaced	# of weight change events	# of ensemble change events	# of nonreadouts (early - late)
J	16	16	1	4	1	36 - 36
J	16	16	4	3	1	39 - 39
J	16	16	3	4	2	41 - 41
J	16	16	2	2	1	38 - 38
J	16	15	+2 -1	2	1	37 - 38
J	16	15	-1	1	1	28 - 29
J	16	16	3	2	1	29 - 29
S	12	12	0	2	-	118 - 118
S	11	11	4	2	2	66 - 66
S	12	12	0	1	-	121 - 121

Supplementary Discussion

Compact Representation in Units Does Not Imply Compact Representation in Neural Modes

We observed compactness in both the "mode" space and the "individual unit" space. Here, we demonstrate that these observations do not simply follow from one another. We consider a two-class classification problem, eliminating the complexity of multiple targets, and we ignore the time dimension. We define x as a random column vector representing centered single-unit activities and consider a logistic regression model $y(x) = f(w^\top x + b)$, where $^\top$ denotes transpose, and w and b are the model parameters. After fitting this model to data and achieving good generalization, we assume w has been determined.

Drawing inspiration from Fig. 3B, we define a representation as compact if the weight vector w is sparse, specifically having a few dominant elements (ideally one, for the sake of this discussion). We introduce matrix A containing principal vectors as columns, forming an orthonormal matrix ($AA^\top = A^\top A = I$, where I is the identity matrix). This allows us to reformulate the logistic regression model in terms of principal components and their projections: $y(x) = f((A^\top w)^\top (A^\top x) + b)$, where $A^\top x$ represents the principal components of x, and $A^\top w$ is the projection of w onto the principal vectors.

If w is sparse, its projection $A^\top w$ will effectively highlight the contribution of the dominant unit(s) in the space of principal vectors. However, whether the resulting weight vector in this transformed space remains compact (sparse) depends on the characteristics of the principal vectors and, thus, on the covariance matrix of the data. This observation underscores that a compact representation in "unit space" does not automatically imply a compact representation in the "mode space".

The essence of this discussion is that the transformation of the basis (through rotation or otherwise) does not necessarily preserve the sparsity of the representation. Thus, the compactness of representations across different spaces (unit vs. mode) is not a trivial matter and requires careful consideration of the underlying data structure and transformation methods used.

Impact of Neural Recording Methods

Differences in neural recording methods, such as threshold crossings versus sorted units, may also influence co-adaptation processes and the stability of neural representations. For instance, we used threshold crossings in monkey J. Threshold crossings may have different stability properties over days than sorted units, which might lead to differences across monkeys. However, isolating contributions of neural recording properties is beyond the scope of what is feasible with these current data. Future

studies exploring how neural recording stability, specifically, influences neural plasticity and decoder adaptation algorithms will be critical for designing BCIs that maintain performance over time.

Supplementary Figures

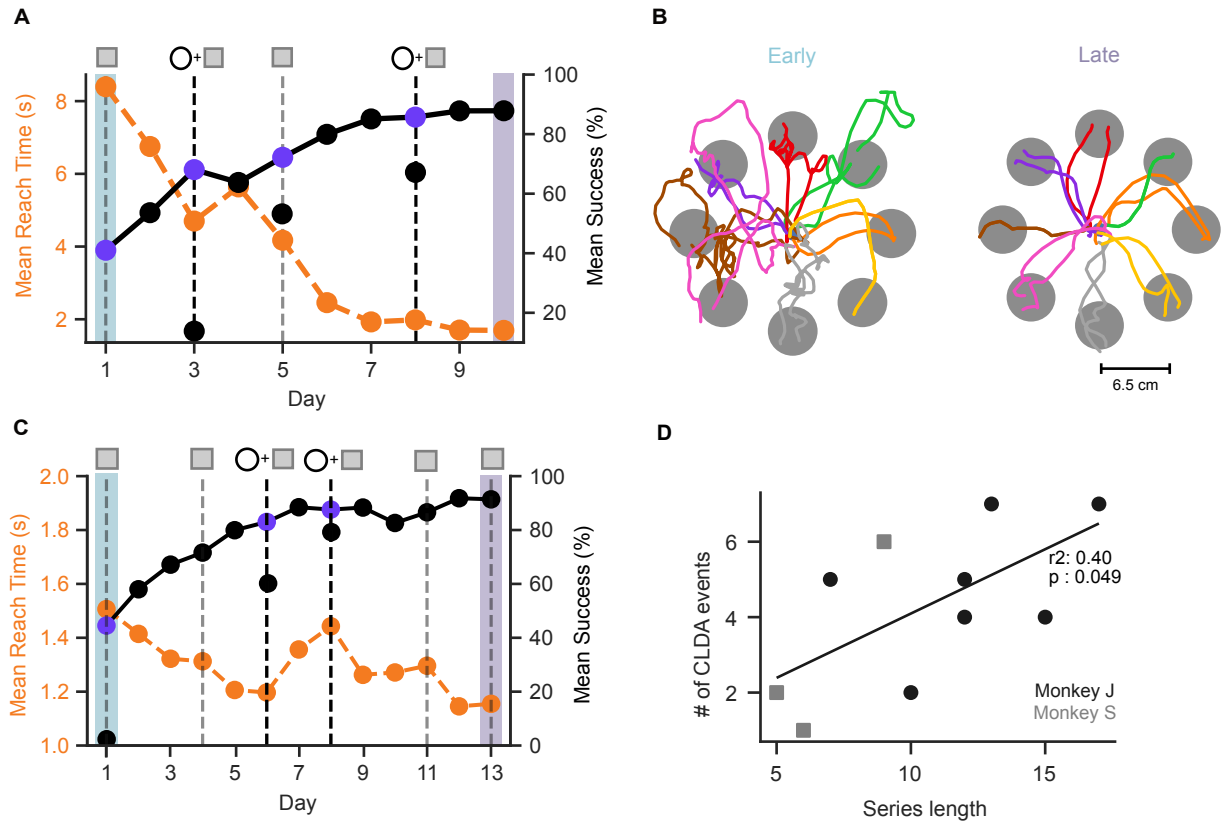

Figure S1. Supplementary figure related to figure 1: example behavior from monkey S and J (A) Success percentage (solid black line) and mean reach time (dashed orange line) across days for an example learning series for monkey S (seba010911_011811). Indigo dots show performance boost after decoder adaptation. Early (blue) and late (purple) training phases are indicated. Vertical dashed lines indicate days where decoder was adapted - gray dashed lines (weight change only) and black dashed lines (readout + weight change). **(B)** Cursor trajectories during early (left) and late (right) training phases for the example series in panel A. **(C)** Same as (A) for another example series from monkey J where decoder was adapted on day 1 for better performance. **(D)** Number of total decoder adaptation events (including both weight changes and readouts + weight changes) plotted against series length for monkey J (black circles) and monkey S (grey squares).

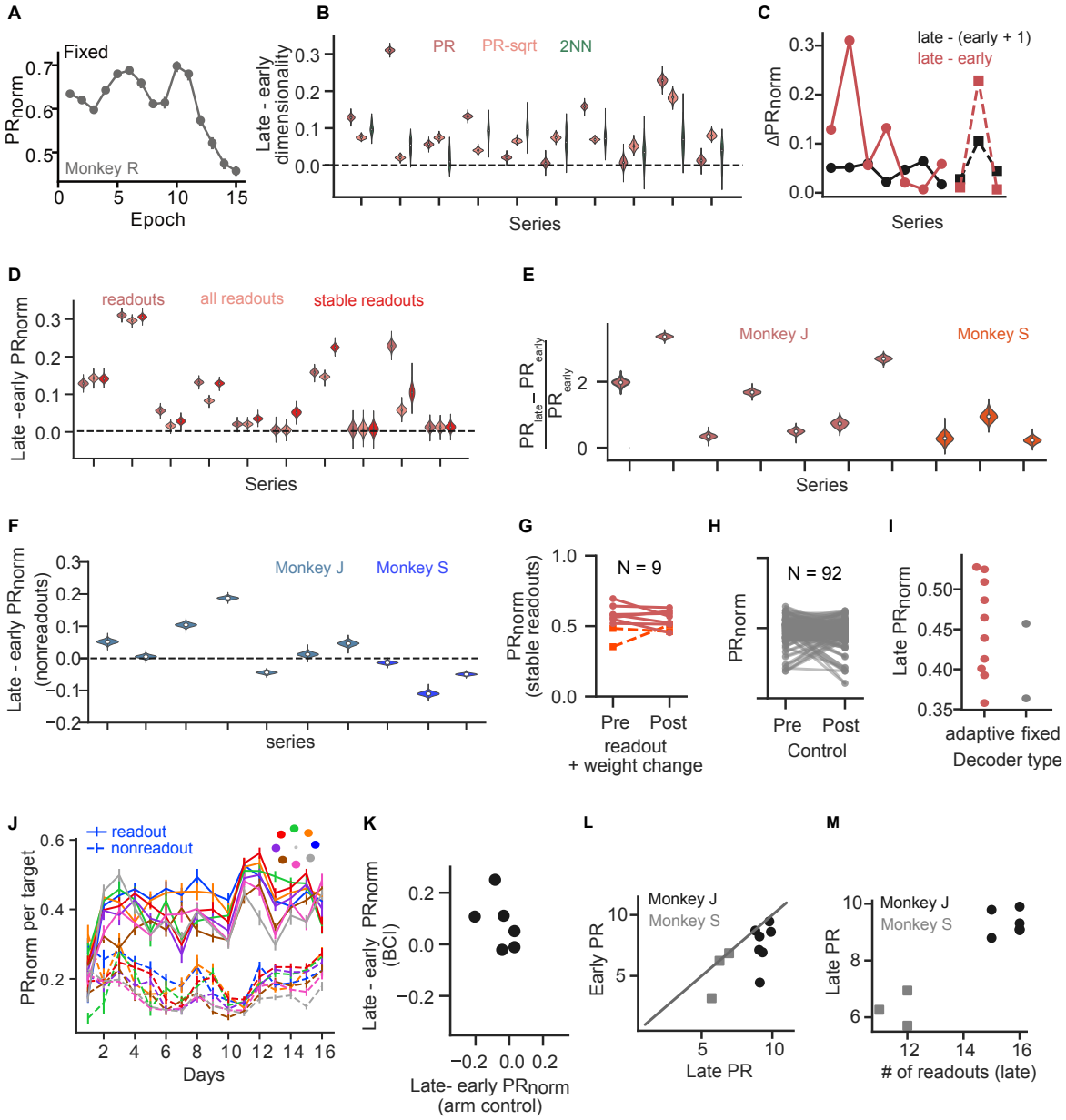

Figure S2. Supplementary figure related to figure 2: dimensionality analysis (A) Normalized participation ratio (PR_{norm}) for fixed decoders (monkey R; data from [2]). Note: the x-axis denotes epoch, which are groups of constant number of trials, consistent with previous analyses [3]. (B) Change in dimensionality of readout units between late and early computed using participation ratio (PR, dark red), participation ratio after applying square root transform (PR-sqrt, pink), two-nearest neighbors (2NN, green). Each distribution estimated using 1000 bootstraps. (C) Late − early PR_{norm} (red) and late − shifted early PR_{norm} (black) for learning series from monkey J (solid lines) and monkey S (dashed lines) (D) Late − early PR_{norm} for readouts used each day ('readouts'; dark red), for all units used as readouts in a learning series—including those that were added or removed—('all readouts'; pink), and for readouts that were stable across all days within a series ('stable readouts'; red). (E) Late − early PR normalized to early PR for readout populations. (F) Late − early PR_{norm} for nonreadout populations. (G) PR_{norm} computed using only stable readout units before and after decoder weight changes alongside readout unit changes across all series for both monkeys J and S ($p = 0.496 (n.s), N = 9$, pre < post, one-sided Wilcoxon signed-rank test). (H) PR_{norm} computed using all readouts for pairs of days where no decoder adaptation occurred ($p = 0.19 (n.s), N = 92$, pre < post, one-sided Wilcoxon signed-rank test). (I) Late PR_{norm} ranges for adaptive decoder series (monkey J and S, $N = 10$) and fixed decoder series (from [2], monkey P and R, $N = 2$). Statistical analyses to compare between fixed and adaptive decoders was not performed due to the small sample size available for fixed decoder studies. (J) PR_{norm} trends per target direction for readouts (solid lines) and nonreadouts (dashed lines) for representative series. (K) Late − early PR_{norm} for BCI control plotted against corresponding series's late - early PR_{norm} of arm control. (L) Unnormalized early PR vs late PR for BCI readouts in monkey J (black dots) and monkey S (grey squares). The grey line represents the identity. (M) Unnormalized late PR vs number of readouts used on late day In panels B,D, E and F, each distribution was estimated using 1000 bootstraps. The first 7 series are from monkey J and last three series are from monkey S.

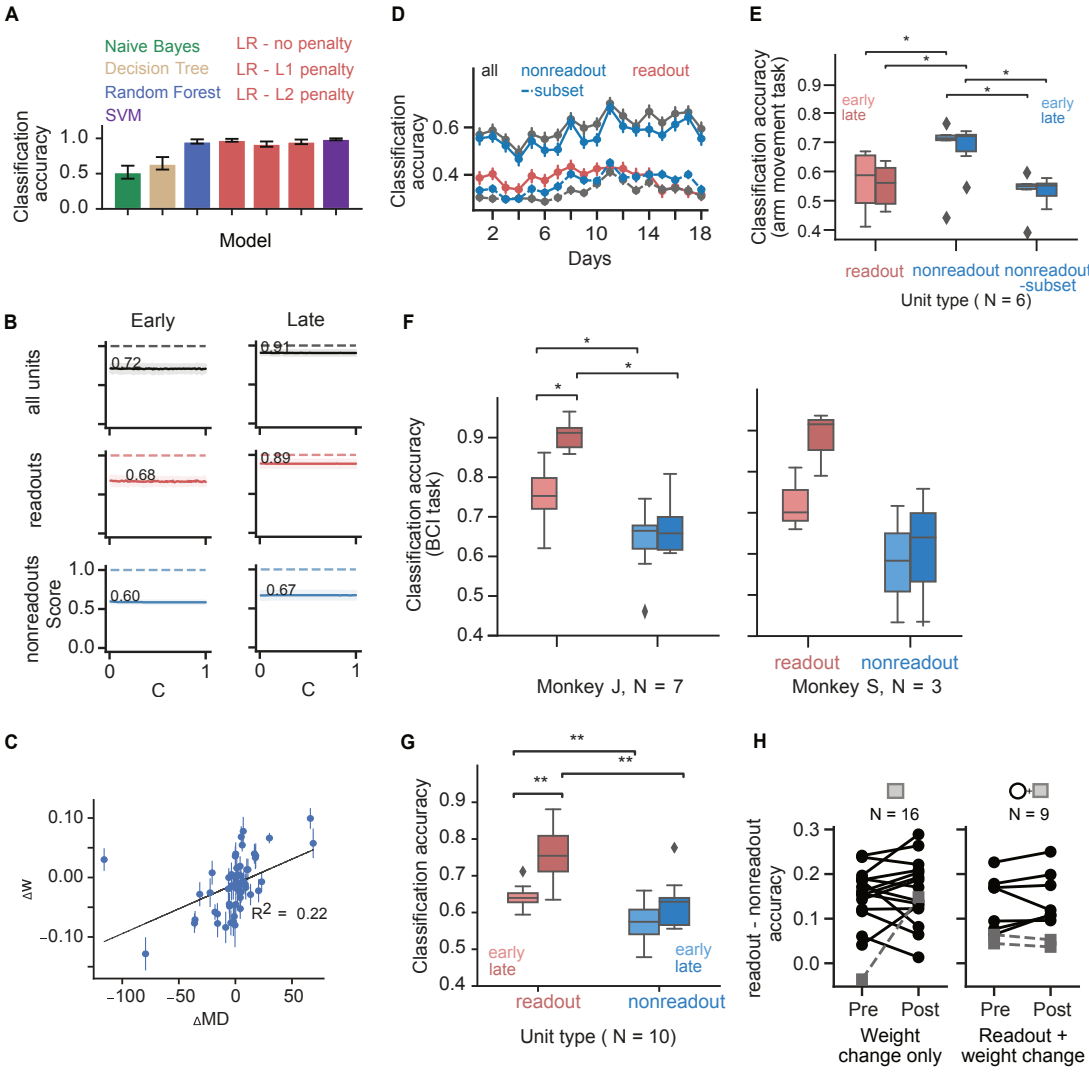

Figure S3. Supplementary figure related to figure 3: credit assignment (A) Classification accuracy obtained from different classifier models when trained on neural data from last day of our representative series. Error bars represent the 95% confidence interval on test accuracy. (B) Early and late classification scores for all units (black), readouts (red) and nonreadouts(blue) plotted as hyperparameter C (related to regularization strength) varies from 1e-5 to 1. (C) Change in logistic regression model coefficients ($\Delta \mathbf{w}$) between early and late day as a function of changes in modulation depth of units from representative series. (D) Classification analysis from all units (black), readouts (red) and nonreadouts (blue) when the animal performed an arm movement task for the representative series. There were significantly more units in the 'all' and 'nonreadout' populations compared to the 'readout' population, which contributes to the large difference in overall classification accuracy. Dashed blue and black lines show the classification accuracy achieved from a subset of nonreadout and all units, respectively, randomly drawn to match the number of readout units. (E) Early vs late classification accuracy for readout (shades of red) and non-readout populations(shades of blue) across all series for arm movement task (readout-early vs nonreadout-early : $p = 0.031(*)$;readout-late vs nonreadout-late : $p = 0.031(*)$; nonreadout-early vs nonreadout subset- early: $p = 0.031(*)$; nonreadout-late vs nonredout-late subset: $p = 0.031(*)$, readout-early vs nonreadout-subset early: $p = 0.21(n.s)$; readout-late vs nonreadout-subset late: $p = 0.84(n.s)$, $N = 6$, Wilcoxon signed-rank test) (F) Early vs late classification accuracy for readout (shades of red) and nonreadout populations (shades of blue) shown separately for each monkey for BCI task (monkey J ($N = 7$: nonreadout-early vs. nonreadout-late: $p = 0.29(n.s)$; readout-early vs. readout-late: $p = 0.0156(*)$; readout-late vs. nonreadout-late: $p = 0.0156(*)$; readout-early vs. nonreadout-early: $p = 0.0156(*)$; monkey S $N = 3$: : nonreadout-early vs. nonreadout-late: $p = 0.5(n.s)$; readout-early vs. readout-late: $p = 0.25(n.s)$; readout-late vs. nonreadout-late: $p = 0.25(n.s)$; readout-early vs. nonreadout-early: $p = 0.25(n.s)$, Wilcoxon signed-rank test) (G) Early vs late classification accuracy for readout (shades of red) and nonreadout populations (shades of blue) from BCI task using neural activity from the go cue to 200 ms after go cue. ($N = 10$: nonreadout-early vs. nonreadout-late: $p = 0.098(n.s)$; readout-early vs. readout-late: $p = 0.0078(**)$; readout-late vs. nonreadout-late: $p = 0.0039(**)$; readout-early vs. nonreadout-early: $p = 0.0078(**)$; Two-sided Wilcoxon signed-rank test) (H) Readout-nonreadout classification accuracy, grouped by the type of change (with (right) and without (left) readout unit changes) across all series and monkeys (Weight change only: p = 0.26 (ns), N = 16 ; readout + weight change: p = 0.25 (ns), N = 9, pre < post, one-sided Wilcoxon signed-rank test)).

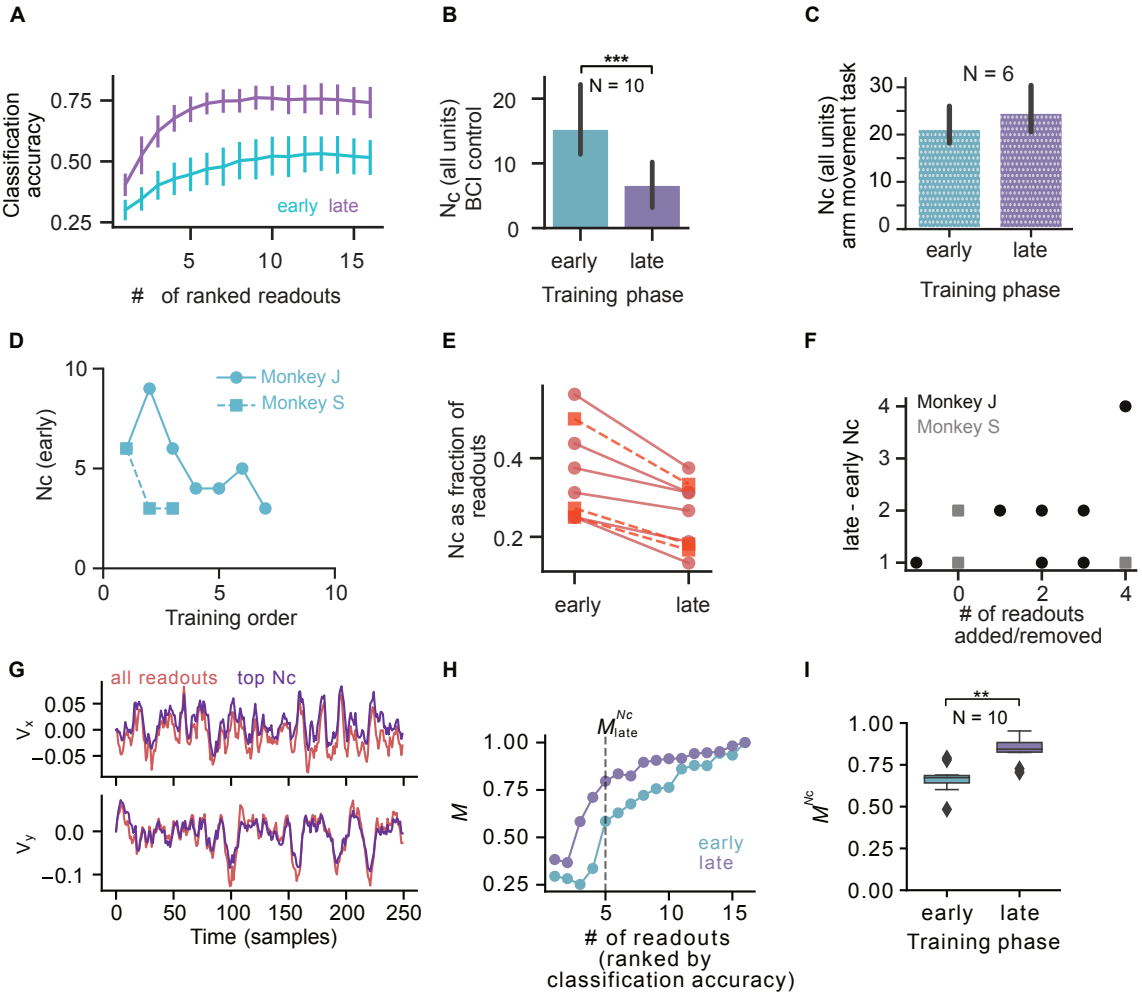

Figure S4. Supplementary figure related to figure 4: compactness (A) Classification accuracy (non-normalized) as units are added in ranked order (ranked NAC) on early (cyan) vs late (purple) day for the representative series (Unnormalized version of Fig 4C). (B) Comparison of N_c^{late} vs N_c^{early} across series while using all units (N_c^{late} is less than N_c^{early}, $p = 0.00098$, $N = 10$, Wilcoxon signed-rank test). (C) Similar to (B) for arm movement task ($p = 0.067 (n.s)$, $N = 6$, Wilcoxon signed-rank test) (D) N_c^{early} (number of units required to decoder 80% of maximum classification accuracy) for readouts early in training as a function of training order, for monkey J (solid lines with round markers) and monkey S (dashed lines with square markers). (E) N_c plotted as a fraction of readouts used for BCI control. (F) late -early N_c plotted as a function of the number of readouts added or removed in a learning series ($R^2 = 0.36$, $p = 0.30$ (n.s). (G) X and Y cursor velocities reconstructed offline using the Kalman filter decoder with all readouts (red) and the top N_c^{late} readout units (purple) identified by target encoding analysis for a representative series. (H) Matching (M) between the velocities reconstructed using all units vs. as units are added in ranked order (ranked NAC) on early (cyan) vs late (purple) day for the representative series. (I) Velocity matching (M^{N_c}) when using top Nc units on early vs late day ($p = 0.0019$ (**) , $N = 10$, Wilcoxon signed-rank test).

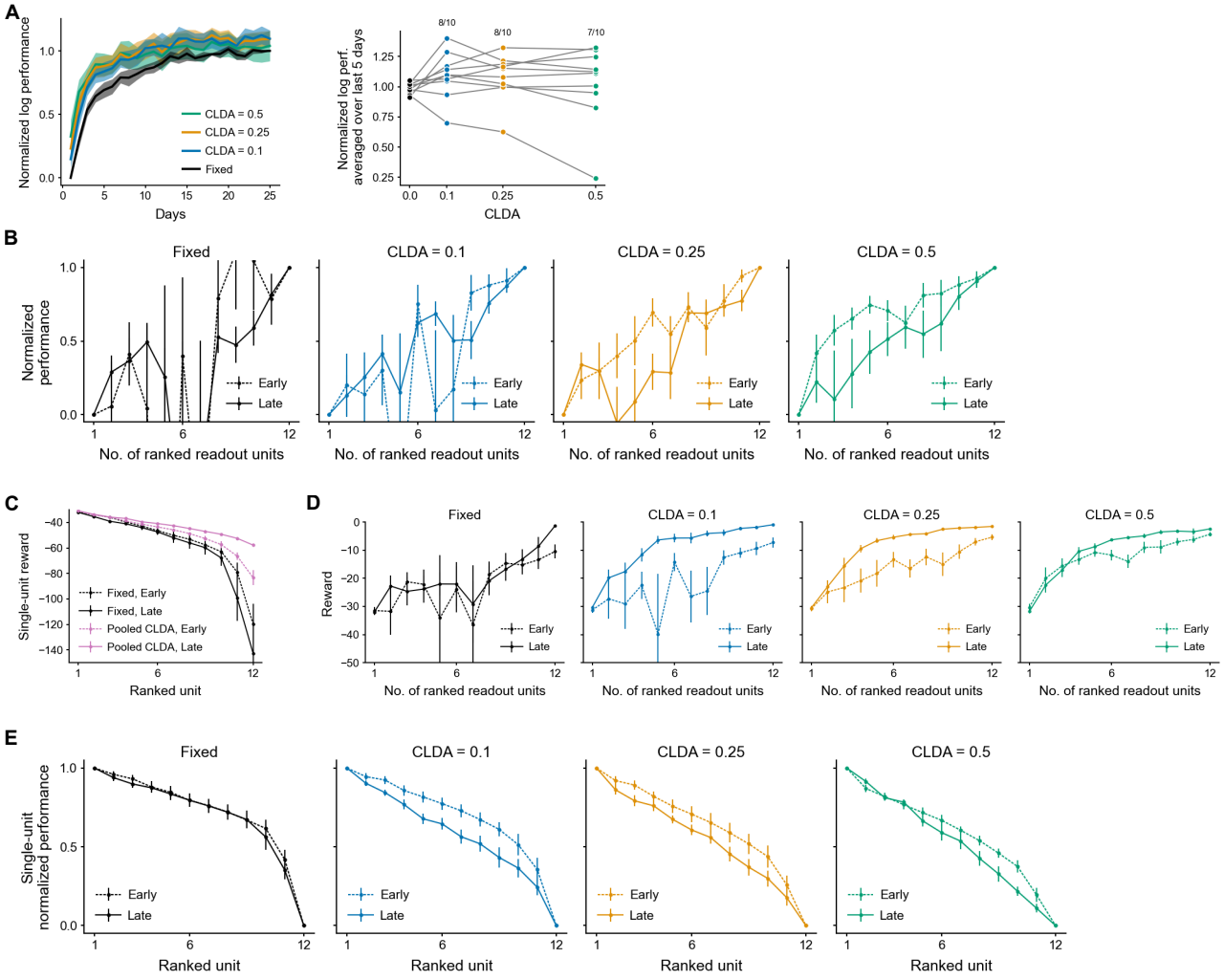

Figure S5. Supplementary figure related to Fig. 5: model. Training performance (A) and compactness (B) when stopping CLDA after a loss criterion was reached (i.e., when loss < 2). Compare with Figs. 5C-D and 5F in the main text. (C-E) Results obtained using the exact same protocol as in Fig. 5 in the main text (i.e., no stopping of CLDA). (C-D) Unnormalized single-unit performances (C) and unit-adding curves (D). Reward, which is the negative of the loss, as a function of number of ranked readout units. Corresponds to Fig. 5E-F in the main text. (E) Single-unit performances for each CLDA intensity. Un-pooled version of Fig. 5E in the main text.

References

[1] A. P. Georgopoulos et al. “On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex”. en. In: *Journal of Neuroscience* 2.11 (Nov. 1982). Publisher: Society for Neuroscience Section: Articles, pp. 1527–1537. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.02-11-01527.1982.

[2] Karunesh Ganguly and Jose M. Carmena. “Emergence of a Stable Cortical Map for Neuroprosthetic Control”. en. In: *PLOS Biology* 7.7 (July 2009). Publisher: Public Library of Science, e1000153. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000153.

[3] Vivek R. Athalye et al. “Emergence of Coordinated Neural Dynamics Underlies Neuroprosthetic Learning and Skillful Control”. In: *Neuron* 93.4 (Feb. 2017), 955–970.e5. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2017.01.016.