Deep learning-based location decoding reveals that across-day representational drift is better predicted by rewarded experience than time

Kipp Freud[1,*], Nathan Lepora[3], Matt W. Jones[2], and Cian O'Donnell[1,4]

[1]School of Computer Science, University of Bristol, Bristol, UK
[2]School of Physiology, Pharmacology & Neuroscience, University of Bristol, Bristol, UK
[3]School of Engineering Mathematics and Technology, University of Bristol, Bristol, UK
[4]School of Computing, Engineering & Intelligent Systems, Ulster University, Derry/Londonderry, UK
[*]corresponding author: km14740@bristol.ac.uk

ABSTRACT

Neural representations of space in the hippocampus and related brain areas change over timescales of days-weeks, even in familiar contexts and when behavior appears stable. It is unclear whether this 'representational drift' is primarily driven by the passage of time or by behavioral experience. Here we present a novel deep-learning approach for measuring network-level representational drift, quantifying drift as the rate of change in decoder error of deep neural networks as a function of train-test lag. Using this method, we analyse a longitudinal dataset of 0.5–475 Hz broadband local field potential (LFP) data recorded from dorsal hippocampal CA1, medial prefrontal cortex and parietal cortex of six rats over ∼30 days, during learning of a spatial navigation task in an unfamiliar environment. All three brain regions contained clear spatial representations which improve and drift over training sessions. We find that the rate of drift slows for later training sessions. Finally, we find that drift is statistically better explained by task-relevant rewarded experiences within the maze, rather than the passage of time or number of sessions the animal spent on the maze. Our use of deep neural networks to quantify drift in broadband neural time series unlocks new possibilities for testing which aspects of behavior drive representational drift.

Introduction

Neural activity patterns associated with sensation, cognition, and action are often not rigid but instead undergo progressive changes over timescales of minutes to weeks — a phenomenon called representational drift[1,2]. Such changes in the activity patterns of populations of neurons can occur even without apparent changes in environment or behaviour[3]. These findings challenge classical models which consider the stability of the engram as the basis of memory's persistence[4].

Representational drift is observed across different primary sensory modalities in the brain. For instance, neural representations in visual cortex exhibit drift in human, rat, and mouse[4–6], with evidence suggesting that deeper areas of the visual cortical hierarchy encode visual information more persistently[7]. Similarly, olfactory representations change over time, even when external stimuli remain constant[8,9]. Representational drift is also evident in multimodal spatial representations, which have been best characterized in hippocampus and interconnected cortical regions. Soon after initial exposure to a new environment, place fields develop in the activity of neurons in the hippocampus and related brain regions[10], but undergo drift both within- and across days[11–14].

The dynamics of drift have been quantified at the level single units, and vary with brain region. Linear classifiers trained to predict odorants on data from single units in the primary olfactory cortex degraded to chance levels after 32 days (though this could be slowed by daily exposure to the same odorant)[8]. In mouse hippocampal CA1, long-timescale drift occurs orthogonally to the representation of context in network space, allowing for consistent readout of contextual information across weeks[15]. One key question is whether drift is a function of experience or time? Khatib et al. (2023)[12] found that within-day drift of spatial representations in dorsal CA1 is driven primarily by active experience rather than time. However Geva et al. (2023)[16] examined drift across days, and found a more nuanced relationship — active experience in a maze drove changes in spatial tuning of dorsal CA1 cells, while time drove changes in the activity firing rates of cells. Importantly, both these studies equated 'experience' with 'time spent on the maze'. This leaves open the question of whether different aspects of the animal's task experience (time doing task, number of task trials, rewards) may have different impacts on representational drift[17].

One other key mechanism modulating hippocampal representations over extended experience is 'replay', the reactivation of place cell activity patterns during sharp-wave ripples (SWRs) during resting and sleep states. Such replay may be implicated

in representational drift as a byproduct of memories being integrated into new sets of neurons as memories are replayed[18]. Consistent with this, Grosmark et al. (2021) found that replay during SWRs was correlated with the long-term stability of spatial representations within single units in the hippocampus[19].

Here, we develop and test methods to disentangle the impacts of time, experience, and SWR-associated replay on representational drift tracked simultaneously across multiple brain regions and ethologically-relevant timescales of 2–3 weeks. We recorded local field potentials (LFPs) from chronically-implanted adjustable microelectrodes spanning three distinct brain regions of six rats; dorsal CA1 of hippocampus (CA1), prefrontal cortex (PFC), and the parietal cortex (PC).

Our approach[20] uses deep convolutional neural networks (CNNs) trained on wavelet decomposition images of this LFP data[21] to decode animal location in a multi-session longitudinal dataset.

This has two main advantages compared to spike-based analyses. First, most spike-based approaches consider firing activity between neurons to be independent, which is a problematic assumption for quantifying representational drift at the ensemble level. Although it is possible to construct spike-based decoding models that include pairwise correlations, fitting them to data is a computation-intensive challenge[22]. The CNN used in this approach contains fully connected layers, which can combine information between channels and electrodes to use information from non-independent activity, giving a more global insight into drift dynamics. Second, our approach requires only minimally-processed input and does not require spike sorting (a computationally-intensive process for detecting action potentials and assigning them to specific neurons[23]). Necessarily, spike sorting discards information in frequency bands outside of the spike range which potentially introduces biases implicit in the algorithm. In this way our approach can allow for a more expansive analysis of neuronal drift than the more classical spike based, single-unit approach[24]. Last, as our approach uses raw LFP data representing the aggregate neural activity at selected brain areas, we are not examining drift at the level of single units. This offers a complementary way of examining drift relative to spike-based electrophysiology or calcium imaging.

Results

Hippocampal, parietal and prefrontal local field potentials code spatial representations that become more informative with behavioural training

We recorded local field potential data from hippocampal CA1 (HPC), prefrontal cortex (PFC), and parietal cortex (PC) from 6 rats as they navigated a memory and decision-making task to find sucrose rewards on a 130 $\times$ 180cm maze for $\sim$30 days (see Figure 1A)[25,26]. Rats were trained to make choices between left and right maze arms, contingent upon the direction of an initial guided turn. During the initial phase of training, the rats were trained using a "match turn" rule, which required that their navigational choices matched their initial guided turn (i.e. if the rats were directed to turn right at the guided turn, they should then turn right again at the choice point). When the rats had achieved task competency (three sessions of at least 85% of trials correct) the rule was reversed to an "opposite turn" rule, where rats were rewarded for making choices that were opposite to their initial guided turn. During these experiments we also recorded rat location via head-mounted LEDs.

We first trained deep CNNs to decode rat maze location using 0.5–475 Hz wide-band LFP data collected from HPC, PFC, and PC, with one model being trained for each rat $\times$ experimental session $\times$ brain region combination. After this, we quantified location decoding performance on a testing set of unseen data that was withheld during training.All decoders performed significantly better than random, with decoders using HPC-only data obtaining an average error of 68.78cm (on a maze of size 164 $\times$ 118 cm), and PFC-only and PC-only decoders obtaining mean errors of 72.41cm and 73.15cm, respectively. Though this decoding error is higher than could be achieved using spiking data, we noted that the error significantly increased with train-test-lag (shown in Figure 2) for all brain regions considered, clearly showing drifting spatial representations being detected by our system. Because of these results, we considered this decoding accuracy to be sufficient for addressing our questions on representational drift.

Second, we trained deep CNNs to decode location using data from all brain regions at once, with one model being trained for each rat $\times$ experimental session combination. Using these models, we obtained an average error of 64.19 cm, significantly outperforming both random location decoding and each of the single-brain-region models.

For all considered rats, we noted a significant decrease in mean decoder error as experimental session increased during the first $\sim$10 experimental sessions of data collection, as shown in Figure 1e (regression coefficient -0.9782, $p = 0.021$; OLS regression). In contrast we did not detect significant further decreases in decoder error after the 10th training session.

To test whether low- vs high-frequency bands carried more location information, we also trained separate models using LFP data either above or below 300 Hz. The vast majority of location information was contained above 300 Hz in the data, with models trained on data in the lower frequency range performing only marginally better than random at same-day decoding. In contrast, models trained on only higher frequency bands $>$ 300 Hz performed almost as well models trained on all data.

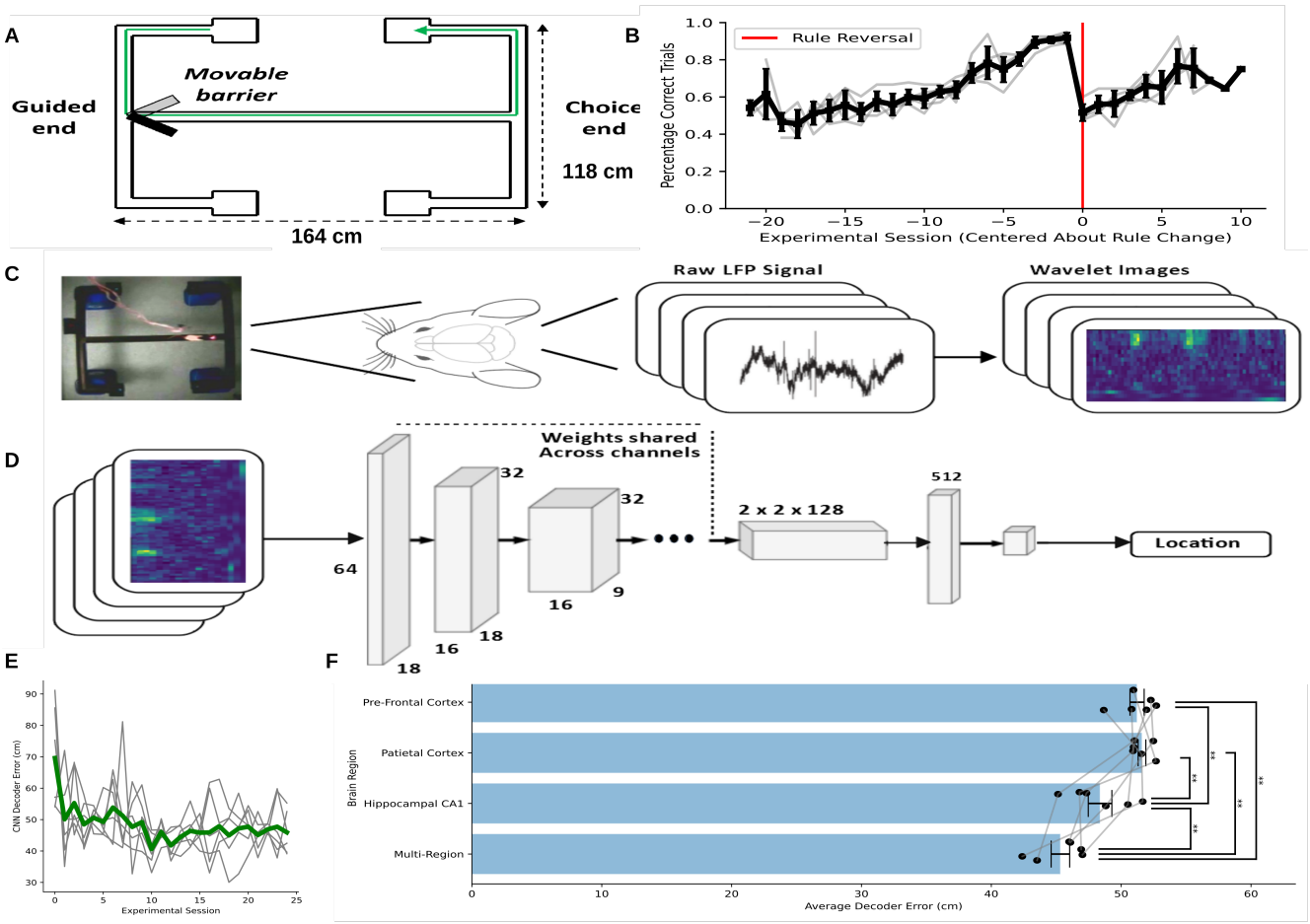

Figure 1. *(A) Rats undertook a memory and decision-making task to find sucrose rewards on a maze over ∼35 days.*
(B) Rats were initially trained under a 'match-turn' rule, but after task competency had been achieved, usually ∼25 days, this rule was reversed to a 'opposite-turn' rule.
(C) Local field potentials were collected from 16 chronically-implanted adjustable tetrodes spanning Hippocampal CA1, Prefrontal Cortex, and Perietal Cortex. These LFPs were transformed into Morelet wavelet images.
(D) Wavelet images from each experiemental session were used with a deep convolutional neural network to decode location.
(E) Decoder error decreased as training session number at which the data was collected increased.
(F) Neural networks which were trained on different brain regions obtain different average error, with models trained on Hippocampal CA1 data performing significantly better than models trained on data from other brain regions, and a multi-region model performing significantly better than every single region models ($p<0.05$, ANOVA with Tukeys correction).

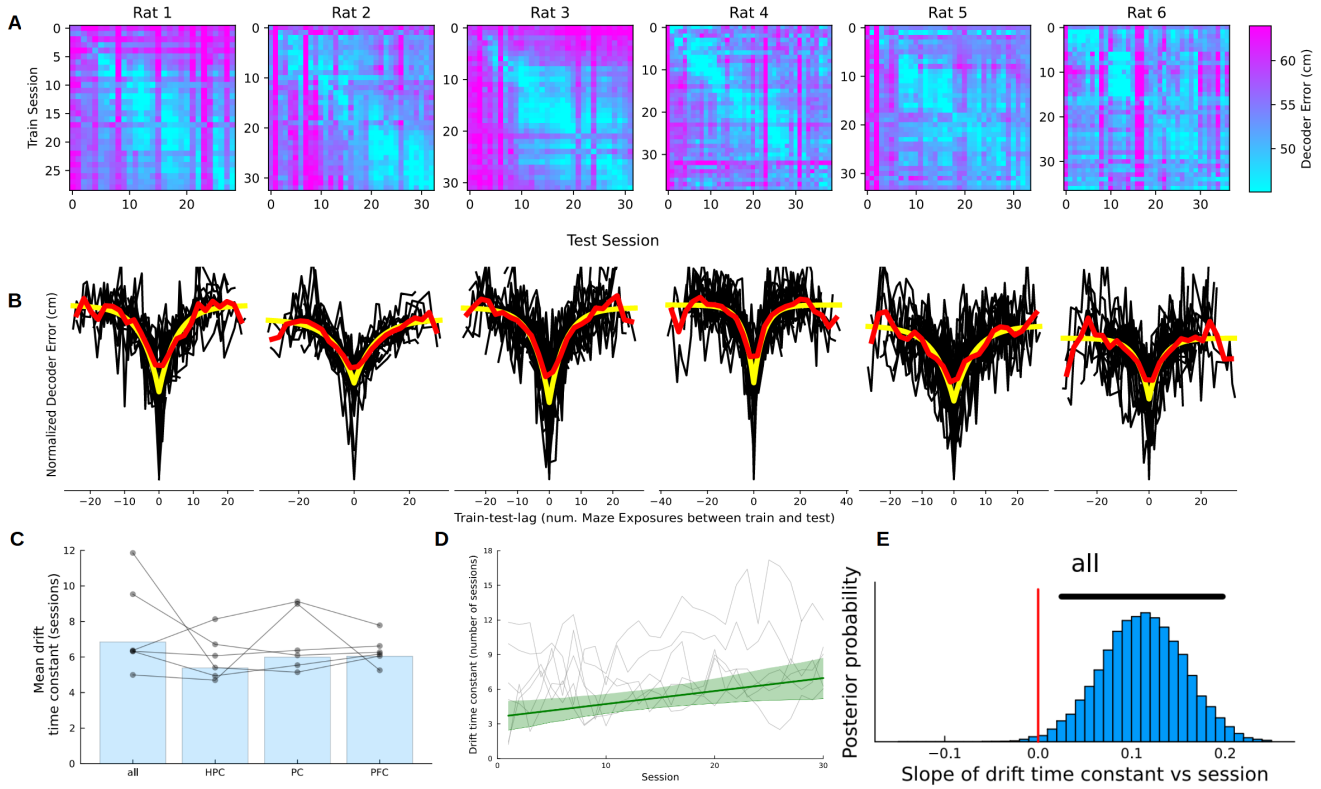

Figure 2. *(A) Accuracy table for all rats, with each subsequent row representing a later session of training data, and each subsequent column representing a later session of testing data. The diagonal represents error with zero train-test-lag i.e. when training and testing on different data collected from the same day.*
(B) Decoder error against train-test lag for the multi-region models of all rats; the mean values of these error curves is shown in red, and fitted exponential decay curves are also shown in yellow.
(C) Decay time constants for fitted exponential decay curves for models trained on different brain regions.
(D) Decay time constants over time for the multi-region models of all rats, with the upwards trend indicating increasing stability. Posterior mean linear fit plotted in dark green with 95% highest density interval in light green.
(E) Posterior probability of the slope parameter for a linear regression between drift time constant vs training session. Black bar indicates 95% highest-density interval.

Representations drift over time

Using results generated from both models trained using data from all considered brain regions (the "multi-region" models), as well as all single-region models, we observed three key findings.

First, we observed that average decoding error increased as train-test lag increased, for both all considered rats and all considered brain region models. This indicates statistically significant representational drift for all regions, even in the low frequency LFP data used ($p < 1 \times 10^{-5}$ for multi-region, CA1, PFC, and PC models).

Second, we observed that mean normalised accuracy was well described by exponential decay curves:

$$P(\Delta t) = P_\infty - (P_\infty - P_0)e^{-\Delta t/\tau}$$

where Δt represents the train-test lag, P_0 represents the performance of the decoder when train-test lag is zero, P_∞ represents the performance of the decoder at very long train-test lags, and τ is the decay constant - representing the train-test lag at which decoder error will rise $(1 - \frac{1}{e}) \approx 63\%$ of the way to its asymptotic value. When fitting these curves, we found mean decay time constants τ of between 5–12 sessions for each rat; τ values did not differ significantly by brain region (Figure 2C; p=0.445, ANOVA).

Representations become more stable over time

For models trained on multi-region data, we found an increase in drift time constants as training day increased for all rats (95% HDI for the slope parameter was [0.025,0.198], probability that the slope <0 was 0.0064) (Figure 2C,D). For single region models, weak evidence was found for an increase in drift time constants (95% HDIs for the slope parameter were [-0.034,0.15] for HPC, [-0.015,0.165] for PC, and [-0.022, 0.138] for PFC, probability that the slope <0 was 0.0725, 0.05295, 0.0984 respectively for HPC, PC and PFC).

These results indicate an increase in stability of spatial representations over time in all considered brain regions. As the rats were learning the task during recording, this result suggests that fast ensemble level drift occurs upon first exposure to a new context, decreasing to a background level as familiarity is gained, or as time from first exposure in increased.

Drift is a function of both replayed and active experiences rather than time

Prior to this section, drift curves such as those shown in Figure 2 have been ordered according to experiment number i.e. points at x = ± 1 come from trials where the train-test pair is separated by ± 1 experiment day. However, experiments were not conducted daily; occasionally single days or whole weekends were skipped. Additionally, not all days consisted of the same number of choice trials or guided trials, and the performance of rats varied by days. Further, a varying number of hippocampal SWR events were detected during rest sessions that occured between each experimental session. We can incorporate this kind of experimental data into our findings to examine to what extent the variability of drift can be explained by different experimental parameters.

To do this, we represent the train-test lag of our data as each of the following experimental parameters:

1. The number of experimental sessions between the given train-test pair
2. The number of days between the given train-test pair
3. The number of guided trials which occured between the given train-test pair
4. The number of incorrect choice trials which occured between the given train-test pair
5. The number of correct (rewarded) choice trials which occured between the given train-test pair
6. The total number of choice trials which occured between the given train-test pair
7. The number of sleeping SWRs which were detected between the given train-test pair

We fit exponential decay curves to these newly arranged data, and examine the error between fitted curves and observed data. For each rat, we divide the error by the mean error of all the fittings for that rat to get a measure of the effect of each fitting. As shown in Figure 3, we noted a significant reduction in fit error when organising by experiential variables when compared with organising by time and number of exposures to the maze. Among experience dependent variables, we obtained the highest reduction of error when ordering by the number of trials and number of correct trials. Notably, we obtain the largest reduction in fit error when ordering by the number of SWR events occurring between experiment sessions.

We fit exponential decay curves to these newly arranged data, and examine the error between fitted curves and observed data. For every rat, we normalized the error by dividing it by the average error from all the curve fittings for that particular rat, to account for some animals having generally higher or lower fit error. This gave us a metric to gauge the impact of each individual fitting. As depicted in Figure 3, a notable decline in fitting error was observed when data was sorted based on certain experiential factors, as opposed to sorting by time or simply the number of maze exposures. Among the variables reliant on experience, the most significant error reduction was achieved when error data was organised by the number of trials and the number of successful trials. Notably, the most substantial decrease in fitting error was observed when arranging data based on the number of sleeping SWR events that took place between experimental sessions.

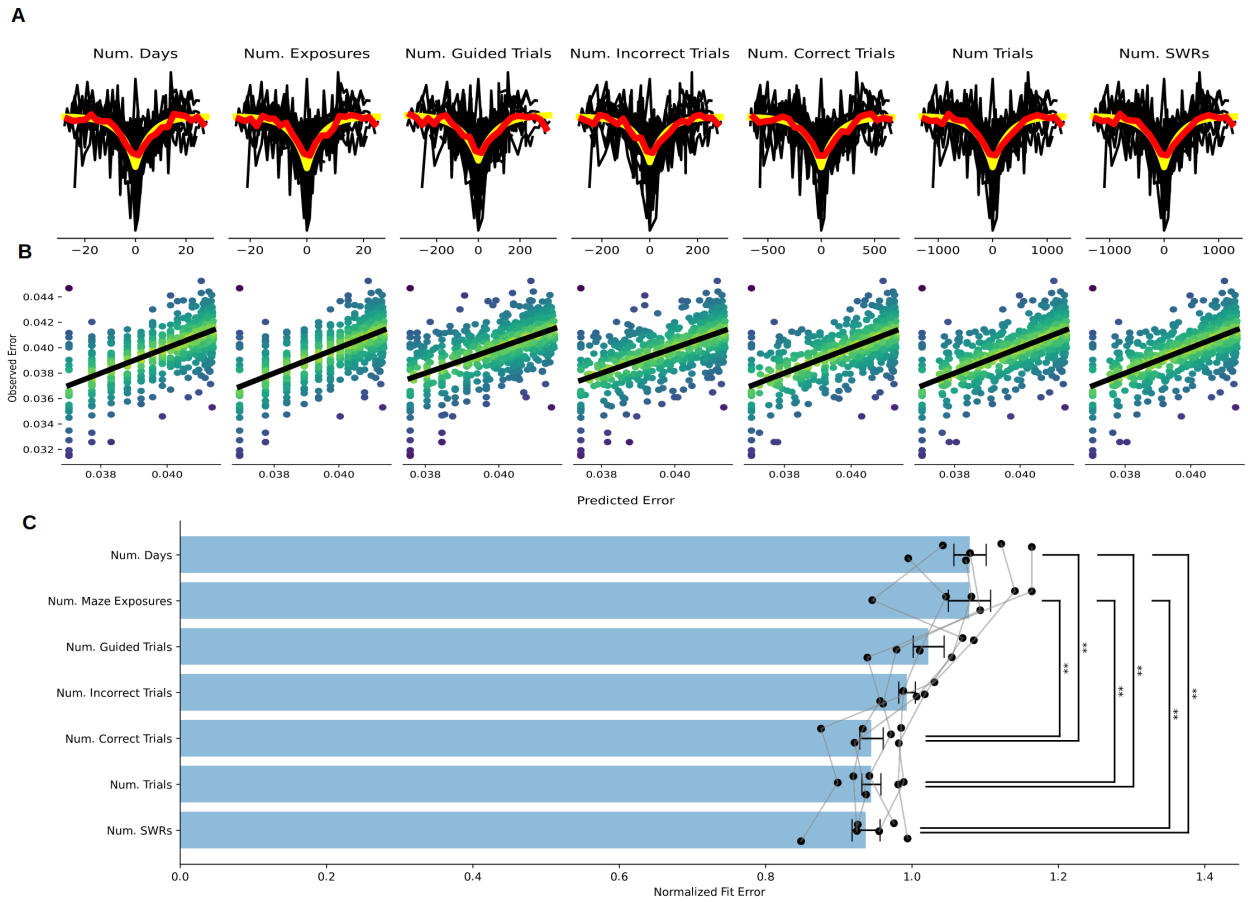

Figure 3. *(A) Decoder error against train-test lag for one rat, with lag being represented by different experience- or time-related experimental variables. The mean values of these error curves is shown in red, and fitted exponential decay curves are shown in yellow.*
(B) Expected vs. observed prediction error for each of the train-test-lag fits presented in A.
(C) Normalised error between observed error values and fitted exponential decay curves, when using each type of lag variable, normalized by the average of all lag variables. We note that when using certain experienced based variables as our train-test lag, we obtain significantly less fit error than when using time based lag, or using a simple exposure count based lag. We also that using the number of sharp wave ripples SWRs as our train test lag parameter results in the lowest fit error.

Discussion

We have presented a method for the quantification of drift which avoids spike sorting by using a deep convolutional neural network trained on wavelet decomposition images of LFP data (Figure 1). Using this approach, we find that across-day drift is well described by exponential decay curves (Figure 2). Moreover, we find that these curves are better able to fit the data when the train-test lag between two experimental sessions is measured in terms of experience-related variables, rather than time-related (Figure 3). Notably, this finding extends to replayed experiences as well as active experience, with the highest reduction in fit error coming when using the number of sleeping SWRs between train and test data as the train-test lag.

It is interesting to note that combined models using data from all brain regions perform significantly better than any single-region model (Figure 1F). This shows distributed information arising from the joint relationships between brain regions in the frequencies being considered[27,28], and also highlights a strength of our CNN based approach, which can discover these relationships in the data. In contrast, a simple naive Bayesian location decoder as often used for location decoding[29] assumes independence of neuron firing activity, and so cannot consider correlations in activity across brain region or neurons.

Early research into the dynamics of representational drift predominantly centered on examining the influence of time on drift dynamics[2,30]. More recent research has examined the contrast of the role of experience versus the passage of time in driving drift via *in vivo* calcium imaging of single neurons[12,16]. Although our results are consistent with these studies, they did not address which aspects of experience best predict drift. Here, we distinguished between both different aspects of active experience and replayed experience. Another recent study using two-photon calcium imaging of hippocampal neural populations in rats performing a navigational task found that drift was slow when reward expectation was high, and lowering the animals' reward expectation sped up representational drift[17]. Importantly, this study examined drift on a shorter 4–day time-scale than the ~ 30 days that we considered. Our study is nevertheless consistent with the results from[17] in the sense that both conclude that reward is a strong driver of drift dynamics.

Our research provides the first evidence that representational drift is driven more by reward related experiences than non reward related experience, and is the first to show a connection between representational drift and hippocampal SWRs.

Notably, the research presented here is novel in that we used LFP to measure drift. Though other research has been conducted which examined drift in local field potentials[31], this work was focused on obtaining stable low dimensional representations of LFP, rather than quantifying the drift dynamics. Other studies which have quantified drift as some measure of the change of spiking behaviors of single units[2,12,16,17]. Our use of LFPs is complementary to these studies, as it captures the aggregate electrical activity of populations of neurons within a localized region. This shift emphasizes the need to consider both localized, single-neuron insights and global, network-level dynamics in order to paint a fuller picture of the mechanisms and influences underlying drift in neural representations.

Our general framework for quantifying representational drift is versatile. By defining drift as a measure of the degradation of quality of a deep neural decoder trained on wavelet images, our exact method can be applied to any brain region, and any representation which can be decoded; calcium imaging data, for example, has been shown to be well suited to decoding using deep networks[21,32]. In this way our method can facilitate research into the drifting behaviours of a wide variety of neural representations and functionalities.

Frey et al. (2021)[21] found that the majority of location information in wide-band LFP lies between frequency bands 469 Hz and 663 Hz. We studied frequencies between 0.5 and 475 Hz, which only partially covered the range of informative frequencies identified by[21]. Further research could ask whether these high frequency components of LFP display different drift dynamics compared to the low-mid frequencies we examined.

In conclusion, the current study presents a novel approach to understanding the intricacies of representational drift. While time and experience have been recurrent themes in prior studies, our findings imply that specific kinds of experiences, especially reward-related and replayed experiences, are critical drivers of drift. Further, our approach using LFP data as a modality for analyzing drift offers a complementary and promising avenue for future research.

References

1. Driscoll, L. N., Duncker, L. & Harvey, C. D. Representational drift: Emerging theories for continual learning and experimental future directions. *Curr. Opin. Neurobiol.* **76**, 102609 (2022).

2. Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N. & Harvey, C. D. Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell* **170**, 986–999 (2017).

3. Micou, C. & O'Leary, T. Representational drift as a window into neural and behavioural plasticity. *Curr. Opin. Neurobiol.* **81**, 102746 (2023).

4. Deitch, D., Rubin, A. & Ziv, Y. Representational drift in the mouse visual cortex. *Curr. biology* **31**, 4327–4339 (2021).

5. Roth, Z. N. & Merriam, E. P. Representations in human primary visual cortex drift over time. *Nat. Commun.* **14**, 4422 (2023).
6. Marks, T. D. & Goard, M. J. Stimulus-dependent representational drift in primary visual cortex. *Nat. communications* **12**, 5169 (2021).
7. Piasini, E. *et al.* Temporal stability of stimulus representation increases along rodent visual cortical hierarchies. *Nat. communications* **12**, 4448 (2021).
8. Schoonover, C. E., Ohashi, S. N., Axel, R. & Fink, A. J. Representational drift in primary olfactory cortex. *Nature* **594**, 541–546 (2021).
9. Barwich, A.-S. & Severino, G. J. The wire is not the territory: Understanding representational drift in olfaction with dynamical systems theory. *Top. Cogn. Sci.* (2023).
10. O'Keefe, J. & Burgess, N. Geometric determinants of the place fields of hippocampal neurons. *Nature* **381**, 425–428 (1996).
11. Ziv, Y. *et al.* Long-term dynamics of ca1 hippocampal place codes. *Nat. neuroscience* **16**, 264–266 (2013).
12. Khatib, D. *et al.* Active experience, not time, determines within-day representational drift in dorsal ca1. *Neuron* (2023).
13. Sheintuch, L. *et al.* Tracking the same neurons across multiple days in ca2+ imaging data. *Cell reports* **21**, 1102–1115 (2017).
14. Mankin, E. A. *et al.* Neuronal code for extended time in the hippocampus. *Proc. Natl. Acad. Sci.* **109**, 19462–19467 (2012).
15. Keinath, A. T., Mosser, C.-A. & Brandon, M. P. The representation of context in mouse hippocampus is preserved despite neural drift. *Nat. communications* **13**, 2415 (2022).
16. Geva, N., Deitch, D., Rubin, A. & Ziv, Y. Time and experience differentially affect distinct aspects of hippocampal representational drift. *Neuron* (2023).
17. Krishnan, S. & Sheffield, M. E. Reward expectation reduces representational drift in the hippocampus. *bioRxiv* DOI: 10.1101/2023.12.21.572809 (2023). https://www.biorxiv.org/content/early/2023/12/22/2023.12.21.572809.full.pdf.
18. Káli, S. & Dayan, P. Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nat. neuroscience* **7**, 286–294 (2004).
19. Grosmark, A. D., Sparks, F. T., Davis, M. J. & Losonczy, A. Reactivation predicts the consolidation of unbiased long-term cognitive maps. *Nat. Neurosci.* **24**, 1574–1585 (2021).
20. Freud, K., Jones, M., Lepora, N. & O'Donnell, C. Deep learning-based decoding of spatial information from limbic-cortical local field potentials reveal drifting spatial representations with increasing stability. In *Conference on Cognitive Computational Neuroscience* (2023).
21. Frey, M. *et al.* Interpreting wide-band neural activity using convolutional neural networks. *Elife* **10**, e66551 (2021).
22. Posani, L., Cocco, S., Ježek, K. & Monasson, R. Functional connectivity models for decoding of spatial representations from hippocampal ca1 recordings. *J. Comput. Neurosci.* **43**, 17–33 (2017).
23. Lewicki, M. S. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Comput. Neural Syst.* **9**, R53 (1998).
24. Zhang, K., Ginzburg, I., McNaughton, B. L. & Sejnowski, T. J. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *J. neurophysiology* **79**, 1017–1044 (1998).
25. Lloyd, K., Becker, N., Jones, M. W. & Bogacz, R. Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats. *Front. computational neuroscience* **6**, 87 (2012).
26. Jones, M. W. & Wilson, M. A. Theta rhythms coordinate hippocampal–prefrontal interactions in a spatial memory task. *PLoS biology* **3**, e402 (2005).
27. Cona, G. & Scarpazza, C. Where is the "where" in the brain? a meta-analysis of neuroimaging studies on spatial cognition. *Hum. brain mapping* **40**, 1867–1886 (2019).
28. Bijsterbosch, J. D. *et al.* The relationship between spatial configuration and functional connectivity of brain regions. *elife* **7**, e32992 (2018).
29. Kloosterman, F., Layton, S. P., Chen, Z. & Wilson, M. A. Bayesian decoding using unsorted spikes in the rat hippocampus. *J. neurophysiology* **111**, 217–227 (2014).

30. Rule, M. E., O'Leary, T. & Harvey, C. D. Causes and consequences of representational drift. *Curr. opinion neurobiology* **58**, 141–147 (2019).

31. Pinotsis, D. A. & Miller, E. K. Beyond dimension reduction: Stable electric fields emerge from and allow representational drift. *NeuroImage* **253**, 119058 (2022).

32. Zhang, Y. *et al.* Decoding pixel-level image features from two-photon calcium signals of macaque visual cortex. *Neural Comput.* **34**, 1369–1397 (2022).

33. Kajikawa, Y. & Schroeder, C. E. How local is the local field potential? *Neuron* **72**, 847–858 (2011).

34. Lin, J. & Qu, L. Feature extraction based on morlet wavelet and its application for mechanical fault diagnosis. *J. sound vibration* **234**, 135–148 (2000).

35. Lee, G., Gommers, R., Waselewski, F., Wohlfahrt, K. & O'Leary, A. Pywavelets: A python package for wavelet analysis. *J. Open Source Softw.* **4**, 1237 (2019).

36. Pnevmatikakis, E. A. Analysis pipelines for calcium imaging data. *Curr. opinion neurobiology* **55**, 15–21 (2019).

37. Stosiek, C., Garaschuk, O., Holthoff, K. & Konnerth, A. In vivo two-photon calcium imaging of neuronal networks. *Proc. Natl. Acad. Sci.* **100**, 7319–7324 (2003).

38. Reddi, S. J., Kale, S. & Kumar, S. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237* (2019).

39. Moré, J. J. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis: proceedings of the biennial Conference held at Dundee, June 28–July 1, 1977*, 105–116 (Springer, 2006).

40. Ge, H., Xu, K. & Ghahramani, Z. Turing: a language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, 1682–1690 (2018).

41. St, L., Wold, S. *et al.* Analysis of variance (anova). *Chemom. intelligent laboratory systems* **6**, 259–272 (1989).

42. Abdi, H. & Williams, L. J. Tukey's honestly significant difference (hsd) test. *Encycl. research design* **3**, 1–5 (2010).

STAR Methods

Data collection

Data was collected from 6 adult, male Long-Evans rats navigating a memory and decision-making task to find sucrose rewards on a 130 × 180cm maze for ∼30 days[25,26]. All procedures were performed in accordance with the UK Animals Scientific Procedures Act (1986) and approved by the University of Bristol Animal Welfare and Ethical Review Board.

The data collection spanned approximately 30 days, and the rats had minimal prior exposure to the maze, having been introduced to it for less than 1 day before the initial recording. The experimental design required the rats to make choices between left and right maze arms, contingent upon the direction of an initial guided turn. During the initial phase of training, the rats were trained using a "match turn" rule, which required that their choices at the choice point match their initial guided turn (i.e. if the rats were directed to turn right at the guided turn, they should then turn right again at the choice point). When the rats had achieved task competency (three sessions of at least 85% of trials correct), which typically occurred after approximately 20-25 days of training, the rule was reversed to an "opposite turn" rule, where rats were expected to make choices that were opposite to their initial guided turn (e.g. if they were initially directed right, they were required to turn left at the choice point).

The experimental sessions each day were divided into two distinct phases. The first phase encompassed a series of guided trials, where the rats were guided at both ends of the maze to move in the correct direction using a barrier blocking the incorrect-choice arm. The number of this type of trial varied on a daily basis. Subsequently, the rats underwent a variable number of trials per day, where they were free to make their own decisions at the choice points. Each session, the total number of trials (including guided trials) undertaken by each rat and the proportion of those trials that resulted in successful outcomes was recorded.

During data collection, 16 chronically-implanted adjustable tetrodes were placed in three distinct brain regions of each rat. These regions included the dorsal CA1 of the hippocampus (CA1), the prefrontal cortex (PFC), and the parietal cortex (PC). Local field potentials (LFPs) originating from each tetrode were bandpass filtered within the frequency range of 0.5 to 475 Hz. These LFPs represent the aggregated activity of neuronal populations and synapses proximate to the tetrode tip. Additionally, the animals' location was tracked using head-mounted LED sampled at 25 Hz.

LFPs were band-pass filtered at 0.5-475Hz and reflect a spatially weighted aggregate activity of populations of neurons and synapses near the electrode tips[33]. LFP data was then processed into wavelet decomposition images using a Morelet wavelet decomposition[34] using the PyWavelets package[35], and wavelet frequencies were downsampled to a rate of 30Hz to discard transient non-local representations. Between experimental sessions, rats were placed in a separate sleeping environment, and hippocampal sharp wave ripples (SWRs) were detected using tetrodes within CA1.

Convolutional Neural Network Decoding

In[21], a generalizable deep learning framework is presented for decoding sensory and behavioural variables directly from wide-band neural data. The approach requires little user input and generalizes across stimuli, behaviours, brain regions, and recording techniques. This framework has been shown to achieve state of the art results on neural localization tasks using Morlet wavelet decompositions of LFP recordings taken from the CA1 pyramidal cell layer in the hippocampus of freely foraging rodents, and is also capable of accurately decoding head direction and speed. We note that this approach to decoding has been shown to be versatile to multiple forms of input including calcium imaging data[36], thus it is natural to assume that the method presented in this paper would also be versatile to these other modalities.

A modified version of the architecture originally presented in[21] is used here. Our TensorFlow 2.10.1 implementation uses deep convolutional networks with 13 downsampling time-distributed (i.e. with shared weights across channels) layers followed by two fully connected layers for each output. A kernel size of 3 was used throughout the model, and the number of filters was kept constant at 64 for the first 10 layers while sharing weights across the channel dimension, then doubling the number of filters for the following 3 layers. For downsampling the input, we used a stride of 2 intermixed between the time and frequency dimension. We used 2D-convolutions which share weights across the channel dimension for the first 10 convolutions and across the time dimension for the last 3 convolutional layers. Sharing weights across channels prevents overfitting to channel-specific features and thus improves generalisation by making sure there exists a global representation of important features e.g. of spikes or other prominent oscillations in the local field potential. The full architecture of the network used is presented in Table 1.

Notably, as this system takes minimally-processed input, it is able to perform sensory decoding without spike sorting (a computationally-intensive process for detecting action potentials and assigning them to specific neurons[23]). Necessarily, spike sorting discards information in frequency bands outside of the spike range which potentially introduces biases implicit in the algorithm. Also, as the system is versatile to different forms of input, it is natural to think that the SLAM system presented here may also be extendable to non-LFP forms of neural input, for example two-photon calcium imaging data[37].

Separate networks were trained for each rat using the training portion of the data generated from a single maze task session. Networks were trained to decode location for 1000 epochs with 250 training steps per epoch. A batch size of 8 was used, as was an AMSGrad optimizer[38] with a learning rate of $7e^{-4}$.

Table 1. Layer by layer architecture of the convolutional model. Note that for the first 10 layers, weights were shared across the channel dimension, while in layers 11-13, weights are shared across the time dimension. We also note than layers 14-15 exist for each output head (i.e distinct outputs have distinct fully connected layers). Order of dimensions: Channels, Time, Frequency.

Layer	Layer Type	Input Dimension	Kernel Size	Stride	Filters
1	2D-Convolution	(16, 64, 18)	(1,3,3)	(1,2,1)	64
2	2D-Convolution	(16, 32, 18)	(1,3,3)	(1,1,2)	64
3	2D-Convolution	(16, 32, 9)	(1,3,3)	(1,2,1)	64
4	2D-Convolution	(16, 16, 9)	(1,3,3)	(1,1,2)	64
5	2D-Convolution	(16, 16, 5)	(1,3,3)	(1,2,1)	64
6	2D-Convolution	(16, 8, 5)	(1,3,3)	(1,1,2)	64
7	2D-Convolution	(16, 8, 3)	(1,3,3)	(1,2,1)	64
8	2D-Convolution	(16, 4, 3)	(1,3,3)	(1,1,2)	64
9	2D-Convolution	(16, 4, 2)	(1,3,3)	(1,2,1)	64
10	2D-Convolution	(16, 2, 2)	(1,3,3)	(1,1,2)	64
11	2D-Convolution	(16, 2, 1)	(3,3,1)	(2,1,1)	128
12	2D-Convolution	(8, 2, 1)	(3,3,1)	(2,1,1)	128
13	2D-Convolution	(4, 2, 1)	(3,3,1)	(2,1,1)	128
14	Fully Connected	(1, 512)	-	-	1024
15	Fully Connected	(1, 1024)	-	-	1/2

Drift Quantification

To quantify drift, we first trained the deep convolutional neural networks (CNNs) described above to decode position from wavelet decomposition images of local field potentials (see Figure1), with distinct networks trained using data from different experimental session.

For each experimental session, separate networks were trained for each rat using data from all brain regions measured (16 tetrodes), as well as three separate networks for each rat trained on data only from CA1 (5 tetrodes), PFC (5 tetrodes) and PC (6 tetrodes). In this way we can examine both local as well as distributed representation dynamics.

Once networks are trained, accuracy data is gathered by calculating the error for networks trained for each day of data and each rat on all other days, generating a table containing error for every train-day/test-day pairing. By calculating decoding accuracy of CNNs trained on data from different days we quantify representational stability of spatial encodings in these regions as a function of train-test-lag. The idea here is that the better a network trained on session i performs on session $i+1$, the more similarities there will be between the spatial representation used by the animal on those days. The end result is, for 30 days of collected data, a 30×30 accuracy table containing the accuracy scores for all train-test day pairings (see Figure 2).

In order to compare drift rates across animals, and across training and test datasets for the same animal, we normalized the accuracy tables by row (dividing each row's elements by their sum) to account for model-to-model variability, and then normalized by column to account for test dataset variability.

As our method for drift quantification method is based on the change of decoding accuracy as a function of time, our method can be applied to a variety of sensory or behavioural variables. Drift can be quantified for any representation in any brain region, so long as decoding is possible.

Drift time constant estimation

The mean normalised accuracy of the tables generated above were well described by exponential decay curves as a function of train-test lag:

$$P(\Delta t) = P_\infty - (P_\infty - P_0)e^{-\Delta t/\tau}$$

where Δt is the train-test lag, P_0 is the performance of the decoder when training and testing is on data from the same session, P_∞ is the aysymptotic performance of the decoder at very long train-test lags, and τ is the drift time constant — the train-test lag at which decoder error will rise by $1/e \approx 63\%$ of the way towards its asymptotic performance.

When pooling the data across training sessions (Figures 2B and 3, we fit the parameters of this function by minimising the squared distance between observed accuracy values and curves using the Levenberg-Marquardt algorithm[39] in the `SciPy` 1.10.0 package.

However we found that this algorithm had numerical instabilities when fitting a different time constant per training session (Figure2D). To handle this we used a Bayesian model to perform the session-by-session regressions, which allowed for regularisation of the permitted time constant values. We used the following model:

$$\begin{aligned}
\bar{P}_0 &\sim \text{Normal}(0.5, 0.1) \\
\bar{P}_\infty &\sim \text{Normal}(0.6, 0.1) \\
\bar{\tau} &\sim \text{TruncatedNormal}(5, 5, 1, 30) \\
P_{0i} &\sim \text{Normal}(\bar{P}_0, 0.2) \\
P_{\infty i} &\sim \text{TruncatedNormal}(\bar{P}_\infty, 0.05, 0.5, 0.75) \\
\tau_i &\sim \text{TruncatedNormal}(\bar{\tau}, 5, 1., 30) \\
\sigma &\sim \text{TruncatedNormal}(0.1, 0.05, 0.01, 0.3) \\
y(i, \Delta t) &\sim \text{Normal}(P_{\infty i} - (P_{\infty i} - P_{0i})e^{-\Delta t/\tau_i}, \sigma)
\end{aligned}$$

where the bar over the symbol $\bar{P}_0$, $\bar{P}_\infty$ and $\bar{\tau}$ denotes that they are grand means across sessions, i indexes the training sessions, σ is the error standard deviation and $y(i, \Delta t)$ are the actual decoder performance values which here we treat as 'data' to be fit. This model performs regularisation via partial pooling of the P_{0i}, $P_{\infty i}$ and τ_i parameters across sessions.

We also used a Bayesian model to perform linear regression on the time constant as a function of training session, pooled across all animals (Figure 2D). Bayesian models were fit using the NUTS Hamiltonian monte carlo algorithm in the `Turing.jl` package[40] in Julia.

Statistical Testing

ANOVA with Tukeys HSD

Statistical testing was undertaken to examine differences between the location decoding error of models trained on data from different brain-regions, between τ values generated from these different brain-region models, and between the fit errors of accuracy data ordered by different experimental variables.

ANOVA[41] was selected as the most appropriate analytical method because it allows for the comparison of the means of three or more independent groups, providing a statistical test of whether or not the means of several groups are all equal. Tukey's

Honestly Significant Difference (HSD) test[42] was employed as our post hoc analysis method. Tukey's test is particularly advantageous because it controls for Type I error across all pairwise comparisons, maintaining the overall error rate at the desired significance level. These tests were conducted using the `pairwise_tukeyhsd` function within the `statsmodels` 0.13.5 package.

Ordinary Least Squares Regression

We used ordinary least squares (OLS) regression to obtain best fit linear equations of the location decoder error against experimental session, and of the fitted τ values against experimental session. Specifically, OLS minimizes the sum of squared deviations from the actual data points, providing a best-fit linear equation. Using this, we obtained p-values of the slope coefficients to determine statistical significance of the gradient differing from zero. Statistical significance was assessed at the standard $p < 0.05$ threshold.

These tests were conducted using the `OLS` module within the `statsmodels` 0.13.5 package.

Bayesian analysis significance

For the Bayesian analysis of drift time constant as a function of training session (Figure2C–E), statistical significance was reported as whether the 95% Highest Density Interval for the slope parameter crossed zero (Figure 2E). In the text we also report the probability that the slope was below zero, calculated as the fraction of monte carlo samples from the posterior distribution that were negative. Small values for this probability indicate the slope is highly likely to be positive.

Acknowledgements

We thank the EPSRC, BBSRC (BB/G006687/1,BB/W001845/1), Wellcome Trust (202810/Z/16/Z), MRC (MR/S026630/1) and Leverhulme Trust (RPG-2019-229) for support. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.