DO MICE GROK? GLIMPSES OF HIDDEN PROGRESS IN SENSORY CORTEX

Tanishq Kumar* **Blake Bordelon** **Cengiz Pehlevan** **Venkatesh N. Murthy** **Samuel J. Gershman**
Harvard University

ABSTRACT

Does learning of task-relevant representations stop when behavior stops changing? Motivated by recent work in machine learning and the intuitive observation that human experts continue to learn after mastery, we hypothesize that task-specific representation learning in cortex can continue, even when behavior saturates. In a novel reanalysis of recently published neural data, we find evidence for such learning in posterior piriform cortex of mice following continued training on a task, long after behavior saturates at near-ceiling performance ("overtraining"). We demonstrate that class representations in cortex continue to separate during overtraining, so that examples that were incorrectly classified at the beginning of overtraining can abruptly be correctly classified later on, despite no changes in behavior during that time. We hypothesize this hidden learning takes the form of approximate margin maximization; we validate this and other predictions in the neural data, as well as build and interpret a simple synthetic model that recapitulates these phenomena. We conclude by demonstrating how this model of late-time feature learning implies an explanation for the empirical puzzle of overtraining reversal in animal learning, where task-specific representations are more robust to particular task changes because the learned features can be reused.

1 INTRODUCTION

Understanding how representations evolve during task learning is a fundamental question in both neuroscience and machine learning. While the dynamics of learning are standard objects of study in both fields, the representational dynamics at play long *after* ceiling performance is reached on the training task are much less commonly studied.

In deep learning, this has changed with the discovery of the grokking phenomenon (Power et al., 2022), whereby neural networks first fit their training data to high accuracy, then many epochs of training later ("overtraining"), abruptly generalize. As a phenomenon that epitomizes how neural networks can behave in sharp and unexpected ways, it has been under intense study (Nanda et al., 2023; Liu et al., 2022b; Varma et al., 2023). Emerging theories posit that generalization occurs due to delayed feature learning taking place when the network has already achieved ceiling accuracy on the training task (Kumar et al., 2023; Lyu et al., 2023).

In psychology, there exists a long line of work on learning tasks at all levels of expertise, including after task mastery (Ericsson et al., 1993; Newell & Rosenbloom, 1981; Fitts & Posner, 1967; Mackintosh, 1969; Richman et al., 1972; Mead, 1973). Many of these works attempt to address the intuitive observation that human experts in a variety of domains continue to learn on tasks at which they have achieved near-ceiling performance. There has also been much work on using deep learning to model perceptual learning tasks, to study the degree to which biological and artificial neural networks exhibit similar behavior on this class of tasks (Wenliang & Seitz, 2018; Bakhtiari, 2019; Yashar & Denison, 2017).

The nature of representational changes at play during overtraining on learned tasks have been tackled at a fine grained (theorems proven about certain classes of deep networks) and coarse grained (qualitative explanations in psychology) levels. Such an understanding, however, is broadly missing

*Correspondence to tkumar@college.harvard.edu.

at an intermediate level: that of experimental neuroscience. This is because there are several difficulties in naively testing the hypothesis that representations continue to evolve during overtraining in cortex. First, for many animals, recording neurons is invasive and risks damaging the animals, so recordings are typically only taken at the end of training to study the end-time learned representation (Yuste, 2015; Kim et al., 2016). Second, experiments in systems neuroscience do not typically design stimuli as "training" and "test" in the way datasets are constructed in deep learning, so making claims about learning and generalization is not possible. We elaborate on further difficulties, as well as the connection of our phenomenon of interest to the related phenomenon of representational drift, in Appendix A.3.

Here, we revisit the neural data from a recent work with precisely the desirable qualities above: (Berners-Lee et al., 2023), who study neural activity during learned odor discrimination task in mouse piriform cortex. The mice are trained to discriminate one target odor from hundreds of nontarget odors, then continue to be trained for weeks on the same task after reaching behavioral mastery ("overtraining"), with neural firing rate data recorded during this time. After overtraining, they are given held-out test examples. Our starting point in this work are two empirical observations in (Berners-Lee et al., 2023) that we seek to model and explain.

- Decoding accuracy of training labels from population activity increases throughout overtraining.
- Mice overtrained for longer tend to do better on held-out test examples after the overtraining period is complete.

Our goal in this work is to understand the representational dynamics in cortex at play underlying these observations, and to what extent, if any, they are similar to those at play in deep neural networks where learning continues after training accuracy reaches ceiling (Power et al., 2022). Our contributions are the following:

- We find that target-nontarget odor class representations in mouse piriform cortex continue actively separating during overtraining while behavior remains unchanged, loosely resembling the "hidden progress" (Barak et al., 2022; Nanda et al., 2023) that often taking place in synthetic deep learning settings.
- We find that the margin of the maximum margin classifier at each day of overtraining is increasing in mouse cortex, implying that points (odor trials) sufficiently close to the optimal decision boundary may be classified incorrectly at the beginning of overtraining, but correctly at the end, despite no outward change in behavior.
- We construct a synthetic model of piriform cortex performing this task exhibits these properties, as well as the grokking phenomenology. We interpret the resulting dynamics and use targeted ablations to trace the continued learning to margin maximization.
- We use our insight to suggest a new, fine-grained and neural explanation for the overtraining reversal effect, an empirical puzzle from experimental psychology in which animals overtrained on a task learn its reversal more quickly than those not.

We also note that Berners-Lee et al. (2023) are not the first to observe such effects experimentally: similar effects in sharpening of already-learned odor representations was also observed by Shakhawat et al. (2014); Kadohisa & Wilson (2006), so that the phenomena we seek to model in this work are likely to be robust at least in the setting of mouse olfaction.

2 Related Work

Lazy/rich learning and margin maximization. We cast our ideas in the language of "lazy" and "rich" regimes of representation learning, introduced in machine learning (Chizat et al., 2019; Jacot et al., 2018). Such characterization of learning regimes has become increasingly popular in computational neuroscience (Chizat et al., 2019; Farrell et al., 2023; Flesch et al., 2021; 2022; Ito & Murray, 2023). In machine learning, lazy learning refers to a trained neural network being well approximated by a kernel method in its initial neural tangent kernel (NTK) and the rich regime is when this approximation breaks down as the network learns task-relevant features and its weights

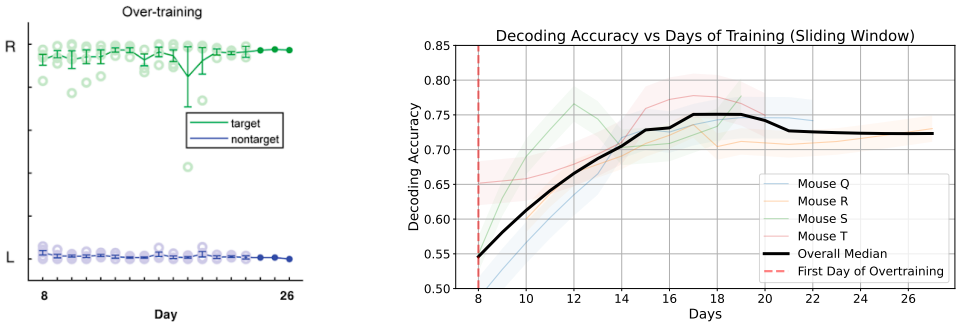


Figure 1: (Left) Behavior of mice on a binary discrimination task of odors, where mice indicate their selected choice by licking left or right. Y-axis is fraction of licks left vs right on a given day. Day 8 is the first day of overtraining. Correct lick for non-target is left, and right for target. The mice can discriminate near-perfectly as overtraining begins. Reproduced with modification from (Berners-Lee et al., 2023). (right) Increase in decoding accuracy from 10-fold linear discriminant analysis for each session. Shaded colored lines in background show standard errors for each mouse.

move far from initialization. In classification problems, margin maximization can be one type of rich learning (Mohamadi et al., 2024; Matyasko & Chau, 2017), where the margin of a classifier is defined as the distance from the decision boundary to the nearest data point. There are some subtleties around defining margin maximization in a setting where a decoder is being retrained on an evolving feature map; we state a precise working definition in Section 3.1.2.

Grokking. Much theoretical work has been done to understand the grokking phenomenon, discovered by Power et al. (2022), in which task-specific representation learning occurs long after training accuracy has saturated (Nanda et al., 2023; Barak et al., 2022; Liu et al., 2022a). Current theories explaining grokking are varied: Liu et al. (2022b) claim it is due to weight norm at initialization, Nanda et al. (2023) suggest it is driven by weight decay, Varma et al. (2023) suggest it is due to competition between memorizing and generalizing neural circuits. A flurry of recent theoretical papers unify these views this as late-time feature learning, where networks begin lazy and then abruptly transition into the rich regime much later during training (Kumar et al., 2023; Lyu et al., 2023). Recent empirical evidence has accumulated in support of this perspective (Edelman et al., 2024; Clauw et al., 2024; Mohamadi et al., 2024), where rich learning often takes the form of margin maximization in classification tasks. For instance, Morwani et al. (2023) and Mohamadi et al. (2024) show that on certain classes of tasks, margin maximization during overtraining provably causes grokking.

3 Revisiting Neural Data From Overtraining

3.1 Setup

Berners-Lee et al. (2023) collected tetrode recordings from posterior piriform cortex in adult mice during overtraining on a binary odor discrimination task (Figure 1). The posterior piriform cortex (PPC) is a key brain region involved in olfactory processing (Blazing & Franks, 2020; Srinivasan & Stevens, 2017). An odor is defined as an n-hot vector of length k, with $n = 3, k = 13$. In other words, a unique combination of three chemicals out of thirteen possible chemicals. The "target" odor is uniquely defined as consisting of three specific chemicals, where non-targets never contain any of these three. The discrimination task mice are trained on is to distinguish the target from the nontarget odors. After an initial training period of 8 days, the mice master the task, reaching ceiling behavioral performance (see Figure 1a), yet training continues in the same way for a further 18 days ("overtraining"). This period is when firing rate data is recorded. After this overtraining period is complete, the mice are exposed to held-out odors ("test/probe" examples), which share one odorant with the target odor and therefore are more difficult to correctly classify.

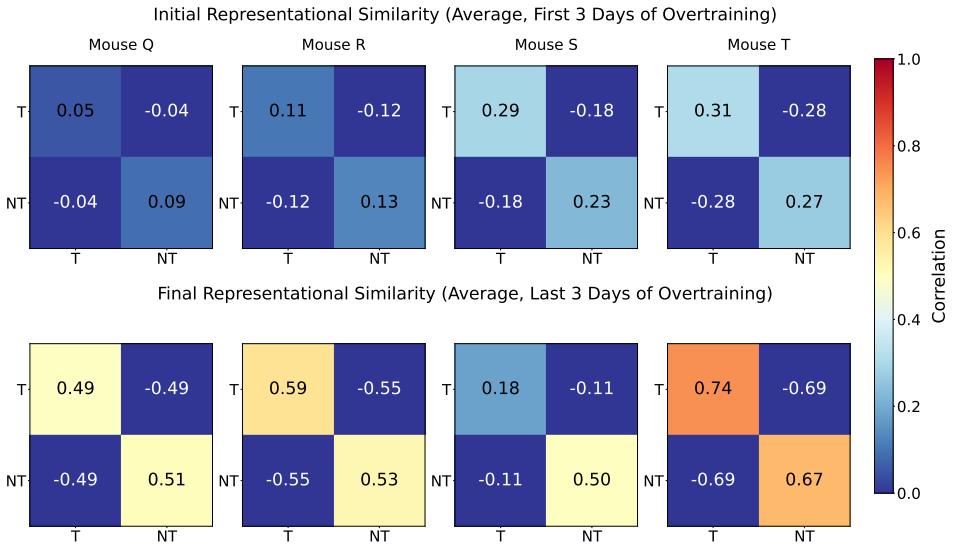


Figure 2: Representational similarity matrices plotting average pairwise correlation between target and nontarget population responses from piriform cortex on the first (top) and last (bottom) days of overtraining. We drop the last few days of data for Mouse T due to inconsistencies in data; methodological details in A.1. We compare the separations (anticorrelation) above to that of random/ablated baselines in Appendix A.5, finding them significantly larger.

3.1.1 Representations of Target/Nontarget Odors Separate During Overtraining

The key object of study throughout the neural data reanalysis will be the population firing rate matrix $X \in \mathbb{R}^{N \times D}$ which contains the neural response vector (activations) for each trial (presented stimulus) on a given-mouse day. There are around $N \approx 200$ trials per day each recording around $D \approx 15$ neurons, and around 15-20 days for each of 4 mice overall. We compute the average dot product (over trials) between the (z-scored) firing rates between target and nontarget trials, finding that representations of target and non-target odors continue to separate over the course of overtraining. We plot the representational similarity matrices at the beginning and end of overtraining for each mouse (Figure 2), with further methodological details deferred to Appendix A.1.

The measurement of such representational similarity metrics is standard in neuroscience (Kriegeskorte et al., 2008), and similar to common interpretability techniques in deep learning that involve probing for directions in activation space (Bau et al., 2019; Olah et al., 2020; Elhage et al., 2021; Zhang & Nanda, 2023) and comparing them, for instance by computing their dot product. The continued separation of class representations between the target and nontarget odors in Figure 2 provides evidence that, despite no visible changes in behavior, there is continued representation learning in mouse PPC during overtraining on this task. Such continued learning may aid in generalization on unseen odors, as indeed Berners-Lee et al. (2023) find empirically to be the case.

What do these task-relevant changes look like visually? In Figure 3, we plot the projection of odor representations on a given session onto the first 2 principal components of the firing rate matrix for that session, which is trials by neurons for each mouse. We see generally that the bottom row has more separated representations than the top row, providing a qualitative view of how target odors are more distinguishable in PPC representation space after overtraining than before. These results provide a glimpse of the representational dynamics at play as decoder accuracy is increasing, broadly consistent with the increasing separation we see quantitatively in Figure 2. Note that this continued representation learning ("rich feature learning") despite no changes in training accuracy ("behavior") is precisely what characterizes the abrupt generalization found in the grokking phenomenology in deep learning (Lyu & Li, 2019; Kumar et al., 2023; Mohamadi et al., 2024).

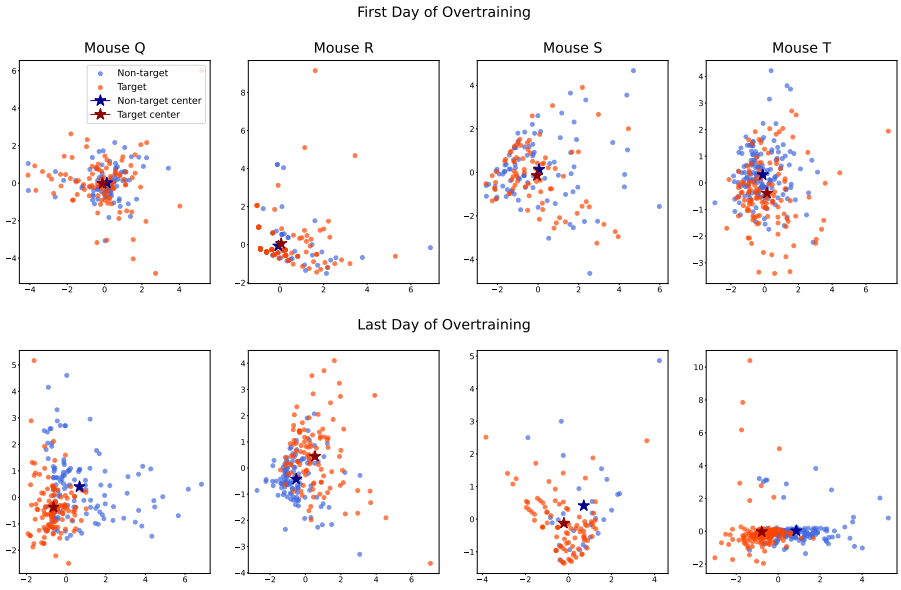


Figure 3: Neural representations (projected onto the top 2 principal components) of target and nontarget odors on the first (top) and last (bottom) day of overtraining for each mouse, showing qualitative separation during overtraining measured quantitatively in the representational similarity matrices in Figure 2. Cluster centers plotted for ease of visual comparison.

3.1.2 Representational Changes Lead to an Improved Maximum Margin Classifier

We now take an alternative perspective on the class separation of odor representations and instead examine the robustness of the learned classifier, as measured by its margin, a classical notion of confidence and robustness from statistical learning theory (Bishop, 2006).

We begin by clarifying a few distinct notions of margin maximization that apply when a decoder is trained on an evolving hidden representation.

1. A linear decoder is trained on a fixed feature representation ϕ for a fixed number of steps, which may or may not involve convergence of the decoder.
2. A linear decoder is trained on a fixed feature representation ϕ to convergence.
3. The feature representation ϕ_t *changes* over time, and at *each time step* t, a linear decoder is trained on the feature representation to convergence.

While the margin of the classifier defined by the decoder can increase as the decoder is trained for more steps on a fixed feature representation as in (1), if the margin of the max-margin (converged) classifier is increasing in (3), it implies the features are evolving to allow for a max-margin classifier with larger margin. This latter notion of margin-maximization that involves feature learning is the notion of margin maximization we will use.

We can define this mathematically as follows. Let $\phi_t(x) \in \mathbb{R}^N$ be the representation at time t for data x, and let $w(t)$ be the parameters of the max margin SVM solution to the label classification problem with $\phi_t(x_i)$ as input point i and y_i as label. Then the margin $\mathcal{M}(t)$ takes the form:

$$\mathcal{M}(t) = \frac{1}{|w(t)|} \quad \text{where} \quad w(t) = \min_{w} |w|^2 \quad \text{such that} \quad y_i w \cdot \phi(x_i)_t \geq 1. \tag{1}$$

In these terms, *our hypothesis is that the evolution on odor representations takes the form of (max) margin maximization*.

While the "margin" of a classifier usually refers to the distance to the single closest point, we take it to refer to the distance to the closest 1% of points, as neural activity underlying individual points

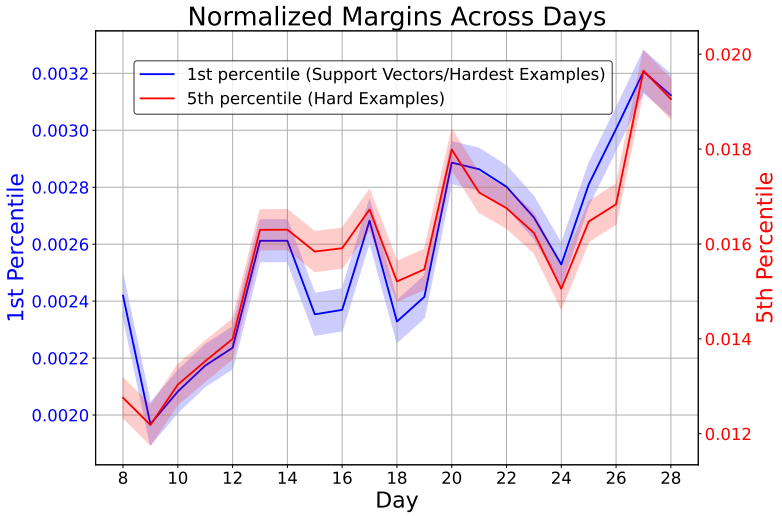


Figure 4: The average distance from the decision boundary within the smallest 1% and 5% distances extracted from a linear support vector machine trained on population data from PPC. We normalize margins for each mice so data are comparable across mice, shaded bands show standard errors.

(trials) is highly variable. We track the closest points to the learned decision boundary in a linear SVM trained on the same population decoding data to convergence with a fixed number of steps throughout. We measure the average distance from this classifier to the closest 1% (hardest examples) as well as 5% (hard examples) of points, finding in Figure 4 that these quantities increase substantially during overtraining.

Therefore, PPC representations evolve to increase the margin of the maximum margin classifier trained on those representations, similar to how many works find that deep neural networks which "grok" during overtraining on classification tasks are implicitly driven to do so by maximizing margin (Morwani et al., 2023; Mohamadi et al., 2024; Lyu & Li, 2019). This margin-maximization perspective is provocative not only because it provides a geometric and intuitive picture of continued learning in this setting, but because it is suggestive of *why* certain generalizing solutions are learned. For instance, Morwani et al. (2023); Mohamadi et al. (2024); Lyu & Li (2019) prove in various settings that the reason that neural networks often solve particular tasks (eg. modular arithmetic) using very specific types of features (eg. Fourier features in Nanda et al. (2023)), is because the regularized optimization trajectories are implicitly biased towards margin-maximizing solutions.

We emphasize that since test data (unseen test odors) is missing during the training period, we cannot make claims about whether the cortical representation begins lazy or rich (ie. evolve significantly during the initial 7-day learning period compared to initialization), and this is an important avenue for future work and limitation of our reanalysis. While the optimization algorithms at play in brains are poorly understood, it is speculatively possible that some form of "implicit-bias" towards generalizing solutions may also be the reason that overtrained mice do not just memorize their training odors, but learn representations that are helpful on difficult, held-out examples.

4 Theory & Modeling

4.1 A synthetic model of mouse piriform cortex

Setup. We construct a synthetic version of the odor discrimination task that captures key computational considerations at an abstract level. We consider a more biologically plausible counterpart in Appendix A.7, finding the same phenomenology persists in the same way. We take our earlier definition of odors as n-hot vectors of odorants (chemicals) in $\{0,1\}^k$ that we pass through a random projection, inspired by the functional role of the glomeruli (Blazing & Franks, 2020), with $n = 10, k = 100$. We train a one-hidden layer multilayer perceptron (MLP) on these embeddings

to classify the one target odor from 200 non-target odors. We use a ReLU nonlinearity to simulate non-negative firing rates. We also hold out 20 test examples that share an odorant (a single entry in the bit string) with the target. These examples are thus closer in both Hamming and projected space to the target, and therefore harder to discriminate. We train the MLP with vanilla gradient descent on a cross-entropy loss, as is standard in classification.

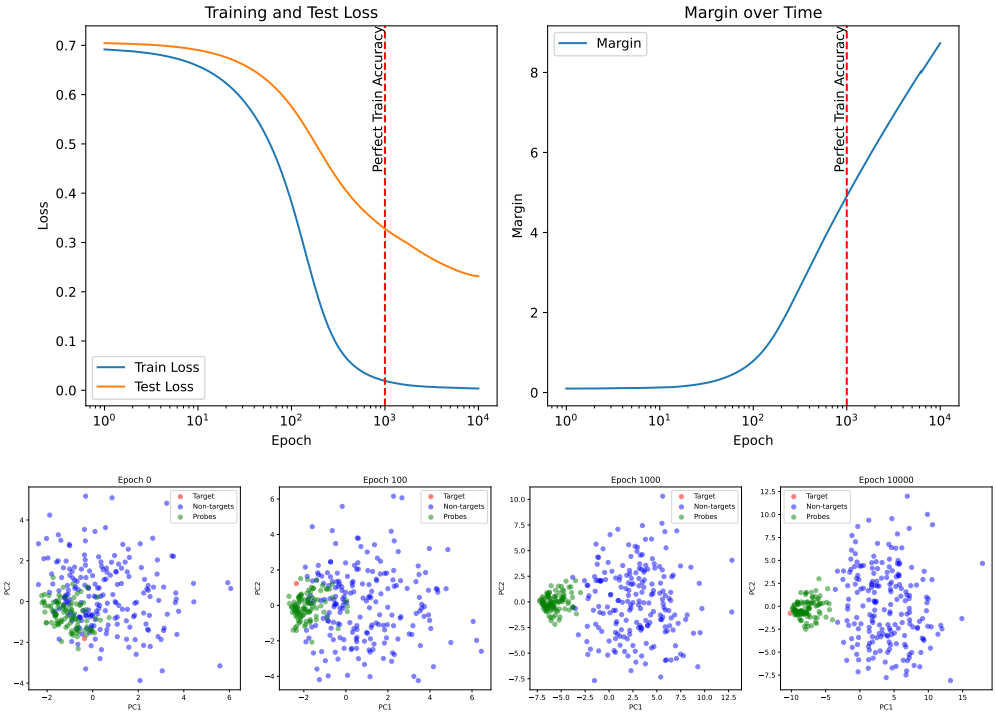


Figure 5: Recapitulating and interpreting mouse piriform cortex dynamics in a simple model. Top left: training and test loss over the course of training. Top right: margin over the course of training. Bottom: projections of the top 2 principal components for several training epochs. Notice how the target and probes (test trials) separate during overtraining (Epochs 1000-9000) despite the target never being trained on a probe example.

Mouse behavior and neural dynamics are captured by an overtrained MLP. This simple model captures the key computational phenomena at play in piriform data that was the subject of analysis in Section 3. In Figure 5, we can see that the test loss (on the probe trials) continues to decrease after training loss plateaus at a low value, in line with a continued increase in the classifier margin. The readout layer of our MLP acts as the "decoder" on the hidden layer representation, which are evolving to allow for an increased max-margin solution to be found. We also see a separation between the target and the nontarget representations emerge during training (up to epoch 1000). However, at epoch 1000, when training loss has converged, the target and probes are not yet easily separable. Strikingly, overtraining on the target-nontarget separation task improves performance on the probe odors, *which are out-of-distribution from the training set*. This happens because the network maximizes margin between the target and nontarget representations during the overtraining period after epoch 1000. This manifests as the red dot (target) gradually separating from the green cluster (test probes) during overtraining. There is an abrupt grokking-like transition in test accuracy (the fraction of probe trials correctly classified) during the overtraining period, Epochs 1000-9000, exactly as we see in Figure 6.

We can test whether margin maximization is the causal factor in this simple model by ablating it and asking whether late-time test loss improvements vanish concomitantly. We do this by replacing the Cross-Entropy objective, which is known to drive late-time margin maximization (Lyu & Li, 2019; Lyu et al., 2023), with a Hinge loss which has a hard-margin objective instead. This means after the model achieves a margin C on each training point, loss on that point vanishes, so the late-time

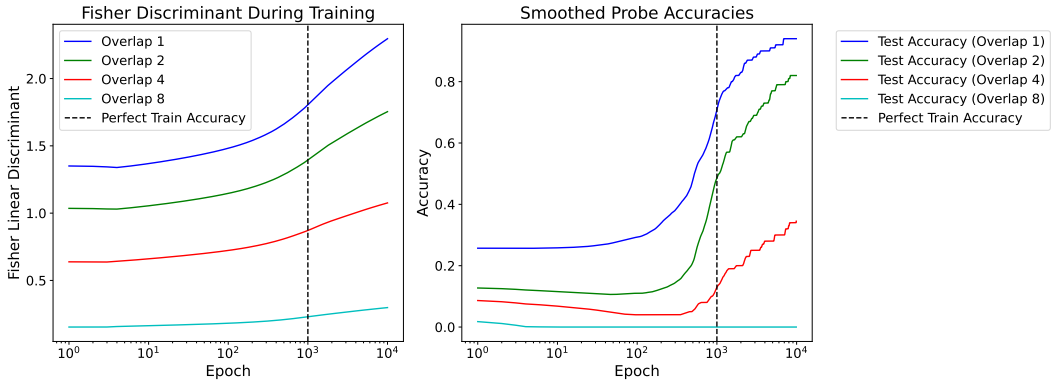


Figure 6: Probe trials are learned in order of increasing overlap with target. (Left) Tracking the Fisher Discriminant (FD) between target and probe classes throughout training. This measures signal to noise in representation space. (Right) The accuracy ("behavioral performance") on increasingly difficult trials over time, exhibiting grokking.

loss incentive for margin maximization is removed. We find in this ablated setting margin saturates approximately when training loss does, and that test-loss saturates at this point in training as well, so that the grokking-like effect is ablated. Mathematical details and plots are in Appendix A.6, supporting the notion of a causal link.

Feature learning in the synthetic model. We increased the dimension of the task "odor" vector as well as the number of nonzero entries so that we can introduce many types of "probe" trials that share increasing amounts of odors (overlap) with the target, allowing us to vary difficulty. Note that our model is never trained on the probes, but monitoring probe accuracy allows us to measure how margin maximization on the original nontarget odors (which are guaranteed to have zero overlap with the target) can be helpful in solving more difficult trials.

In Figure 6, we track the Fisher Discriminant (FD) between target and probe classes, as well as the accuracy on discriminating the two throughout training. The FD is a statistical quantity that uses the linear combination of the features that maximize the separation between two classes, and represents the direction in the feature space along which the projected class means are maximally separated while minimizing the variance within each class. This is the quantity that allows the linear discriminant analysis decoding accuracy to increase; the portion of the lines after low train loss is reached mimics how the median LDA posterior probability on the correct class increases in Figure 1 (right). Mathematical details about the Fisher Discriminant and its relation to margin, LDA, and more are deferred to Appendix A.4.

Possible biological drivers of late time learning. Certain types of loss functions have been proven in a deep learning setting to drive late-time margin maximization (Soudry et al., 2017; Lyu & Li, 2019). Here, we speculate on biologically plausible components of the objective in cortex that may plausibly serve a similar role. A simple but plausible driver of late-time learning is the set of energetic/metabolic constraints on biological organisms. Such constraints been suggested to give rise to a variety of neural phenomena (Sterling & Laughlin, 2015), for example disentangled representations (Pehlevan et al., 2017; Olshausen & Field, 1996; Plumbley, 2003; Whittington et al., 2023). Pertinently, such constraints, in the form of weight decay in machine learning, are often the driving force for grokking in some settings (Nanda et al., 2023; Liu et al., 2022b). These constraints force neural networks to deviate from using a lazy (kernel) method, and force the weights to move in task-relevant directions (Lyu & Li, 2019; Kumar et al., 2023). Another possibility in Berners-Lee et al. (2023) is a very small supervised error signal from occasional mistakes made by the mice for idiosyncratic reasons that could plausibly be driving late-time learning (See Figure 1).

Finally, a key plausible driver of late-time learning in cortex is the existence of unsupervised terms in the implicit objective of organisms that penalize low confidence predictions. For instance, cross-entropy loss is nonzero even when an example is correctly classified because it can be seen as penalizing the *uncertainty* of model predictions, rather than merely *correctness*. The fact that the loss is nonvanishing even when predictions are correct serves as a signal to guide task-relevant

learning during overtraining. If biological learning systems store uncertainty in predictions and/or use them for behavior, as indeed is known to be the case in rodent cortex (Kepecs et al., 2008) and brains more generally (Lak et al., 2017; Fetsch et al., 2014; Komura et al., 2013), then such implicit penalties on uncertainty will serve a similar functional role to an unsupervised term that keeps loss nontrivial even as training accuracy saturates. This nonvanishing loss during behavioral plateau can plausibly drive continued feature learning during overtraining.

4.2 Rich Learning During Overtraining can Explain Overtraining Reversal

Does our model of the dynamics underlying learning during overtraining make any new predictions outside our immediate setup? One concrete prediction of late-time feature learning is faster adaptation under task reversal for any learned task. In machine learning this means fast adaptation of a predictor to distribution shift of the form $y(x) \mapsto -y(x)$. If a network learns features to compute $y(x)$, it is possible many such features can be reused to compute $-y(x)$. This is not true if the network learns $y(x)$, for instance, in the lazy regime, in which case y was fit using the initial features, which were not task-dependent, so cannot be advantageously "repurposed" for a reversed variant of the task. This particular prediction about reversal learning may seem as oddly specific to a machine learning audience, but in fact overtraining reversal has been a topic of study by experimental pyschologists for decades (Mackintosh, 1969; Richman et al., 1972; Mead, 1973). Theoretical work in psychology offers qualitative cognitive explanations in terms of attention allocation (Lovejoy, 1966), where our model of overtraining naturally posits a stronger, more fine-grained, neural model for overtraining reversal that is the first of its kind to our knowledge. We begin by defining the phenomenon.

The Overtraining Reversal Effect. The reversal effect is the empirical finding that animals that are overtrained on a task more quickly adapt when the task is reversed. Concretely, when animals are given a perceptual discrimination problem and the rewarded mapping is reversed, those animals that were overtrained on the original mapping adapt to the task reversal in fewer trials than those that were not overtrained, see (Mackintosh, 1969) for further discussion.[1] We propose a simple explanation for this phenomenon in terms of feature learning in the modern machine learning sense.

A Mathematical Model. We illustrate this with the simplest possible network model that exhibits this effect, a two layer linear network (Saxe et al., 2013). Reversal learning proceeds in two steps. First, we train a two layer linear network with N hidden neurons $f(\boldsymbol{x}) = \frac{1}{N\gamma_0}\boldsymbol{w}^2 \cdot \boldsymbol{h}(\boldsymbol{x})$, where $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{W}^1\boldsymbol{x}$ is optimized for T steps on task $y(x)$. After this, a new readout vector $\boldsymbol{v}$ is introduced to give a new output $f_{\text{rev}}(\boldsymbol{x}) = \frac{1}{N\gamma_0}\boldsymbol{v}\cdot\boldsymbol{h}(\boldsymbol{x})$ which is trained on the reversed task $-y(\boldsymbol{x})$. The matrix $\boldsymbol{W}^1$ can be updated in the second phase of learning as well. At large width N, we can invoke mean field theory ideas (Bordelon & Pehlevan, 2022, see Appendix F.1.1), allowing us to describe the hidden kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{N}\boldsymbol{h}(\boldsymbol{x}) \cdot \boldsymbol{h}(\boldsymbol{x}')$ dynamics in both phases of training. For whitened data and squared error loss, it suffices to compute the projection of the kernel $\boldsymbol{K}$, $K_y = \boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{y}$, and the projection of the predictions $f_y = \boldsymbol{y}^\top \boldsymbol{f}$ along the direction of the labels $\boldsymbol{y}$, which gives:

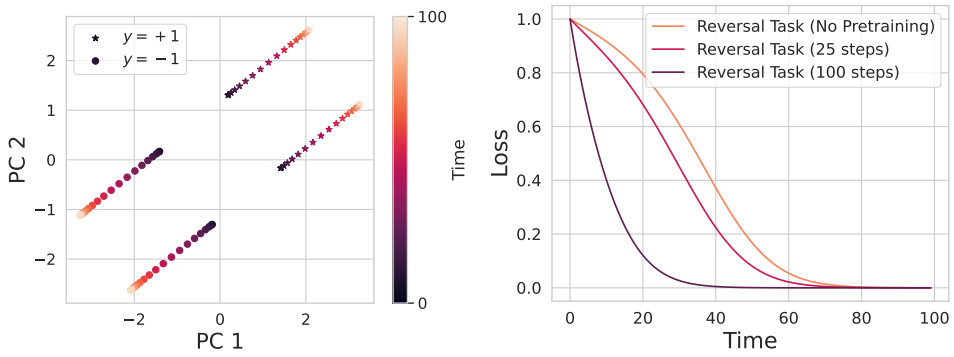
$$K_y(t) = \sqrt{1 + \gamma_0^2 f_y(t)^2}\ ,\ \partial_t f_y(t) = 2\sqrt{1 + \gamma_0^2 f_y(t)^2}\ (y - f_y(t)). \tag{2}$$

We show that the kernel evolves only in this $\boldsymbol{y}\boldsymbol{y}^\top$ direction in Figure 7 (a), where points separate by their target category. These dynamics are run from time $t = 0$ to time T, which gives a terminal value of the kernel alignment $K_y(T)$. In the second phase of learning ($t > T$) where the new output f_{rev} is fit, the predictions on the reversal task in the $\boldsymbol{y}$ direction evolve as

$$\partial_t f_{\text{rev},y}(t) = \sqrt{(K_y(T) + 1)^2 + 4\gamma_0^2 f_{\text{rev},y}(t)^2}\ (-y - f_{\text{rev},y}(t)). \tag{3}$$

We note that larger kernel-target overlap $K_y(T)$ leads to faster training in this second phase $t > T$, as rate of learning on the second task is a direct function of alignment on the first task. We verify this with simulations in Figure 7 (b). Since $K_y(T)$ is monotonically increasing with T, longer training on the first task will accelerate learning of the reversed task, providing an explanation for overtraining reversal. The key intuition is that adaptation time depends on alignment $K_y = K_{-y}$, where $\boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{y} = (-\boldsymbol{y}^\top)\boldsymbol{K}(-y)$, so overtraining leads to faster adaptation.

[1] We note that there has been debate on whether the phenomenon is general across species and tasks (Beck et al., 1966; Warren, 1978).



(a) Pretraining Feature Movement ($t < T$) (b) More Pretraining = Faster Reversal ($t > T$)

Figure 7: A simple two layer model explains faster reversal learning after overtraining on the original task. (a) We plot the original hidden representations (projected from N hidden dimensions to the top two principal components) of a width $N = 250$ network for 4 odor mixtures, with two odors positive and two odors negatively labeled. (b) When training a new readout on the reversal task $-y(x)$ after pretraining for t steps on the original labels $y(x)$, we find that additional pretraining steps T accelerate learning the reversal task, consistent with theory.

5 Conclusion

We find that recent work in mouse olfaction provides evidence for rich sensory learning taking place under the veil of a behavioral plateau in mice. This suggests that difficult tasks that require significant changes in learned representations may benefit from overtraining, even if behavioral performance remains outwardly unchanging. It also offers a glimpse of shared structure between artificial and biological networks, where phenomenologically similar dynamics appear in both cortex and MLPs.

Our work has a number of limitations. A key limitation of our work is that the neural evidence we draw upon is observational, and new experiments in systems neuroscience are required to validate this hypothesis. We describe in detail a proposed experimental setup in Appendix A.2. Another is that the number of neurons recorded per mouse-day (around 20) is small, so that trends can be noisy (see Appendix A.5 for an example). A third limitation is that we consider only olfactory cortex, and only one type of perceptual discrimination task – whether such late-time learning is a universal property across animals and tasks remains to be seen. We hope our hypotheses drive further work on learning dynamics in cortex during overtraining in animals. Our study raises several new questions: What are the classes of tasks in which such overtraining is beneficial, and what do they have in common? Can deep learning models of cortex hint at the answer? What is the nature of the implicit objective in sensory cortex driving this late-time representation learning? Altogether, our work aims toward a more complete characterization of the behavior of sensory cortex on perceptual discrimination tasks, and to unify the mechanics underlying such behavior with existing theory and observations in deep learning.

6 Ethics Statement

Our work models the dynamics of learning during overtraining in mice and MLPs, aiming towards a more complete characterization of learning in both artificial and biological neural networks. We do not note any important ethical consequences in particular.

7 Reproducibility Statement

We model the data of (Berners-Lee et al., 2023), which is publicly available for open reproduction, and will open source our code to train and interpret artificial networks.

REFERENCES

Shahab Bakhtiari. Can deep learning model perceptual learning? *The Journal of Neuroscience*, 39 (2):194, 2019.

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.

Carol A Barnes, Matthew S Suster, Jiemin Shen, and Bruce L McNaughton. Multistability of cognitive maps in the hippocampus of old rats. *Nature*, 388(6639):272–275, 1997.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

C. H. Beck, J. M. Warren, and R. Sterner. Overtraining and reversal learning by cats and rhesus monkeys. *Journal of Comparative and Physiological Psychology*, 62(2):332–335, 1966. doi: 10.1037/h0023674.

Alice Berners-Lee, Elizabeth Shtrahman, Julien Grimaud, and Venkatesh N Murthy. Experience-dependent evolution of odor mixture representations in piriform cortex. *PLoS Biology*, 21(4): e3002086, 2023.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0-387-31073-2.

Robin M Blazing and Kevin M Franks. Odor coding in piriform cortex: mechanistic insights into distributed coding. *Current opinion in neurobiology*, 64:96–102, 2020.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

Lars Buesing, Johannes Bill, Bernhard Nessler, and Wolfgang Maass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS computational biology*, 7(11):e1002211, 2011.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Kenzo Clauw, Daniele Marinazzo, and Sebastiano Stramaglia. Information-theoretic progress measures reveal grokking is an emergent phase transition. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

Benjamin L Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*, 2024.

Neil Elhage, A Nanda, Chris Olah, Shan Carter, Nick Cammarata Gabriel Schubert Wang, and Curtis Hernandez. A mathematical framework for transformer circuits. In *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/.

K Anders Ericsson, Ralf Th Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363–406, 1993.

Matthew Farrell, Stefano Recanatesi, and Eric Shea-Brown. From lazy to rich to exclusive task representations in neural networks and neural codes. *Current opinion in neurobiology*, 83:102780, 2023.

Christopher R Fetsch, Roozbeh Kiani, William T Newsome, and Michael N Shadlen. Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron*, 83(4):797–804, 2014.

Paul M Fitts and Michael I Posner. *Human Performance*. Brooks/Cole Publishing Company, 1967.

Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Rich and lazy learning of task representations in brains and neural networks. *BioRxiv*, pp. 2021–04, 2021.

Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270, 2022.

David J. Freedman, Maximilian Riesenhuber, Tomaso Poggio, and Earl K. Miller. Comparison of categorization of stimuli in the monkey prefrontal cortex: Frontal eye field and inferior convexity. *Journal of Neurophysiology*, 90(2):670–682, 2003.

Pieter M Goltstein, Sandra Reinert, Tobias Bonhoeffer, and Mark Hübener. Mouse visual cortex areas represent perceptual and semantic features of learned visual categories. *Nature neuroscience*, 24(10):1441–1451, 2021.

Takuya Ito and John D Murray. Multitask representations in the human cortex transform along a sensory-to-motor hierarchy. *Nature neuroscience*, 26(2):306–315, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Mikiko Kadohisa and Donald A Wilson. Separate encoding of identity and similarity of complex familiar odors in piriform cortex. *Proceedings of the National Academy of Sciences*, 103(41): 15206–15211, 2006.

Clifford G Kentros, Naveen T Agnihotri, Samantha Streater, Robert D Hawkins, and Eric R Kandel. Increased attention to spatial context increases both place field stability and spatial memory. *Neuron*, 42(2):283–295, 2004.

Adam Kepecs, Naoshige Uchida, Hatim A Zariwala, and Zachary F Mainen. Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210):227–231, 2008. doi: 10.1038/nature07200. URL `https://pubmed.ncbi.nlm.nih.gov/18690210/`.

Tae Hoon Kim, Yajun Zhang, Julien Lecoq, Jason C Jung, Jun Li, Hongkui Zeng, and Mark J Schnitzer. Long-term optical access to an estimated one million neurons in the live mouse cortex. *Cell Reports*, 17(12):3385–3394, 2016.

Yusuke Komura, Akifumi Nikkuni, Naofumi Hirashima, Takuya Uetake, and Atsushi Miyamoto. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, 16(6):749–755, 2013.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Kamesh Krishnamurthy, Ann M Hermundstad, Thierry Mora, Aleksandra M Walczak, and Vijay Balasubramanian. Disorder and the neural representation of complex odors. *Frontiers in Computational Neuroscience*, 16:917786, 2022.

Tanishq Kumar, Blake Bordelon, Samuel J Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. *arXiv preprint arXiv:2310.06110*, 2023.

Armin Lak, Katsuhiro Nomoto, Mehdi Keramati, Masamichi Sakagami, and Adam Kepecs. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology*, 27(6):821–832, 2017.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022a.

Ziming Liu, Eric J Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. *arXiv preprint arXiv:2210.01117*, 2022b.

Elijah Lovejoy. Analysis of the overlearning reversal effect. *Psychological Review*, 73(1):87, 1966.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *The Twelfth International Conference on Learning Representations*, 2023.

Nicholas John Mackintosh. Further analysis of the overtraining reversal effect. *Journal of Comparative and Physiological Psychology*, 67(2p2):1, 1969.

Emily A Mankin, Fraser T Sparks, Begum Slayyeh, Robert J Sutherland, Stefan Leutgeb, and Jill K Leutgeb. Neuronal code for extended time in the hippocampus. *Proceedings of the National Academy of Sciences*, 109(47):19462–19467, 2012.

Alexander Mathis, Dan Rokni, Vikrant Kapoor, Matthias Bethge, and Venkatesh N Murthy. Reading out olfactory receptors: feedforward circuits detect odors in mixtures without demixing. *Neuron*, 91(5):1110–1123, 2016.

Alexander Matyasko and Lap-Pui Chau. Margin maximization for robust classification using deep learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 300–307. IEEE, 2017.

Philip G Mead. The effect of overtraining on reversal shift behavior in rats reinforced with electrical brain stimulation. *Physiological Psychology*, 1(4):330–332, 1973.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J Sutherland. Why do you grok? a theoretical analysis of grokking modular addition. *arXiv preprint arXiv:2407.12332*, 2024.

Depen Morwani, Benjamin L Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: case studies in algebraic tasks. *arXiv preprint arXiv:2311.07568*, 2023.

K Nakamura, Y Gu, D Kato, et al. Representational drift in the mouse visual cortex. *Nature Communications*, 12:5186, 2021.

Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Allen Newell and Paul S Rosenbloom. Mechanisms of skill acquisition and the law of practice. In John R Anderson (ed.), *Cognitive skills and their acquisition*, pp. 1–55. Lawrence Erlbaum Associates, 1981.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. In *Distill*, 2020. https://distill.pub/2020/circuits/.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Cengiz Pehlevan, Sreyas Mohan, and Dmitri B Chklovskii. Blind nonnegative source separation using biological neural networks. *Neural computation*, 29(11):2925–2954, 2017.

Mark D. Plumbley. Algorithms for nonnegative independent component analysis. *IEEE Transactions on Neural Networks*, 14(3):534–543, 2003.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Shanshan Qin, Shiva Farashahi, David Lipshutz, Anirvan M Sengupta, Dmitri B Chklovskii, and Cengiz Pehlevan. Coordinated drift of receptive fields in hebbian/anti-hebbian network models during noisy representation learning. *Nature Neuroscience*, 26(2):339–349, 2023.

Charles L Richman, Karol Knoblock, and Wayne Coussens 1. The overtraining reversal effect in rats: A function of task difficulty. *The Quarterly journal of experimental psychology*, 24(3): 291–298, 1972.

Michael E Rule, Timothy O'Leary, and Christopher D Harvey. Causes and consequences of representational drift. *Current opinion in neurobiology*, 58:141–147, 2019.

Nicole C. Rust and James J. DiCarlo. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, 30(39):12978–12995, 2010.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Carl E Schoonover, Sarah N Ohashi, Richard Axel, and Andrew JP Fink. Representational drift in primary olfactory cortex. *Nature*, 594(7864):541–546, 2021.

Amin MD Shakhawat, Carolyn W Harley, and Qi Yuan. Arc visualization of odor objects reveals experience-dependent ensemble sharpening, separation, and merging in anterior piriform cortex in adult rat. *Journal of Neuroscience*, 34(31):10206–10210, 2014.

D Soudry, E Hoffer, and N Srebro. The implicit bias of gradient descent on separable data. arxiv. *arXiv preprint arXiv:1710.10345*, 2017.

Shyam Srinivasan and Charles F Stevens. A quantitative description of the mouse piriform cortex. *bioRxiv*, pp. 099002, 2017.

Peter Sterling and Simon Laughlin. *Principles of Neural Design*. MIT Press, 2015.

Merav Stern, Kevin A Bolding, LF Abbott, and Kevin M Franks. A transformation from temporal to ensemble coding in a model of piriform cortex. *Elife*, 7:e34831, 2018.

LT Thompson and PJ Best. Long-term stability of the place-field activity of single units recorded from the dorsal hippocampus of freely behaving rats. *Brain research*, 509(2):299–308, 1990.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Peter Y Wang, Cristian Boboila, Matthew Chin, Alexandra Higashi-Howard, Philip Shamash, Zheng Wu, Nicole P Stein, LF Abbott, and Richard Axel. Transient and persistent representations of odor value in prefrontal cortex. *Neuron*, 108(1):209–224, 2020.

J. M. Warren. Overtraining and extradimensional shift learning by cats. *Bulletin of the Psychonomic Society*, 12(3):177–178, 1978. doi: 10.3758/BF03329663.

Li K Wenliang and Aaron R Seitz. Deep neural networks for modeling visual perceptual learning. *Journal of Neuroscience*, 38(27):6028–6044, 2018.

James CR Whittington, Will Dorrell, Surya Ganguli, and Timothy Behrens. Disentanglement with biological constraints: A theory of functional cell types. In *The Eleventh International Conference on Learning Representations*, 2023.

Amit Yashar and Rachel N Denison. Feature reliability determines specificity and transfer of perceptual learning in orientation search. *PLoS Computational Biology*, 13(12):e1005882, 2017.

Rafael Yuste. From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, 16(8): 487–497, 2015.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

Yaniv Ziv, Laurie D Burns, Eric D Cocker, Elizabeth O Hamel, Kunal K Ghosh, Lacey J Kitch, Abbas El Gamal, and Mark J Schnitzer. Long-term dynamics of ca1 hippocampal place codes. *Nature neuroscience*, 16(3):264–266, 2013.

A APPENDIX

A.1 METHODOLOGICAL AND IMPLEMENTATION DETAILS

The raw data consists of around 70 sessions, each representing a recording of 20 random neurons on a given mouse-day. The data is spikes over time for each neuron on each trial, of which there are 200-300 in a session. After odor onset, mice are given 2.5s to lick in response. We count the number of spikes in the final 500ms part of this window and decode on this number, which gives us a $N \times T$ neurons by trials spike count matrix to decode on for each mouse-day. We drop neurons with low (less than 4) number of spikes across all trials classifying them as unresponsive to the odors, we z-score this matrix and use 10-fold cross validation to get decoding accuracy, over 20 iterations. Since session to session data is very noisy, we plot smoothed statistics using a sliding window and plot standard error of all statistics. It should be noted the last few days of mouse T overtrainining are dropped because of notable inconsistencies in neural data such as the decoding accuracy plummeting to far below chance while behavior remained high, suggesting an unknown and idiosyncratic cause for unreliable decoding accuracy in this mouse.[2] Note that the activity matrix was z-scored across trials, so every neuron contributed equally in our analysis.

A.2 TESTABLE PREDICTIONS & EXPERIMENTAL PROPOSALS

A.2.1 A CONCRETE PROPOSAL FOR PIRIFORM CORTEX

While reanalysis of existing data suggests that rich learning during overtraining is plausible, reanalysis on its own insufficient on their own to confidently establish a new empirical phenomenology. In particular, since the focus of (Berners-Lee et al., 2023) is studying an increase in target-selectivity and how training history affects the form of the learned representations, targeted experiments attempting to falsify our hypothesis may be helpful. We outline what one such experiment in piriform cortex may look like. Note that such an experimental setup can be applied to any part of sensory cortex in which at least hundreds of neurons can be recorded over several (10+) contiguous days of overtraining.

We propose an experiment of a similar flavor to that of (Berners-Lee et al., 2023), with several important changes. These changes include:

- Using several, as opposed to one, target odors would prevent the mice from memorizing the target and thus would make the task require more learning of odor identity, leaving more space for rich learning during overtraining.
- We would include a spectrum of difficult, held-out, examples, of $n-1$ types if an odorant is an n-hot vector of chemicals. Then, we would define j-probe trial as those that would share j elements with the target, allowing us to modulate distance in both Hamming and projected space, and therefore difficulty of classification as measured by distance to boundary.
- Out of every 200 trials on a given mouse-day, $n-1$ would be probe trials, one for each type. This would be included during the initial training period of 8 days, as well as during and after overtraining, which is not done in (Berners-Lee et al., 2023). This would ensure the test set has minimal impact on learning compared to the training set (which would be $200/n$ times more frequent).
- We would also ensure mice have as much time to recover between overtraining sessions as learning sessions, so the same amount of rich learning can potentially take place, and that conditions during overtraining are kept as similar to those seen during learning, as possible, to compare feature learning on equal footing.

A.2.2 IMPORTANT EXPERIMENTAL DETAILS FOR GENERAL SENSORY DISCRIMINATION TASKS

In seeking to test our hypothesis, there are a few key experimental details that apply broadly to experiments on sensory cortex beyond olfaction.

[2](Berners-Lee et al., 2023) include these unreliable decoding trials in their analysis.

- The task needs to involve diverse and novel stimuli. This can be measured by ensuring the ability to decode stimuli from population activity is not significantly higher than change before learning.
- The task needs to be difficult, but possible to master. This is a tricky requirement: here, mice reach up to 90% behavioral performance, which is close to the upper bound possible given noise and fluctuating conditions. Conversely, the otherwise relevant work in vision of (Goltstein et al., 2021) trains mice to discriminate gratings based on frequency and orientation. This task is not fully learnable, as the mice reach ceiling of 70%. Since a primary part of our hypothesis is that this late-time rich learning can persist when task performance is near-ceiling and thus in the absence of large error signal, an error rate of 30% is too high to test our hypothesis.
- It is crucial to separate a part of the data that is dedicated toward learning the task, namely the "training" set, and a part of the data dedicated toward testing generalization, namely the "test" set. As in the piriform setting, constructing some test examples to be particularly challenging by construction is likely to elucidate whether useful yet hidden representation learning takes place in sensory cortex during overtraining. Test trials should be infrequent so their effect on learning is small, but should be included throughout learning so that notions of training and test performance can be determined for the entire training-overtraining-test timeline.
- Chronic recordings should include the evolution of population activity both during training and overtraining. In this work, we cannot examine whether representations begin lazy or rich during learning, since data and probe trials are only available during the overtraining, and not the training, period. We note that mice are good organisms for such a task because larger animals, like macaques, take much longer and are more expensive to train, and chronic and invasive recordings during training can run the risk of causing serious damage to the organism. Conversely, smaller organisms like fruit flies cannot be trained to solve difficult perceptual discrimination tasks with in-vivo neural recordings, in any meaningful sense.

On such setups, our theory predicts continued increases in decoding accuracy for category label for several days after training performance saturates and stops changing noticeably, and separation between task-relevant class representations.

A.3 Continued Introduction

A third difficulty, in addition to the two mentioned in the main introduction, is that to see genuine representation learning during overtraining, the task must be difficult, but possible to master. In the deep learning setting of grokking, for instance, this is made precise by specifying that grokking occurs under a certain regime of task-model alignment (Kumar et al., 2023). Many settings in experimental neuroscience either study passive exposure (no perceptual task learning) (Rust & DiCarlo, 2010; Freedman et al., 2003), use tasks which are easy (high initial decoding accuracy) for sensory cortex (Wang et al., 2020), or use tasks that are too hard in the sense that the animal does only slightly above chance by the end (Goltstein et al., 2021).

Closely related, but distinct from our desired phenomenology of study is that of representational drift. In neuroscience, our understanding of the stability of neural network representations over time has also been revised. Classical work in neuroscience operated under the implicit assumption that neural representations were supported by stable neuronal activity (Barnes et al., 1997; Thompson & Best, 1990). While many neurons do show stable responses to stimuli, the recent discovery of "representational drift" shows that the population activity underlying some task-relevant stimulus can change over time (Kentros et al., 2004; Mankin et al., 2012; Ziv et al., 2013; Rule et al., 2019; Schoonover et al., 2021). Representational drift is commonly conceived as a diffusive process (Qin et al., 2023), and there are arguments that such drift can offer advantages in cortical processing, or confer robustness (Nakamura et al., 2021; Buesing et al., 2011). Our focus is instead on the *learning* of representations during overtraining on a task, and therefore our phenomenology of interest is orthogonal to, but may potentially co-occur with, representational drift.

A.4 MATHEMATICAL DETAILS ON FISHER DISCRIMINANT, LDA, AND MARGIN

Here we provide some background on the key statistical quantities we track, both in piriform cortex and our synthetic model, and their relations to each other. The Fisher Linear Discriminant (FD) is a statistical quantity that measures the ratio of cluster center separation to intra-cluster variance. Informally, it quantifies the signal to noise (SNR) in a binary classification problem in representation space, and is a quantity that Linear Discriminant Analysis (LDA) relies on for classification. It is particularly useful when the goal is to reduce the dimensionality of data while preserving as much class-discriminative information as possible.

Consider a dataset consisting of n samples, each of which belongs to one of two classes: C_1 and C_2. Let $x_i \in \mathbb{R}^d$ denote the d-dimensional feature vector of the i-th sample. The FD seeks a projection vector $w \in \mathbb{R}^d$ such that when the data is projected onto this vector, the separation between the projected means of the two classes is maximized relative to the projected variance within each class.

Formally, the projection of a data point x onto the vector w is given by $y_i = w^\top x_i$.

The objective of the FD is to find w that maximizes the SNR

$$J(w) = \frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2},$$

where $\mu_1 = w^\top m_1$ and $\mu_2 = w^\top m_2$ are the means of the projected points for classes C_1 and C_2, respectively, with $m_1 = \frac{1}{n_1}\sum_{x_i \in C_1} x_i$ and $m_2 = \frac{1}{n_2}\sum_{x_i \in C_2} x_i$ being the mean vectors of the classes in the original feature space. The terms s_1^2 and s_2^2 represent the variances of the projected points for classes C_1 and C_2, respectively.

To express $J(w)$ in terms of the original data, we can define the within-class scatter matrix S_w and the between-class scatter matrix S_b as

$$S_w = \sum_{x_i \in C_1} (x_i - m_1)(x_i - m_1)^\top + \sum_{x_i \in C_2} (x_i - m_2)(x_i - m_2)^\top,$$

$$S_b = (m_1 - m_2)(m_1 - m_2)^\top.$$

Then, we have that the Fisher criterion can then be rewritten as

$$J(w) = \frac{w^\top S_b w}{w^\top S_w w}.$$

Then to maximize $J(w)$, we differentiate it with respect to w and set the derivative to zero. This leads to the generalized eigenvalue problem given by

$$S_b w = \lambda S_w w.$$

The solution w that maximizes $J(w)$ is the eigenvector corresponding to the largest eigenvalue of the matrix $S_w^{-1} S_b$.

The FD is closely related to Linear Discriminant Analysis (LDA) in the following sense. In LDA, the assumption is that the data from each class is normally distributed with a common covariance matrix Σ. The LDA decoding rule assigns a new observation x to the class C_1 or C_2 based on which class's linear discriminant function is maximized with

$$\delta_k(x) = x^\top \Sigma^{-1} m_k - \frac{1}{2} m_k^\top \Sigma^{-1} m_k + \log \pi_k,$$

where π_k is the prior probability of class C_k, and $k \in \{1, 2\}$. The LDA solution involves projecting the data onto a lower-dimensional space using the same criterion as the Fisher Linear Discriminant, making the FD effectively equivalent to LDA when the data follows a Gaussian distribution

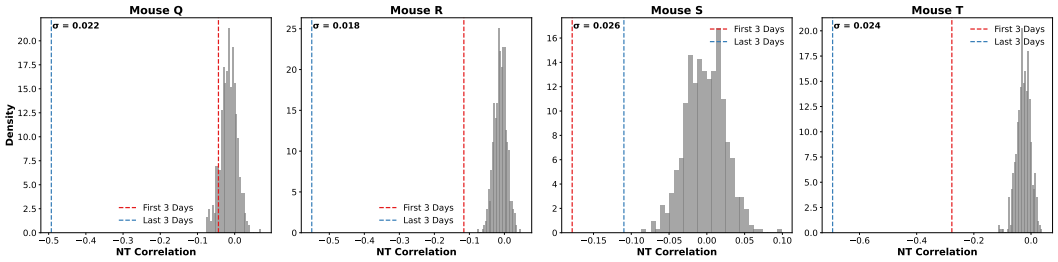


Figure 8: Cluster separation (anticorrelation between representations for classes) over 500 random firing rate matrices mean-matched with our z-scored firing rate matrix. σ denotes standard deviation. More negative is better.

with equal covariance matrices across classes. This is why we track the Fisher Discriminant in our synthetic model: because we see mean and median LDA posterior probability on the correct class increase during overtraining, suggesting that SNR as measured by Fisher Discriminant is also increasing.

The connection between FD and margin maximization can be understood in simpler, more qualitative terms: by examining the concept of separating hyperplanes in the context of binary classification. In margin-based methods such as SVMs, the goal is to find a hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest points from each class (support vectors). The FD also seeks to maximize the separation between classes, but it does so by maximizing the ratio of the between-class variance to the within-class variance. It can be thought of as a "soft margin" of sorts.

If the data is linearly separable, the direction w found by FD will tend to align with the direction that maximizes the margin between classes, although FD does not explicitly enforce the maximum margin criterion as SVM does. However, in many cases, the Fisher criterion leads to a solution that approximates the maximum margin direction, particularly when the distributions of the classes are approximately Gaussian with equal covariances, which is why in practice these two quantities often tend to vary together, as indeed they do both in our neural data and our synthetic model. Regardless, they measure different things.

A.5 Representational Separation Ablation

Since we z-score the firing rates, the diagonals of the correlatation matrices will be positive by definition as they take the form $||z||$. However, the expectation of the off-diagonals is zero for a set of random firing rates, with variance depending on dimension. If we are claiming there is significant separation in representations, we must check that the anticorrelation between the two firing rates is far from chance for our dimension $D \approx 15$. In Figure 8, we plot the distribution of anticorrelations for a mean-matched firing rate matrix with firing rate vectors drawn from a Gaussian for each mouse, and draw a line to represent the anticorrelation (representational quality, since the task is to discriminate two classes) both before and after the overtraining period. We plot these in Figure 8.

We can see generally that 1) overtraining causes separation, in the sense that representations are more anticorrelated after overtraining compared to training, and that 2) even at the beginning of overtraining (red lines) the separation is nontrivial compared to chance. This makes sense because the mice can perform the task at ceiling, so we expect representations at the beginning of overtraining (after a week of overtraining) to be much better than chance. The caveats are that the data is noisy: for instance, Mouse S has anticorrelation slightly decrease compared to the beginning of overtraining. Such limitations of the data emphasize that our findings are *suggestive* rather than conclusive, and we aim to inspire further experimental work and pose our findings as *hypotheses* that this data suggest are *plausible* rather than definitively validated.

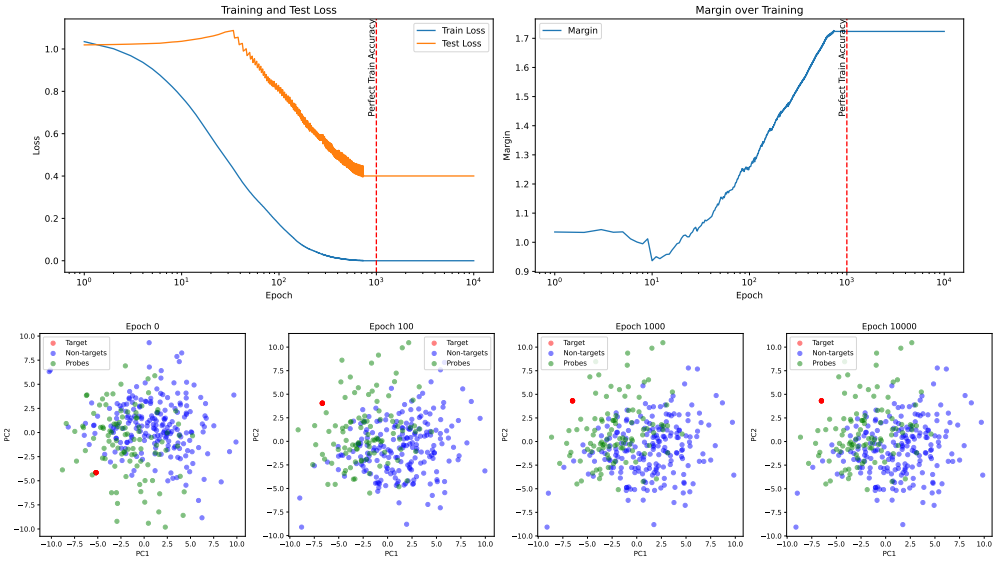


Figure 9: Loss function ablation for synthetic model that traces asymptotic margin maximization induced by cross-entropy as the cause for late-time decrease in test loss/late-time separation of representations. "Late-time" refers to when training loss plateaus at a low value.

A.6 Hard Margin Ablations

We begin with a review of the Cross-Entropy vs Hinge loss definitions. Let $y \in \{0, 1\}$ represent the true class label, and $\hat{y} \in [0, 1]$ represent the predicted probability for the positive class (i.e., $P(y = 1 \mid x)$). The cross-entropy loss L_{CE} for a single example is defined as:

$$L_{\text{CE}}(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})).$$

In hinge loss, the labels y are typically encoded as $y \in \{-1, +1\}$, and the model outputs a score $f(x) \in \mathbb{R}$, where the sign of $f(x)$ determines the class prediction. For simplicity, let's assume $f(x)$ is the output of a linear model, $f(x) = w \cdot x + b$.

The hinge loss L_{hinge} for a single example is defined as:

$$L_{\text{hinge}}(y, f(x)) = \max(0, C - y \cdot f(x)),$$

where C is the hard margin parameter to be specified, which we take as $C = 1$. The key point is that in a Hinge objective, loss vanishes after the learned classifier classifies each training example with margin at least C, whereas loss does not vanish in this discontinuous way for the Cross-Entropy objective, driving continued late-time learning. We can see in Figure 9that test loss stops improving when train loss does, and this is coincedent with an abrupt ceasation in the increase in margin, in contrst to the continued margin maximization driven by Cross-Entropy in the main text.

A.7 Biologically Plausible Model

We now repeat the above experimental setup of our synthetic model with an architecture that takes piriform cortex more seriously (Stern et al., 2018; Krishnamurthy et al., 2022; Mathis et al., 2016).

Our biologically plausible model consists of three main processing layers plus an output layer, with additional feedback connections to model recurrent dynamics. The first layer serves as a sensory input layer, analogous to inputs from the olfactory bulb. It projects the input to a higher-dimensional space, using ReLU activation followed by dropout with probability 0.2 to maintain sparse coding, a key feature of sensory representations in the piriform cortex. The second layer acts as a dense associative layer, modeling the extensive recurrent connectivity between pyramidal cells. This layer maintains the expanded dimensionality from layer one and includes a parallel feedback pathway. The feedback is implemented through a separate linear transformation followed by ReLU activation,

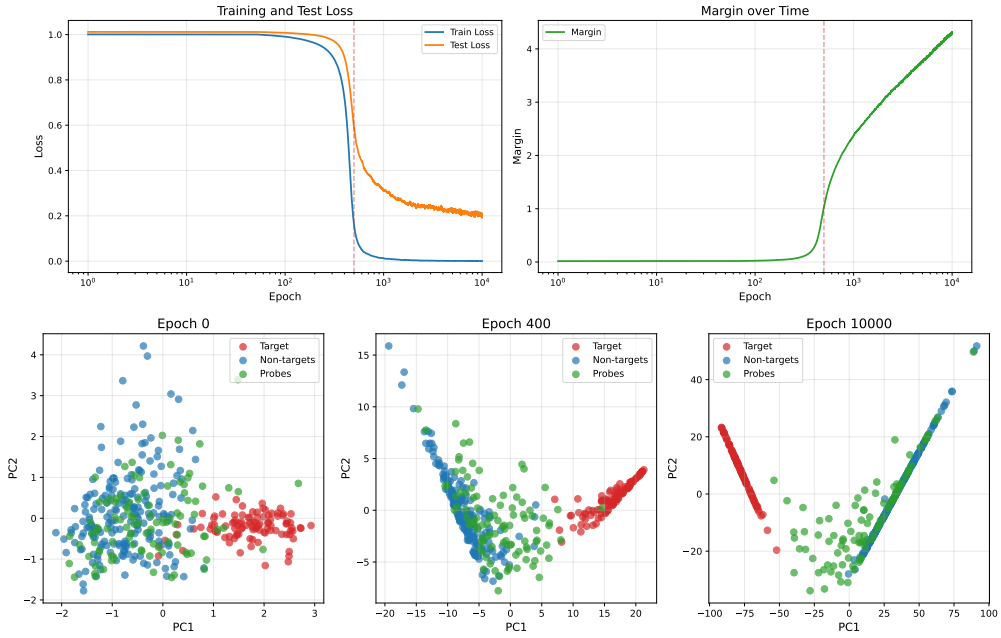


Figure 10: Architecture ablation for biologically plausible model of piriform cortex trained on the same binary classification task to separate target and nontarget odors. We see loss on held out probe odors continues to decrease after low train loss is achieved, since cross-entropy loss is once again used as the objective.

and the result is added back to the main pathway via a skip connection. Both the main and feedback pathways use the same dimensionality to preserve the associative nature of this layer. The third layer implements modulatory control, inspired by inhibitory interneuron circuits. This layer reduces dimensionality compared to layer two and uses LeakyReLU activation followed by a higher dropout rate (0.3) to model inhibitory effects. The reduction in dimensionality reflects the modulatory rather than representational role of this layer. The output layer performs a final linear transformation to a single scalar value (since we're training on binary classification). We train without weight decay and learning rate 1e-2. We also add Gaussian noise to the forward pass to simulate stochastic activity in sensory cortex, hence why there are multiple target odor representations.

Results are shown in Figure 10. The persistence of the phenomenology identified in the main text suggests that the objective function and task structure, rather than architectural details, are the key here.