Object manifold geometry across the mouse cortical visual hierarchy

Emmanouil Froudarakis[1,4,5,*], Uri Cohen[2], Maria Diamantaki[1], Edgar Y. Walker[4,5,6], Jacob Reimer[4,5], Philipp Berens[5,6,7], Haim Sompolinsky[2,3], and Andreas S. Tolias[4,5,8,*]

[1]Institute of Molecular Biology and Biotechnology,
Foundation for Research and Technology, Hellas, Heraklion, Greece
[2]Edmond and Lily Safra Center for Brain Sciences,
Hebrew University of Jerusalem, Israel
[3]Center for Brain Science,
Harvard University, Cambridge, MA, USA
[4]Department of Neuroscience,
Baylor College of Medicine, Houston, TX, USA
[5]Center for Neuroscience and Artificial Intelligence,
Baylor College of Medicine, Houston, TX,USA
[6]Institute for Ophthalmic Research,
University of Tübingen, Germany
[7]Department of Computer Science,
University of Tübingen, Germany
[8]Department of Electrical and Computer Engineering,
Rice University, Houston, TX, USA

Abstract

Despite variations in appearance we robustly recognize objects. Neuronal populations responding to objects presented under varying conditions form object manifolds and hierarchically organized visual areas untangle pixel intensities into linearly decodable object representations. However, the associated changes in the geometry of object manifolds along the cortex remain unknown. Using home cage training we showed that mice are capable of invariant object recognition. We simultaneously recorded the responses of thousands of neurons to measure the information about object identity across the visual cortex and found that lateral areas LM, LI and AL carry more linearly decodable object information compared to other visual areas. We applied the theory of linear separability of manifolds, and found that the increase in classification capacity is associated with a decrease in the dimension and radius of the object manifold, identifying the key features in the geometry of the population neural code that enable invariant object coding.

*To whom correspondence should be addressed; E-mail: frouman@imbb.forth.gr, astolias@bcm.edu

1 Introduction

Object recognition is an ethologically-relevant task for many animals. This is a challenging problem because an individual object can elicit myriads of images on the retina due to so-called nuisance transformations such as changes in viewing distance, projection, occlusion and illumination. The collection of neural responses associated with a single object is known as the object manifold. A prevailing hypothesis is that along the visual hierarchy, object manifolds are gradually untangled to produce increasingly invariant object representations, which are linearly decodable [1]. This hypothesis is primarily based on work in non-human primates, which is a powerful model to study object recognition especially given the similarities in visual perception among primates. These studies have revealed that the selectivity for object identity increases as visual signals are conveyed from primary visual cortex (V1) to inferotemporal cortex [2] [3]. Despite this significant progress, the underlying changes in the geometry of the object manifolds along the visual cortical hierarchy that leads to object recognition and a circuit-level mechanistic understanding of how they are generated remain largely unknown. The mouse animal model is ideally suited to dissect circuit mechanisms due to its genetic tractability and the numerous methods available to perform large scale recordings, manipulations and anatomical tracing with cell-type precision [4] [5]. Therefore developing visually guided behaviors in rodents is important [6] and identifying the relevant network of visual areas involved in object recognition analogous to the ventral stream of primates is critical. In this direction, we developed a novel automatic high-throughput training paradigm and demonstrated that mice can be trained to perform a two-alternative forced choice (2AFC) object classification task, which is typically used in primates to test object identification. While visually-guided operant behavioral tasks have been used previously in mice [7] [8] [9], here we show that mice can also learn to correctly discriminate objects under a 2AFC paradigm. Critically, this capability persisted even when they were presented with previously-unseen transformation of objects demonstrating that mice are capable of invariant object recognition.

To systematically study how objects are encoded in the mouse visual system, we simultaneously recorded the activity of thousands of neurons across all cortical visual areas of the mouse: primary (V1), anterolateral (AL), rostrolateral (RL), lateromedial (LM), lateral intermediate (LI), posteromedial (PM), anteromedial (AM), posterior (P), postrhinal (POR) and laterolateral anterior (LLA) visual areas, while presenting images of moving objects undergoing numerous identity-preserving transformations such as rotation, scale and translation across different illumination conditions. By decoding the identity of the objects from the recorded neural activity using a linear classifier, we found that the lateral extrastriate visual areas (LM, AL, LI) carried more linearly decodable information about object identity compared to V1 and all other higher order areas we studied. The large scale recordings provide the opportunity to study the changes in the geometry of object manifolds along the cortex associated with invariant object coding. We applied the recently developed theory of linear separability of manifolds to our neural recordings and found that in areas LM and AL the increase in classification capacity is associated with improved manifold geometry, where both the manifold radii and dimensions are reduced compared to other visual areas. Additionally, by recording simultaneously from many visual areas, we found that the population dynamics differed across the visual hierarchy, with information about object identity accumulating faster in areas that were more object selective compared to V1.

2 Results

2.1 Mice are capable of invariant object recognition

We generated movies of 3D objects by varying their location, scale, 3D pose and illumination in a continuous manner across time (**Fig. 1a**, **Supp. Movie 1**). We developed a 2AFC automatic home cage training system in which water restricted mice had to lick a left or a right port depending on the object that was shown on a small monitor in front of their cage (**Fig. 1b**). Upon a correct choice, animals immediately received a small amount of water reward. Naive animals initially licked the left and right probes at random, but within two weeks of training, animals learned to preferentially lick the correct port matched to object identity (**Fig. 1c**); trained animals maintained consistent performance on the task across days (**Fig. 1d**). An important property of object recognition is the ability to generalize across views of objects that have never been seen before. After the animals learned to discriminate objects from the movie clips - which contained a specific set of object transformations, new movie clips with unique parameters across translation, scale, pose and illumination were presented to the animals. We could not detect any differences in performance between the previously seen object transformations (**Fig. 1e**, familiar transformations) and novel object transformations (**Fig.1e**, novel transformations). This ability to generalize across identity-preserving transformations indicated that mice learned an internal object-based model and did not rely simply on low-level features of the rendered movies they observed during training. Importantly, a linear classifier trained on the pixel intensities of the rendered movies performed at chance level (**Fig. 1e**), indicating that the discrimination of these objects cannot be simply solved using low-level strategies based on pixel intensity differences between the images.

If mice are capable of discriminating between objects, there should exist a set of areas along their visual processing hierarchy that can extract this information. It has been suggested that one way of extracting the object information irrespective of its transformations is to have neural representations for each object that are untangled, i.e. can be read-out using a linear decoder [1]. To test this idea, we used transgenic mice expressing GCamp6s in pyramidal neurons and recorded the activity from hundreds of neurons in each visual areas separately or from thousands of neurons across the whole visual cortical hierarchy of the mouse using a large field of view microscope ([5], **Fig. 1f, g**), while the animals passively viewed the moving objects (**Fig. 1a**). We identified the borders between visual areas using wide-field retinotopic mapping as previously described [10] [11] [12] (**Fig. 1f**, **Materials and Methods**). Neurons in all of the identified visual areas showed significantly more reliable responses when compared to neurons that were not assigned to any visual area (**Supp. Fig. 1a**).

2.2 Lateral visual areas carry more linearly decodable object identity information

To measure how linearly discriminable the responses to the different objects were, we used cross-validated logistic regression to classify the object identity from the responses of neurons in each visual area. As expected, discriminability increased as a function of the number of neurons sampled (**Fig. 2a**), but only the higher visual areas, LM, LI and AL, showed consistently higher discriminability levels compared to V1 responses (**Fig. 2a, b, c**). In contrast, areas RL, AM, P, POR and LLA had significantly lower discriminability levels when compared to V1 and this effect was independent of the number of objects (**Fig. 2b**). The differences in decoding between these areas persisted at the single neuron as well (**Fig. 2d, Supp. Fig. 2a**).

We performed several control analyses: first, our results might be due to differences in the retinotopic coverage across areas. As has been reported before [11], the coverage of the visual

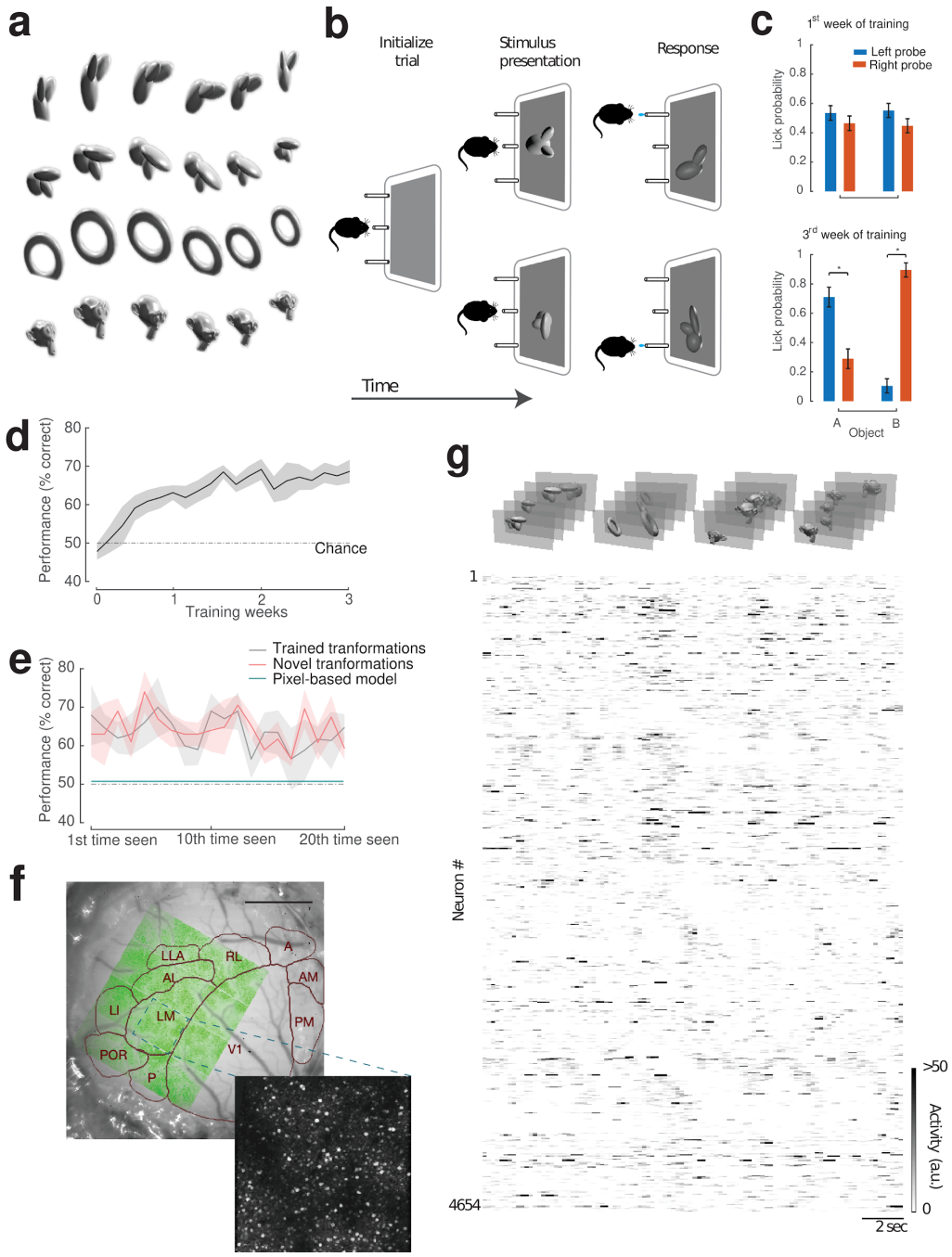


Figure 1: Experimental procedure for behavioral training and two-photon imaging. (a) Single frames from movies with the objects that were presented to the animals. (b) Behavioral training sequence. (c) Probability of licking either probe during the early training period (upper bar plot) and later training period (lower bar plot) for 1 animal. Error bars represent S.E.M. Student t-test * $p < 0.05$ (d) Performance as a function of training time, N = 8 animals. (e) Performance across repetitions of previously seen (gray) and previously unseen (red) object trajectories during one session. N = 6 animals. Gray dashed line represents chance level. Green line represents performance of a pixel based linear classifier. For both (d) and (e) shaded areas represent S.E.M. (f) Example large field of view recording (green) with area boundaries overlaid. Scale bar represents 1mm. A small inset depicts the two-photon average image for a small segment of the large field of view captured with the mesoscope. (g) Example responses of all neurons to moving objects (shown on top) from the recording shown in (f). Each clip is presented for 3-5 seconds before a short pause switches to a new clip that might be the same or a different object identity.

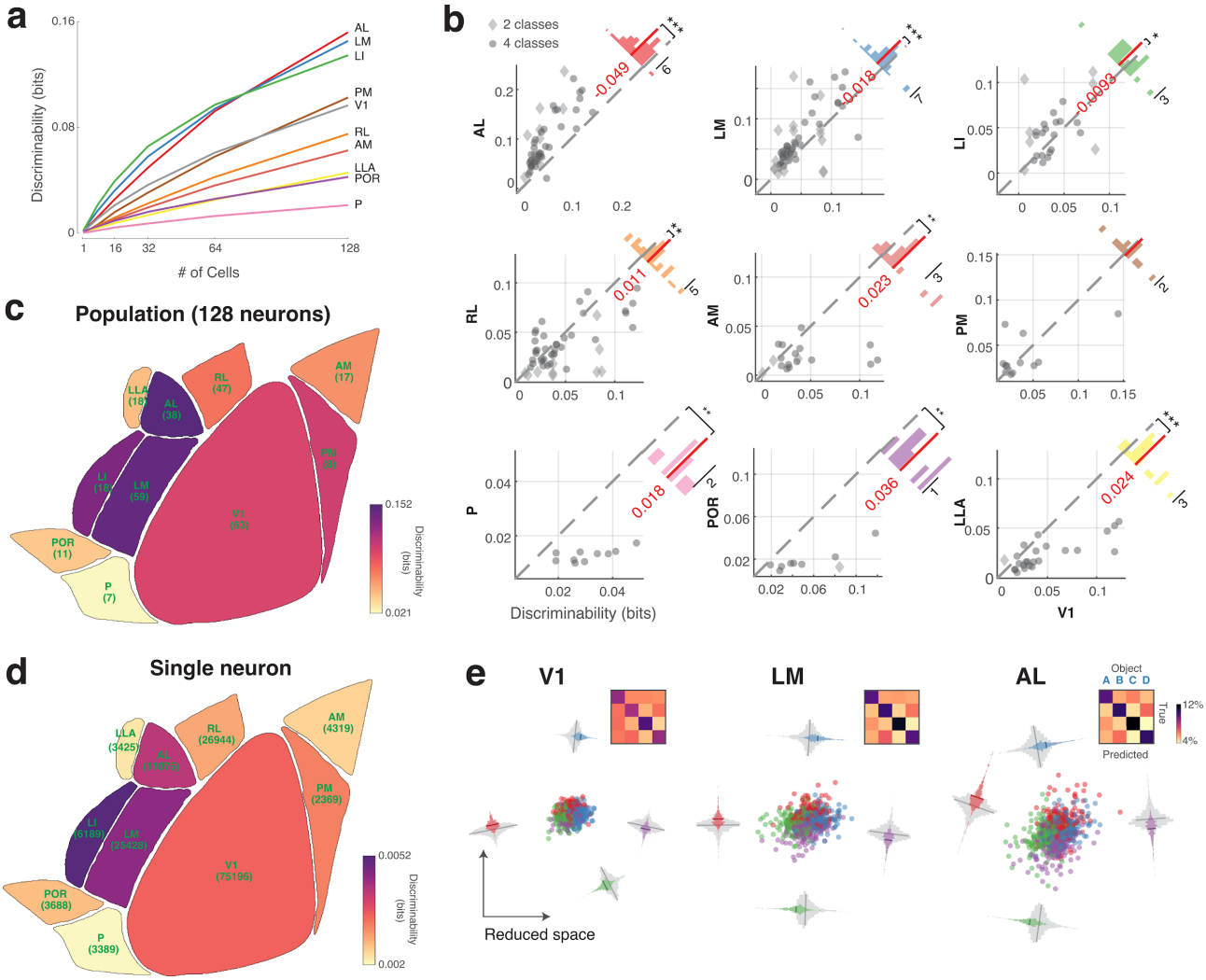


Figure 2: Object identity decoding across the visual hierarchy. (a) Discriminability of object identity as a function of the number of neurons sampled. Each line represents the average across all recorded sites. (b) Scatter plot of the discriminability of different areas with a population of 128 neurons compared to V1 for all the recording sites. Insert histogram represents the difference between the discriminability of each area and V1. Red line and number indicate the mean difference. Diamonds represent the results with 2 objects whereas circles represent the results with 4 objects. Outliers have been omitted for better visualization. Wilcoxon signed rank test *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. (c) Average discriminability of all visual areas with a population of 128 neurons. The number below each area represents the recording sites sampled. (d) Same as in (c) but when using a single neuron at a time to decode the object identity. The number below each area represents the cells sampled. (e) Low-dimensional representation of the 128-dimensional neural activity space, illustrating the separation of the responses to four different objects for three example areas. Each dot represents the average of the activity in one 500msec bin. The side histograms represent the distances of the data projected onto each of the four object category axes for the same-class (colored) and different-class (gray). Each insert represents the confusion matrix after decoding.

field is different between visual areas. To control for this, in some experiments we also mapped the receptive fields (RF) of all recorded neurons using a dot stimulus (see **Materials and Methods**) and repeated our decoding analysis using only neurons from each area with RF centers within the same ∼20 degree area of visual space. When we restricted our analysis in this way, areas LM, LI and AL still showed significantly higher discriminability (**Supp. Fig. 2b**). Another potential confound might be differences in receptive field sizes across areas [13] [12]. An area with larger receptive fields might be better at representing objects simply because more neurons are responding to the object at any moment. Indeed, when we examined decoding performance conditioned on the object size, we observed an increase in discriminability for all visual areas as a function of object size (**Supp. Fig. 3a**), in agreement with the increased performance we found when sampling from more neurons (**Fig. 2a**). However, if changes in receptive field size alone are responsible for increased object discriminability, we would expect that area PM, which has very large receptive fields [13], would also have high object discriminability. This was not the case in our data (**Fig. 2**). To further investigate the influence of receptive field size on discriminability, we modeled the effect of changing receptive field (RF) size in a simulated population of neurons using either pixel intensities or the output of filters learned by a sparse coding model of natural images [14]. Increasing the size of the receptive fields by either scaling or pairwise linearly combining them (**Supp. Fig. 4a**, see **Materials and Methods**) led to either a decrease in discriminability or had no significant effect, respectively (**Supp. Fig. 4b**). Using pixels as an input to the model also resulted in low discriminability irrespective of the number of pixels used (**Supp. Fig. 4b,c**). These results argue that our *in vivo* results cannot be trivially explained by a simple pixel model or differences in the receptive field sizes across visual areas. Additionally, higher visual areas have been reported to have different temporal frequency selectivities [15] [8] [16]. To determine whether the range of speeds that objects were moving in the movies that we showed influenced our results, we computed the decoding performance for each area as a function of the object speed, but did not find any significant differences (**Supp. Fig. 3b**). Therefore, we interpret the increase in discriminability in AL, LI, and LM indicating that these visual areas are particularly involved in the processing of visual object information with neural representations that are easier to decode (**Fig. 2e**).

2.3 Lateral visual areas show responses that are more invariant to nuisance transformations

An important property of visual areas that extract information about object identity is generalization to out-of-distribution data, such as adding background clutter to the stimuli. To assess the effect of background clutter, we first trained a logistic regression decoder on responses to objects with only a gray background as previously used. We then evaluated the performance of the decoder on the responses to movies in which we embedded the objects on top of background clutter (**Fig. 3a**, **Supp. Movie 2**). While the discriminability decreased for all visual areas when compared to noise-free stimuli (**Fig. 3b**), areas LM and AL maintained significantly higher discriminability compared to V1 and all other visual areas (**Fig. 3a,b,c**), indicating that in addition to being highly invariant to changes in the appearance of the object, the object representation in these areas is also more robust than in V1 and other visual areas to clutter. We also studied the relationship between discriminability and reliability of the neural responses. Although the decoding performance of the objects without the background correlated well with the reliability of the responses for both V1 and the lateral visual areas, when background noise was introduced this relationship broke down for V1 but not for the lateral visual areas (**Supp. Fig. 1b**).

An additional test of invariant object recognition is the ability of the neural representation to

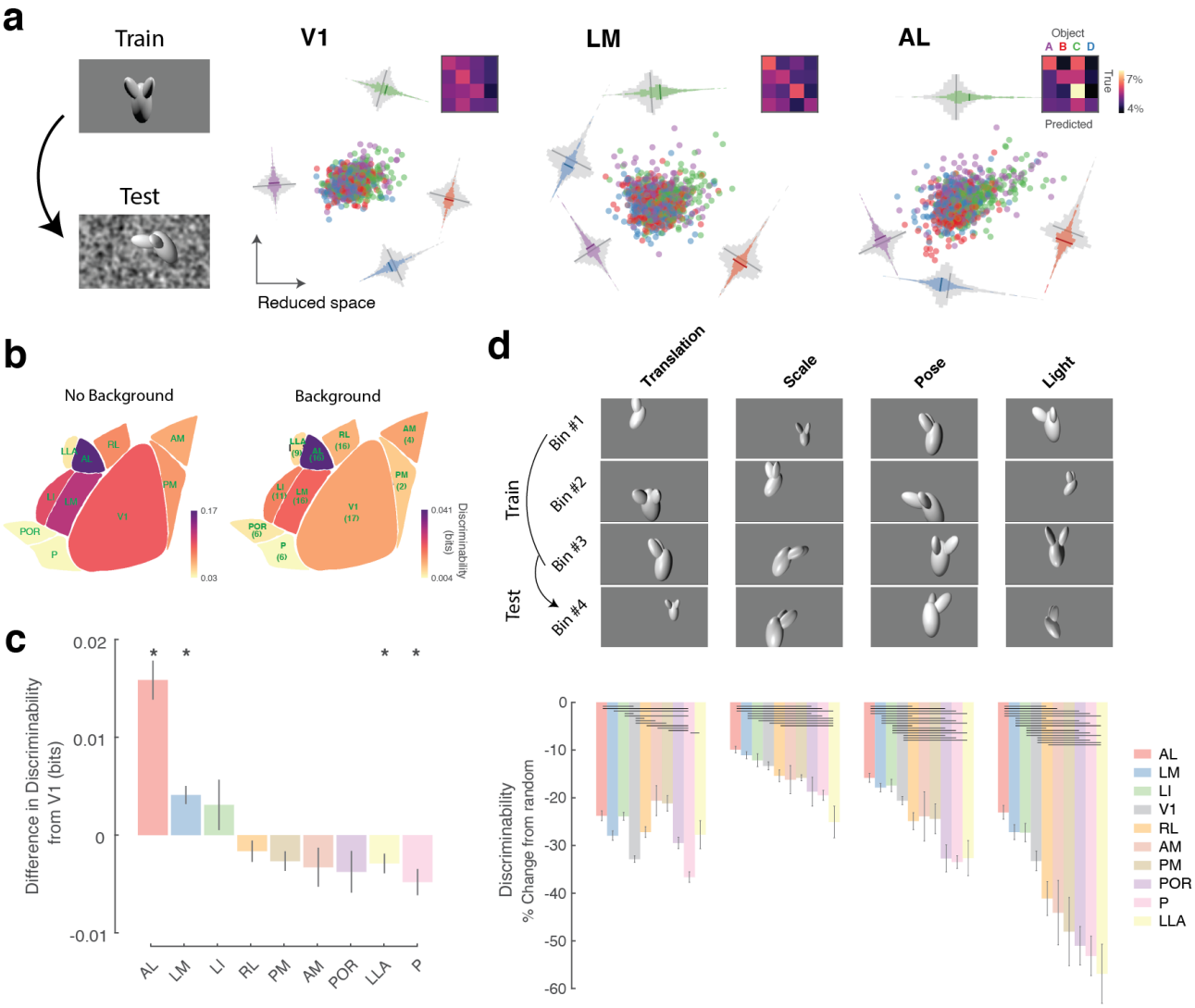


Figure 3: Generalization performance across background noise and identity-preserving transformations. (a) Generalization test across background noise. The decoder was trained on the responses to objects without background and tested on the responses to objects that contained background noise. Low-dimensional representation of the responses to the object w/ background are shown on the right similar to Figure 2e. Each insert represents the confusion matrix after decoding. (b) Average discriminability of all visual areas for objects w/o and w/ background, on the same recorded sites. (c) Bar plot indicating the difference in discriminability between all visual areas and V1 on the responses to objects w/ background. Kruskal-Wallis with multiple comparisons test $*p < 0.001$. (d) Top: Example parameter space of the four nuisance classes: Translation (x/y), Scale, Pose (tilt/rotation) and Light (four light sources). The decoder was tested on a parameter space of each of the four nuisance variables that had not been part of the training set. Bottom: Bar plot indicating the performance when testing on untrained parameter space, compared to the performance of the random sampling across all classes. Lines indicate $p < 0.05$ Kruskal-Wallis with multiple comparisons test.

generalize across new image transformations — not used during training — that preserve object identity such as changes in position, scale, pose and illumination [1] [17] [3] [18]. Specifically an object decoder built on a subset of the nuisance parameter space, i.e. a limited range of translations, sizes, and rotations, should generalize across new nuisance parameters. To test this we split the data into four non-overlapping bins for each of the nine continuously-varying parameters that defined the object stimulus (for example for the size object parameter into very small, small, medium, and large objects; **Fig. 3d top**), while the remaining parameters were randomly sampled. For each parameter, we then used data from three of the bins to train the decoder, and tested the prediction performance on the held-out data bin. We compared this performance to a baseline discriminability using a 4-fold cross validation, when the values for each parameter were randomized before binning so that the training and test set both spanned the same parameter range. Comparing the out-of-distribution test set performance to this baseline allowed us to assess the ability of the decoder to generalize, and thus the invariance of the representation in each area (**Fig. 3d bottom**; negative values). Areas AL, LM and LI consistently showed the best generalization performance (smallest reduction in performance for out-of-distribution test set vs baseline), when changing scale, pose and light (**Fig. 3d bottom**). Interestingly, that was not true for translation. The larger receptive field sizes of areas PM and AM [13] [12] might contribute to the improved translation invariance that we observed relative to the other parameters.

2.4 Changes in the geometry of object manifolds along the cortical hierarchy

Chung and colleagues [19] recently developed the theory of linear separability of manifolds and defined a measure called the classification capacity which quantifies how well a neural population supports object classification. The classification capacity measures the ratio between the number of objects and the size of the neuronal population that is required for reliable binary classification of the objects, and is tightly related to the geometry of a neuronal population responding to an object presented under varying nuisance transformations with respect to the identity of the object (object manifold). In deep neural networks trained on object classification tasks, it has been shown that the classification capacity improves along the network's processing stages [20]. Our data, consisting of responses of large neuronal populations in different visual areas to objects under various transformations, are well suited for applying this method to characterize the object manifolds in different visual areas. We used the neuronal responses of 128 simultaneously recorded neurons from each visual area to four objects under the identity-preserving transformations introduced earlier (object position, scale, pose and illumination conditions, with and without background noise). In agreement with our decoding results, we found that the classification capacity increased in higher visual areas AL and LM compared to V1, but decreased in the rest of the areas (**Fig. 4a, b**). The theory of linear separability of manifolds [19] also enabled us to characterize the associated changes in the geometry of the object manifolds to understand how object invariant representations arise along the processing hierarchy [20] (i.e. relate the manifolds' classification ability to the geometry of object manifolds). In particular, classification capacity depends on the overall extent of variability across the encoding dimensions, the radius of the manifold, but also the number of directions in which this variability is spread, the dimension of the manifold. These geometric measures influence the ability to linearly separate the manifolds (**Fig. 4c**). In our results, we find that the increase in classification capacity can be traced to changes in the manifolds' geometry, both as a decrease of the dimension and radius of object manifolds (**Fig. 4d**).

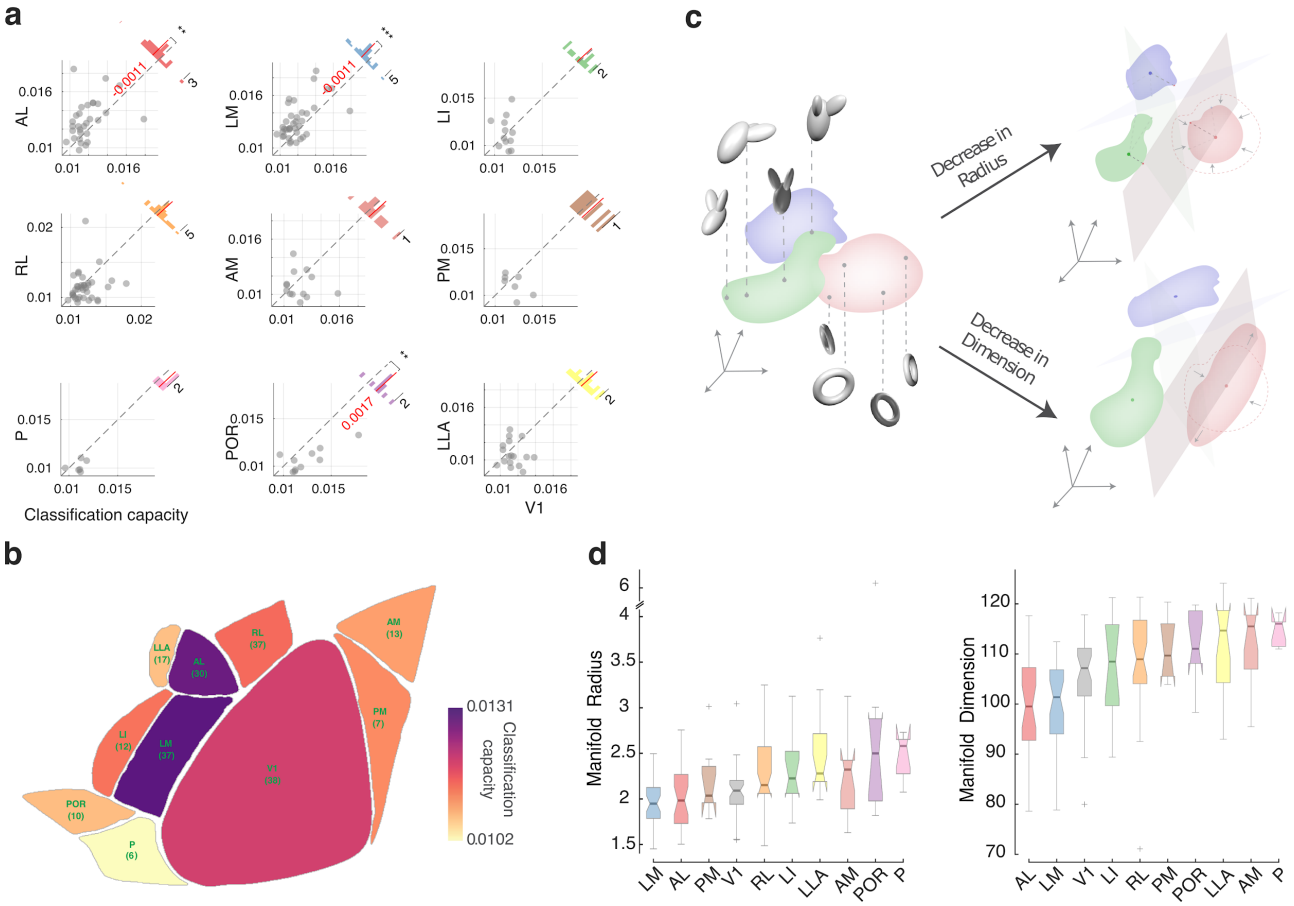


Figure 4: Classification capacity and geometry of manifolds across the visual hierarchy. (a) Scatter plot of the classification capacity of different areas compared to V1 for 4 objects. Insert histogram represents the difference between the classification capacity of each area and V1. Red line and number indicate the mean difference. Wilcoxon signed rank test *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. (b) Average classification capacity of all visual areas with a population of 128 neurons. The number below each area represents the recording sites sampled. (c) Illustration of low dimensional representations of object manifolds for two visual areas. Left: each point in an object manifold corresponds to neural responses to an object under certain identity-preserving transformations. Right: demonstration of two possible changes in the manifold geometry in a higher order area, reduction of the radius of one manifold through reduction of its extent in all directions (top) and reduction of the dimension of one manifold by concentrating variability at certain elongated axis, reducing the spread along other axes. Such changes have predictable effects on the ability to perform linear classification of those objects. (d) Box plots of the manifold radius (left), and manifold dimension (right) of all areas, sorted in ascending order of the median value.

2.5 Temporal dynamics and cross-area dependencies

One question that arises is how these visual areas are able to form invariant representations that can generalize across background noise or nuisance parameters. One way for these areas to optimize the representations is by taking advantage of the temporal continuity that exists for natural objects by integrating information over time [21] [22]. We analyzed the temporal dynamics of the decoding performance of random samples of 50 simultaneously recorded neurons for objects overlaid on background noise. From one trial to the next the nuisance parameters varied continuously but the object identity was preserved (cis trials) or switched (trans trials) (**Fig. 1g**). When we compared the discriminability as a function of time for cis/trans trials, we found that indeed in the trials in which the identity of the object was switched (trans trials), discriminability was overall lower across all visual areas in the early phase of the trials compared to the late phase of the trials, providing evidence for temporal integration during a trial (**Supp. Fig. 5**). In the late period discriminability in area AL was significantly closer to the discriminability levels of the cis trials than all other visual areas, suggesting that activity in AL more quickly evolved to more disentangled representations (**Supp. Fig. 5b**, Early/Late). We also studied the correlations between the representations of objects across multiple visual areas. If information about object identity propagates across areas, then we expect to find significant temporal correlations in the evolution of object discriminability across these areas. We estimated each area's confidence about the identity of the object at each time point, as the distance of the population activity from the decision boundary (**Fig. 5a**), and we examined the evolution of this metric across time in each area. Specifically, we estimated the distance to the decision boundary at different moments within the trial for the class that was presented. This decision boundary was a linear hyperplane in the 128 dimensional neural activity space (**Fig. 5a**). We then computed the correlation between the resulting temporal vectors of the score values across all simultaneously recorded visual areas (**Fig. 5b**, Score Correlation). The highest correlations in this moment-to-moment discriminability score were between AL, LM, RL and V1 (**Fig. 5c**).
Given that activities of neurons across areas can co-fluctuate because of global brain states, these score correlations could just be the result of raw activity correlations across areas. To test this we computed the activity correlations between the responses of pairs of neurons across visual areas. We observed a different correlation pattern that was distinct from the structure of the score correlation (**Fig. 5b**). Moreover, we measured the strength of the linear relationship between each pair of areas after adjusting for relationships with the rest of the areas. To this end, we computed the partial score correlations. The correlation pattern remained largely unchanged with strong dependencies between V1-LM, V1-RL and LM-AL suggesting that these areas work together as a network of areas specialized for object recognition (**Fig. 5c**). Interestingly, we did not find a strong relationship between V1-AL (**Fig. 5c**).

3 Discussion

The ability to recognize, discriminate, and track objects across time is a key adaptive trait that is fundamental to identifying food items or conspecifics [23]. The ability to recognize objects has been observed not only in higher mammals such as humans and monkeys, but also rodents, birds, fish and insects [24] [25] [26] [27] [28] [6]. While the implementation of how object information is extracted from the visual scene may vary across species, the computational problem remains the same: construct an invariant representation of objects under a wide range of identity-preserving transformations. While there is plenty of evidence that mice can detect novel objects [9], and that mice rely on their vision to hunt crickets [29], until our study there was no direct evidence that mice are capable of invariant object recognition.

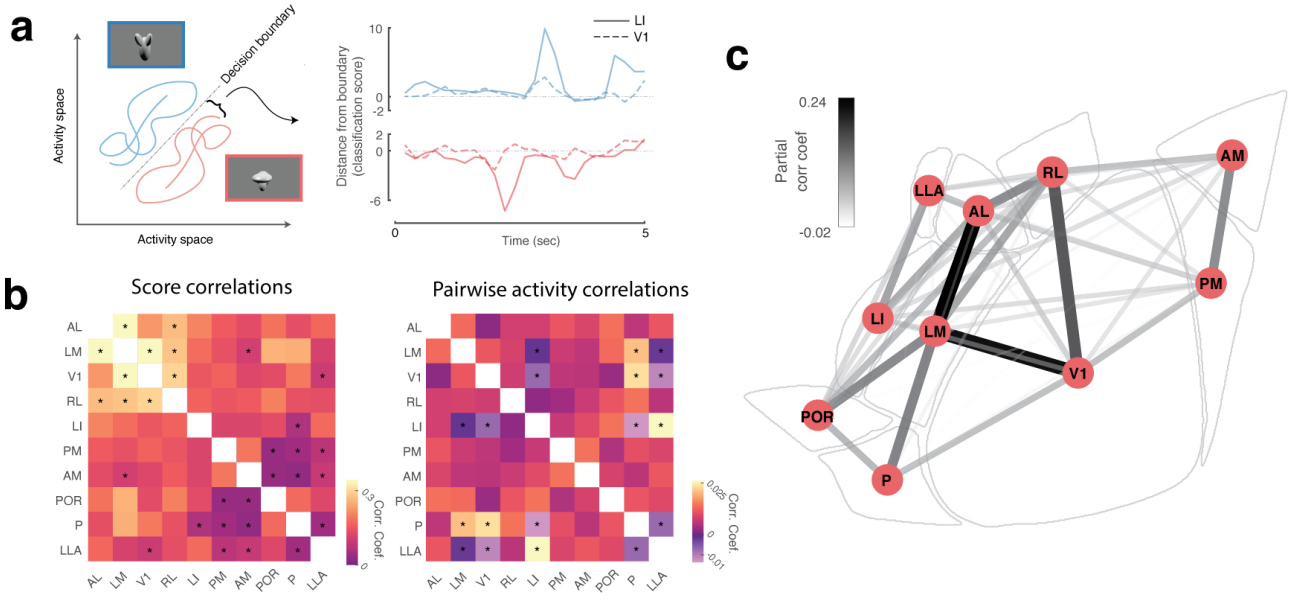


Figure 5: Temporal dynamics and cross-area dependencies. (a) Schematic representation of the classification scores as the distances of the response trajectories to the decision boundary (left) and their resulting temporal dependencies across different areas (right). (b) Score correlations across all recorded areas (left) and raw pairwise correlations of the single neuron activity between areas (right). Significance was estimated by bootstrapping across all correlations, * $p < 0.025/45$. (c) Schematic representation of the score partial correlation coefficients between areas.

In this work, we showed that mice can be trained to recognize unfamiliar objects in a 2AFC paradigm (**Fig. 1**). Similar tasks have been developed for rats [6], but mice have not been reported to perform such a task. That might be related to the fact that even though mice and rats can achieve similar performance levels, mice are slower to train [30]. Our unique training approach involves minimal interactions with the animals since the training system is part of their housing. Within a few weeks animals learn to discriminate objects and can show generalization across unseen objects poses and clutter establishing that mice are capable of invariant object recognition (**Fig. 1d**).

To identify how animals are able to extract object identity, we analyzed the activity of thousands of neurons of all known visual cortical areas of the mouse. We found that the decoding performance varied across the visual hierarchy where a set of lateral visual areas carried more linearly decodable information about the object identity. Importantly, these areas retained the information about object identity even in difficult visual conditions such as clutter and across new identity-preserving transformations despite that the linear classifiers were not trained under those conditions. Our results agree with the hypothesis that object representations become untangled and more linearly separable as information progresses through the visual hierarchy. This process might be beneficial as a simple readout mechanism can be employed to drive behavior. It is important to note that a biologically plausible readout mechanism could involve only from a small set of projection neurons in order to extract object identity. We found that information carried by single neurons also increased progressively across the hierarchy of V1-LM-LI in lateral visual cortex in agreement with electrophysiology studies in the rat [18] (**Supp. Fig. 2a**). Importantly, our richly varied object stimuli cannot be easily discriminated by a simple pixel model (**Fig. 1b, Supp. Fig. 4b, c**) and the ability of mice to separate such objects likely depends on more complex computations in cortical circuits. Therefore, analogous to primates, hierarchically organized visual areas in mice untangle pixel intensities into more linearly decodable object representations.

However, the associated changes in the geometry of the object manifolds along the visual cortex remained unknown. To this end, we characterized how the geometry of the object manifolds changed across the visual hierarchy, using the newly developed theory of linear separability of manifolds [19] [20]. We found that the two lateral visual areas LM and AL showed increased classification capacity with object manifolds becoming smaller and having lower dimensionality (**Fig. 4**). While the classification capacity and radius of object manifolds has not been previously quantified along the visual processing hierarchy, our results on the dimensionality of the neural population agree with previous work. Different methods have been used to quantify the dimensionality of the population responses which also showed that it decreases along the visual hierarchy of monkeys [31] [32]. However, critically the theory of the linear separability of manifolds differs from these previous methods as it quantifies the geometrical properties of the object response manifolds which contribute to the ability to perform linear decoding. This enabled us to determine that the dimension of the object manifold decreases from primary visual cortex to higher visual areas in a way which allows for linear decoding of objects using smaller number of neurons. The higher visual areas of the mouse [33] [12], have distinct spatio-temporal selectivities [15] [16] and project to different targets [34]. Based on these differences in their selectivities, projection and chemoarchitectonic patterns, efforts have been made to separate areas into ventral and dorsal pathways analogous to those described in primates [35] [34] [36] [37]. Specifically, areas such as LM, LI, P and POR areas are hypothesized to comprise the ventral stream whereas areas AL, RL, AM and PM comprise the dorsal stream. In rats, lateral visual areas LM, LI and LL have been shown to carry progressively more information about objects [18] [38] [39]. However, the areas of the mouse that might be involved in extraction of object information are not known. We found that higher visual areas AL, LM and LI had significantly more information about object identity than V1, with area AL consistently outperforming all other areas which is inconsistent with the current assumption that AL is part of a distinct dorsal pathway. Strong interactions between anatomically defined dorsal and ventral pathways in rodents might be particularly important for object detection and discrimination given the importance of navigation in rodents [40]. To that effect, both areas AL and LM show faster accumulation of information about object identity in noisy conditions (**Supp. Fig. 5**) that could result in the increased temporal stability that has reported recently in higher visual areas of the rat [41]. Moreover, the correlations we report in decoding confidence between areas AL and LM (**Fig. 5c**), could be the result of recurrent processes that have been suggested to play a significant role during object recognition [42] [43] [44]. These object-selective dependencies, particularly with area AL showing strong correlations with areas LM and LI, do not share the same structure as have been reported with more parametric stimuli [35], which could be due to objects having a statistical structure closer to the preferences of these lateral visual areas. Interestingly, area LI which is believed to be a high visual area did not consistently outperform areas LM and AL and had great variability in the discrimination levels (**Fig. 2b, 3c, 4a**). We also found that for larger populations, area LI carries less information than LM (**Fig. 2a, b**). This variability could be due to the fact that these experiments were done in awake passive viewing animals without a relevant behavioral task that would necessitate the engagement of LI.

Future experiments are required to determine how these different areas work together to extract information about objects that might be used to guide behavior. First, utilizing an ethologically-relevant task while mapping the activities across visual areas might provide even stronger evidence of hierarchy [29]. Second, in order to establish a more causal relationship between visual areas and behavior, it will be important to combine behavioral performance with causal manipulations of neural activity. Finally, neural networks models and the inception loop methodology will enable the characterization of the specific features that drive populations of neurons in these different visual areas [45] [46] [47].

In summary, we offer evidence that mice share similarities with other mammals in their ability to recognize objects. By recording the activity from ∼300000 neurons across the whole visual system of the mouse, in this paper we have deciphered for the first time for any species how object manifold geometry is transformed to become more separable thus identifying key features of the population code that enable invariant object coding. Given the panoply of tools available, the mouse has the potential to become a powerful model to dissect the circuit mechanisms of object recognition.

4 Acknowledgments

This work was supported by the the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003 (AST). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government. Also supported by R01 EY026927 (AST), NEI/NIH Core Grant for Vision Research (T32-EY-002520-37) and NSF NeuroNex grant 1707400 (AST). HS is partially supported by the Gatsby Charitable Foundation, the Swartz Foundation, the National Institutes of Health (Grant No. 1U19NS104653) and the MAFAT Center for Deep Learning. PB is supported by the the Deutsche Forschungsgemeinschaft (DFG, BE5601/4, SFB 1233 "Robust Vision" 276693517, Cluster of Excellence 2064 "Machine Learning - New Perspectives for Science" 390727645) and the German Ministry for Education and Research (FKZ 01GQ1601, 01IS18039A).

5 Author contributions

EF: Conceptualization, Methodology, Validation, Software, Data Curation, Formal Analysis, Investigation, Writing - Original Draft, Visualization, Supervision, Project administration; **UC**: Methodology, Validation, Formal Analysis, Writing - Original Draft; **MD**: Investigation, Validation, Formal Analysis, Writing - Review & Editing; **EYW**: Software; **JR**: Methodology, Investigation, Writing - Review & Editing; **PB**: Validation, Writing - Review & Editing; **HS**: Validation, Writing - Review & Editing, Supervision, Funding acquisition; **AST**: Conceptualization, Methodology, Validation, Supervision, Funding acquisition, writing - Review & Editing

6 Materials and Methods

Animal preparation and two photon imaging All procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of Baylor College of Medicine. We used 25 adult mice expressing GCaMP6s in excitatory neurons via either SLC17a7-Cre, Dlx5-Cre, Ai75, Ai148, Ai162 or CamKII-tTA transgenic lines. Animals were initially anesthetized with Isoflurane (2%) and a ∼4mm craniotomy was made over the right visual cortex as previously described [48]. The animals were head-mounted above a cylindrical treadmill and calcium imaging was performed using Chameleon Ti-Sapphire laser (Coherent, Santa Clara, CA) tuned to 920 nm. We recorded calcium traces by using either a large field of view mesoscope [5] equipped with a custom objective (0.6 NA, 21mm focal length) with a typical field of view of ∼2500x2000μm, or a two-photon resonant microscope (Thorlabs, Newton, NJ) equipped with a Nikon objective (1.1 NA, 25X) with a typical field of view or ∼500x500μm. Laser power after

the objective was kept below ∼60mW. We recorded data from depths of 100–380 μm below the cortical surface. Imaging was performed at approximately ∼5-12Hz for all scans. Imaging data were motion corrected, automatically segmented and deconvolved using the CNMF algorithm [49]; cells were further selected by a classifier trained to detect somata based on the segmented cell masks.

Behavioral training The mice are trained in a 2 alternative forced choice task in response to moving objects that are presented on a small 7" monitor that is located in front of their home cage. In total, 4 objects are used in the experiments, however it should be noted that 2 objects are presented to each mouse. The training procedure is illustrated in Figure 1. Briefly, naive water restricted mice are placed in a modified cage that has three ports and a monitor on one side of the box. The center port has a proximity sensor, and the two other ports on either side of the central port are used to detect licks and are coupled to a computerized valve-controlled liquid delivery that can deliver liquid volumes with 1uL resolution. The task is as follows: Mice initiate a trial by placing their snout in close proximity to the central port for ∼200-500msec. A stimulus that can be one of two objects is presented on the monitor that is ∼1.5" in front of the animal. The animal has to report the identity of the object by licking one of the side ports. Each port is allocated to the identity of the same object throughout the training. If the animal licks the correct port, then a small water reward ∼5-12μl is delivered almost immediately which the animals consume. A new trial can be started thereafter. If the animal licks the wrong port, a short delay 4-10 seconds is added and the screen turns black. A new trial can start after the delay. Animals have free access to food, and the most of the water that they receive comes from their training. The training periods in which animals can initiate tasks are restricted to 4-8 hours a day. At the start of the training animals are shown the same clip for each object that contains the same set of transformations. Once animals reach performance levels, new clips with unique transformations are added. At the end of their training they have seen between 10-20 unique 10s clips of unique object transformations. For the generalization test, at the start of a new session a whole new set of 10 clips are used for each object and the performance was compared to the session that preceded.

Receptive field mapping We mapped the location and size of the receptive fields of the neurons using black and white squares that each covered ∼8-10o of the visual field. The squares were presented across the entire range of the monitor in random order for 150-200 ms each. To map the receptive fields of individual neurons we averaged the first 500 ms of the activity of a cell across all repetitions of the stimulus for each location. We fit the resulting 2D map using an elliptic 2D Gaussian. For each neuron we computed the SNR as the ratio of the variance of this image within three SD of the receptive field center to the variance of the image outside of the three SD of the receptive field center.

Visual area identification We generated retinotopic maps of all the visual areas using wide-field imaging. The signals from GCamp6s were captured using either a custom epifluorescence setup or two-photon imaging. For the epifluorescence, brain was illuminated with a high power LED (Thorlabs) and the emitted signal was bandpass filtered at nm and captured at a rate of 10 Hz with a CMOS camera (MV1-D1312-160-CL, PhotonFocus, Lachen, Switzerland). For the two-photon retinotopic mapping we sampled the activity from a 2.4x2.4mm area with large field of view two photon microscope [5] at a rate of ∼5Hz. We stimulated with upward and rightward drifting white bars (speed: 9-18deg/sec, width: 10-20deg) on black background that had their size and speed constant relative to the mouse perspective as previously described. Additionally, within the bar we had drifting gratings with a direction opposite to the movement of the bar. Images from either the epifluorescent or the two-photon setups were analyzed by a

custom-written code in MATLAB to construct the 2D phase maps for the two directions. We used the resulting retinotopic maps to identify the borders and delineate the visual areas as previously described [11] [12].

Stimulus generation and visual stimulation In this study we used four synthesized three-dimensional objects that were rendered in Blender (www.blender.org). Two of the objects were built to match the objects used in [6] and the other two were already existing models within Blender. We varied the following parameters of the objects: X and Y location (Translation), magnification (Scale), tilt and axial rotation (Pose) and variation of either the location or energy of 4 light sources (Light). The different object parameters were varied continuously over time in order to generate a cohesive object motion. Objects were rendered either on a gray background, or on a gaussian noise pattern with a fixed seed between objects. The long rendered movie was split into smaller 10 second clips. A short 3-5 second segment from 150-380 clips for each object were presented in a random sequence to the left eye with a 25" LCD monitor positioned ∼15cm away from the animal. A small number of clips were repeated multiple times in order to estimate the reliability of the neural responses.

Model For the V1 model, we used the filter responses to a set of 256 localized and oriented filters obtained with Independent Component Analysis (ICA) on 12x12 patches randomly sampled from natural images from the van Hateren database as previously described [48] **Supp. Fig. 4a**. In order to control for the increase in size of the receptive fields in higher visual areas, we also created an enlarged version of these receptive fields (**Supp. Fig. 4a**, ICA 150%) by scaling the initial ICA filters. That method has the disadvantage of changing the spatial tuning properties and thus to create more realistic receptive fields, we linearly combined pairs of the ICA group depending on their response similarity to natural movies. This procedure created larger receptive fields that were more elongated and often had corner-like structure (**Supp. Fig. 4b**, ICA multi). As an additional control, we also used a 144x144 grid of pixels as filters when comparing with the behavioral data (**Fig. 1e**) and 32x32 when comparing to the neural data (**Supp. Fig. 4b, c**). The filter responses were half-wave rectified and squared and were used as a scale to sample from the Weibull distribution with a shape parameter optimized in order to create similar reliability levels to the in-vivo data. Finally, we used the resulting responses to train the same decoder as the in vivo data.

Decoding and discriminability We used a one-versus-all logistic regression classifier to estimate the decoding error between the neural representations of 2-4 objects of 200-500 ms scenes. Each scene was represented as an N-dimensional vector of neural activity for each response scene. In almost all of the cases we used a 10 fold cross-validation in which the performance of the decoder was tested on 10% of the data that were held out during training. When generalizing across the background noise in Figure 3, the decoder was trained on 90% of the data with the no-background objects and tested on 10% of the data with the background objects. For the generalization across object parameters in Figure 3 we used a 4 fold cross-validation in which the decoder was trained on 75% of the data, and tested on 25% with a unique parameter range. In order to compare the decoding accuracy of a one vs all decoder between experiments that had different number of objects (2 and 4), we converted the decoding error to discriminability, the mutual information (measured in bits) between the true class label c and its estimate, by computing

$$MI(c,\hat{c}) = \sum_i \sum_j p_{ij} log_2 \frac{p_{ij}}{p_{i.} p_{.j}}$$

where p_{ij} is the probability of observing true class i and predicted class j and $p_{i.}$ and $p_{.j}$ denote the respective marginal probabilities. Using the mutual information also doesn't have the nonlinear

Classification capacity and geometry of manifolds An object manifold is defined by the neuronal population responses to an object under different conditions (i.e. identity-preserving transformations). The ability of a downstream neuron to perform linear classification of object manifolds depends on the number of objects, denoted P, and the number of neurons participating in the representation, denoted N. *Classification capacity* denotes the critical ratio $\alpha_c = P/N_c$ where N_c is the population size required for a binary classification of P manifolds to succeed with high probability [19]. This capacity can be interpreted as the amount of information about object identity coded per neuron in the given population. Capacity α_c depends on the radius of each of the manifolds, denoted R_M, representing the overall extent of variability (relative to the distance between manifolds), and their dimension, denoted D_M, representing the number of directions in which this variability is spread. These geometric measures are defined through the alignment of the hyperplane (in the representation N-dimensional space) that separates positively labelled from negatively labelled manifolds. This hyperplane is uniquely determined by a set of *anchor points*, one from each manifold, that lie exactly on the separating plane. As the classification labels are randomly changed, the identity of the anchor points change; these changes, along with the dependence of the hyperplane orientation on the particular position and orientation of the manifolds, give rise to a statistical distribution of anchor points. Averaging the extent and directional spread of the anchor points with this distribution determines the manifolds radii and dimensions, respectively. Knowledge of manifold radius and dimension is sufficient to predict classification capacity using the relation $\alpha_c = \alpha_{Balls}(R_M, D_M)$ where α_{Balls} is a closed-form expression describing capacity of D-dimensional balls of radius R [19].
The separability of manifolds depends not only on their geometries but also on their correlations. For manifold classification with random binary labeling, clustering of the manifolds in the representational space, as expected for real-world object representations, hinders their separability, and the theory of manifold classification has been extended [20] to take these correlations into account in evaluating α_c.
Here we used the methods and code from [20] to analyze the geometry of the object manifolds (i.e. manifold radius and dimension) as well as estimate classification capacity of neuronal populations in the different cortical areas. As those methods depend on the correlation structure of the objects, we analyzed neural representations for data-sets of 4 objects (i.e. omitted data-sets where only 2 objects are available). At each session of simultaneously recorded neurons we have sub-sampled from the available population 128 neurons; the subsequent analysis was repeated 10 times with different choices of neurons, and we report the average results across this procedure. Each object manifold is defined by neural responses to an object at non-overlapping 500ms time windows, using the entire range of nuisance parameter space, as well as responses with and without background noise. This analysis was performed at each visual area for sessions where more than 128 neurons are available. The baseline to which classification capacity is compared is the value expected by structure-less manifold which is $2/M$, where M is the number of samples (i.e. time windows where the object was presented).

References

[1] Dicarlo James and Cox David. "Untangling invariant object recognition". In: *Trends in Cognitive Sciences* 11 (8 Aug. 2007), pp. 333–341. ISSN: 1364-6613. DOI: 10.1016/j.tics.2007.06.010. PMID: 17631409.

[2] Hung Chou et al. "Fast Readout of Object Identity from Macaque Inferior Temporal Cortex". In: *Science* 310 (5749 Nov. 2005), pp. 863–866. ISSN: 0036-8075. DOI: 10.1126/science.1117593. PMID: 16272124.

[3] Rust N and Dicarlo J. "Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT". In: *Journal of Neuroscience* 30 (39 Sept. 2010), pp. 12978–12995. ISSN: 0270-6474. DOI: 10.1523/jneurosci.0179-10.2010. PMID: 20881116.

[4] Fenno Lief, Gunaydin Lisa, and Deisseroth Karl. "Mapping Anatomy to Behavior in Thy1:18 ChR2-YFP Transgenic Mice Using Optogenetics". In: *Cold Spring Harbor Protocols* 2015 (6 June 2015), pdb.prot075598. ISSN: 1940-3402. DOI: 10.1101/pdb.prot075598. PMID: 26034299.

[5] Sofroniew Nicholas et al. "A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging". In: *eLife* 5 (June 2016). DOI: 10.7554/elife.14472. PMID: 27300105.

[6] Zoccolan D et al. "A rodent model for the study of invariant visual object recognition". In: *Proceedings of the National Academy of Sciences* 106 (21 May 2009), pp. 8748–8753. ISSN: 0027-8424. DOI: 10.1073/pnas.0811583106. PMID: 19429704.

[7] Nicole M. Procacci et al. "Context-dependent modulation of natural approach behaviour in mice". In: *Proceedings. Biological Sciences* 287.1934 (Sept. 9, 2020), p. 20201189. ISSN: 1471-2954. DOI: 10.1098/rspb.2020.1189.

[8] Han Xu, Vermaercke Ben, and Bonin Vincent. "Segregated encoding of spatiotemporal features in the mouse visual cortex". In: (Oct. 2018). DOI: 10.1101/441014.

[9] Leger Marianne et al. "Object recognition test in mice". In: *Nature Protocols* 8 (12 Nov. 2013), pp. 2531–2537. ISSN: 1754-2189. DOI: 10.1038/nprot.2013.155. PMID: 24263092.

[10] Fahey Paul et al. "A global map of orientation tuning in mouse visual cortex". In: *bioRxiv* 745323 (Aug. 2019). DOI: 10.1101/745323.

[11] Garrett M et al. "Topography and Areal Organization of Mouse Visual Cortex". In: *Journal of Neuroscience* 34 (37 Sept. 2014), pp. 12587–12600. ISSN: 0270-6474. DOI: 10.1523/jneurosci.1124-14.2014. PMID: 25209296.

[12] Wang Quanxin and Burkhalter Andreas. "Area map of mouse visual cortex". In: *The Journal of Comparative Neurology* 502 (3 2007), pp. 339–357. ISSN: 0021-9967. DOI: 10.1002/cne.21286. PMID: 17366604.

[13] Murgas Kevin et al. "Unique spatial integration in mouse primary visual cortex and higher visual areas". In: *J. Neurosci* 40 (May 2020), pp. 1862–1873. DOI: 10.1101/643007.

[14] J H van Hateren and A van der Schaaf. "Independent component filters of natural images compared with simple cells in primary visual cortex." In: *Proceedings of the Royal Society B: Biological Sciences* 265.1394 (Mar. 7, 1998), pp. 359–366.

[15] Andermann Mark et al. "Functional Specialization of Mouse Higher Visual Cortical Areas". In: *Neuron* 72 (6 Dec. 2011), pp. 1025–1039. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2011.11.013. PMID: 22196337.

[16] Marshel James et al. "Functional Specialization of Seven Mouse Visual Cortical Areas". In: *Neuron* 72 (6 Dec. 2011), pp. 1040–1054. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2011.12.004. PMID: 22196338.

[17] Hénaff Olivier, Goris Robbe, and Simoncelli Eero. "Perceptual straightening of natural videos". In: *Nature Neuroscience* 22 (6 Apr. 2019), pp. 984–991. ISSN: 1097-6256. DOI: 10.1038/s41593-019-0377-4.

[18] Tafazoli Sina et al. "Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex". In: *eLife* 6 (Apr. 2017). DOI: 10.7554/elife.22794. PMID: 28395730.

[19] Chung Sueyeon, Lee Daniel, and Sompolinsky Haim. "Classification and Geometry of General Perceptual Manifolds". In: *Physical Review X* 8 (3 July 2018). DOI: 10.1103/physrevx.8.031003.

[20] Cohen Uri et al. "Separability and Geometry of Object Manifolds in Deep Neural Networks". In: *Nature Communications* 11 (May 2019), p. 746. DOI: 10.1101/644658.

[21] Dicarlo James, Zoccolan Davide, and Rust Nicole. "How Does the Brain Solve Visual Object Recognition?" In: *Neuron* 73 (3 Feb. 2012), pp. 415–434. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2012.01.010. PMID: 22325196.

[22] Orlov Tanya and Zohary Ehud. "Object Representations in Human Visual Cortex Formed Through Temporal Integration of Dynamic Partial Shape Views". In: *The Journal of Neuroscience* 38 (3 Dec. 2017), pp. 659–678. ISSN: 0270-6474. DOI: 10.1523/jneurosci.1318-17.2017. PMID: 29196319.

[23] Jones A and Ratterman N. "Mate choice and sexual selection: What have we learned since Darwin?" In: *Proceedings of the National Academy of Sciences* 106 (Supplement_1 June 2009), pp. 10001–10008. ISSN: 0027-8424. DOI: 10.1073/pnas.0901129106. PMID: 19528643.

[24] Bevins Rick and Besheer Joyce. "Object recognition in rats and mice: a one-trial non-matching-to-sample learning task to study 'recognition memory'". In: *Nature Protocols* 1 (3 Oct. 2006), pp. 1306–1311. ISSN: 1754-2189. DOI: 10.1038/nprot.2006.205. PMID: 17406415.

[25] Blaser Rachel and Heyser Charles. "Spontaneous object recognition: a promising approach to the comparative study of memory". In: *Frontiers in Behavioral Neuroscience* 9 (July 2015). DOI: 10.3389/fnbeh.2015.00183. PMID: 26217207.

[26] Newport Cait, Wallis Guy, and Siebeck Ulrike. "Object recognition in fish: accurate discrimination across novel views of an unfamiliar object category (human faces)". In: *Animal Behaviour* 145 (Nov. 2018), pp. 39–49. ISSN: 0003-3472. DOI: 10.1016/j.anbehav.2018.09.002.

[27] Soto Fabian and Wasserman Edward. "Mechanisms of object recognition: what we have learned from pigeons". In: *Frontiers in Neural Circuits* 8 (Oct. 2014). DOI: 10.3389/fncir.2014.00122. PMID: 25352784.

[28] Werner Annette, Stürzl Wolfgang, and Zanker Johannes. "Object Recognition in Flight: How Do Bees Distinguish between 3D Shapes?" In: *PLOS ONE* 11 (2 Feb. 2016), e0147106. DOI: 10.1371/journal.pone.0147106. PMID: 26886006.

[29] Hoy Jennifer et al. "Vision Drives Accurate Approach Behavior during Prey Capture in Laboratory Mice". In: *Current Biology* 26 (22 Nov. 2016), pp. 3046–3052. ISSN: 0960-9822. DOI: 10.1016/j.cub.2016.09.009. PMID: 27773567.

[30] Santiago Jaramillo and Anthony M Zador. "Mice and rats achieve similar levels of performance in an adaptive decision-making task". In: *Frontiers in systems neuroscience* 8 (2014), p. 173.

[31] Brincat Scott et al. "Gradual progression from sensory to task-related processing in cerebral cortex". In: *Proceedings of the National Academy of Sciences* 115 (30 July 2018), E7202–E7211. ISSN: 0027-8424. DOI: 10.1073/pnas.1717075115. PMID: 29991597.

[32] Lehky Sidney et al. "Dimensionality of Object Representations in Monkey Inferotemporal Cortex". In: *Neural Computation* 26 (10 Oct. 2014), pp. 2135–2162. ISSN: 0899-7667. DOI: 10.1162/neco_a_00648. PMID: 25058707.

[33] Glickfeld Lindsey and Olsen Shawn. "Higher-Order Areas of the Mouse Visual Cortex". In: *Annual Review of Vision Science* 3 (1 Sept. 2017), pp. 251–273. ISSN: 2374-4642. DOI: 10.1146/annurev-vision-102016-061331. PMID: 28746815.

[34] Wang Q, Sporns O, and Burkhalter A. "Network Analysis of Corticocortical Connections Reveals Ventral and Dorsal Processing Streams in Mouse Visual Cortex". In: *Journal of Neuroscience* 32 (13 Mar. 2012), pp. 4386–4399. ISSN: 0270-6474. DOI: 10.1523/jneurosci.6063-11.2012. PMID: 22457489.

[35] Smith Ikuko et al. "Stream-dependent development of higher visual cortical areas". In: *Nature Neuroscience* 20 (2 Jan. 2017), pp. 200–208. ISSN: 1097-6256. DOI: 10.1038/nn.4469. PMID: 28067905.

[36] Wang Q, Gao E, and Burkhalter A. "Gateways of Ventral and Dorsal Streams in Mouse Visual Cortex". In: *Journal of Neuroscience* 31 (5 Feb. 2011), pp. 1905–1918. ISSN: 0270-6474. DOI: 10.1523/jneurosci.3488-10.2011. PMID: 21289200.

[37] Wang Q and Burkhalter A. "Stream-Related Preferences of Inputs to the Superior Colliculus from Areas of Dorsal and Ventral Streams of Mouse Visual Cortex". In: *Journal of Neuroscience* 33 (4 Jan. 2013), pp. 1696–1705. ISSN: 0270-6474. DOI: 10.1523/jneurosci.3067-12.2013. PMID: 23345242.

[38] Vermaercke Ben et al. "Neural discriminability in rat lateral extrastriate cortex and deep but not superficial primary visual cortex correlates with shape discriminability". In: *Frontiers in Neural Circuits* 9 (May 2015). DOI: 10.3389/fncir.2015.00024. PMID: 26041999.

[39] Vermaercke Ben et al. "Functional specialization in rat occipital and temporal visual cortex". In: *Journal of Neurophysiology* 112 (8 Oct. 2014), pp. 1963–1983. ISSN: 0022-3077. DOI: 10.1152/jn.00737.2013. PMID: 24990566.

[40] Emmanouil Froudarakis et al. "The Visual Cortex in Context". In: *Annual Review of Vision Science* 5.1 (2019), pp. 317–339.

[41] Eugenio Piasini et al. "Temporal stability of stimulus representation increases along rodent visual cortical hierarchies". In: *Nature Communications* 12.1 (July 21, 2021), p. 4448. DOI: 10.1038/s41467-021-24456-3. URL: https://www.nature.com/articles/s41467-021-24456-3.

[42] Kar Kohitij et al. "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior". In: *Nature Neuroscience* 22 (6 Apr. 2019), pp. 974–983. ISSN: 1097-6256. DOI: 10.1038/s41593-019-0392-5.

[43] Tang Hanlin et al. "Recurrent computations for visual pattern completion". In: *Proceedings of the National Academy of Sciences* 115 (35 Aug. 2018), pp. 8835–8840. ISSN: 0027-8424. DOI: 10.1073/pnas.1719397115. PMID: 30104363.

[44] Tang Hanlin et al. "Spatiotemporal Dynamics Underlying Object Completion in Human Ventral Visual Cortex". In: *Neuron* 83 (3 Aug. 2014), pp. 736–748. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2014.06.017. PMID: 25043420.

[45] Bashivan Pouya, Kar Kohitij, and Dicarlo James. "Neural Population Control via Deep Image Synthesis". In: *Science* 364 (Nov. 2018), p. 9436. DOI: 10.1101/461525.

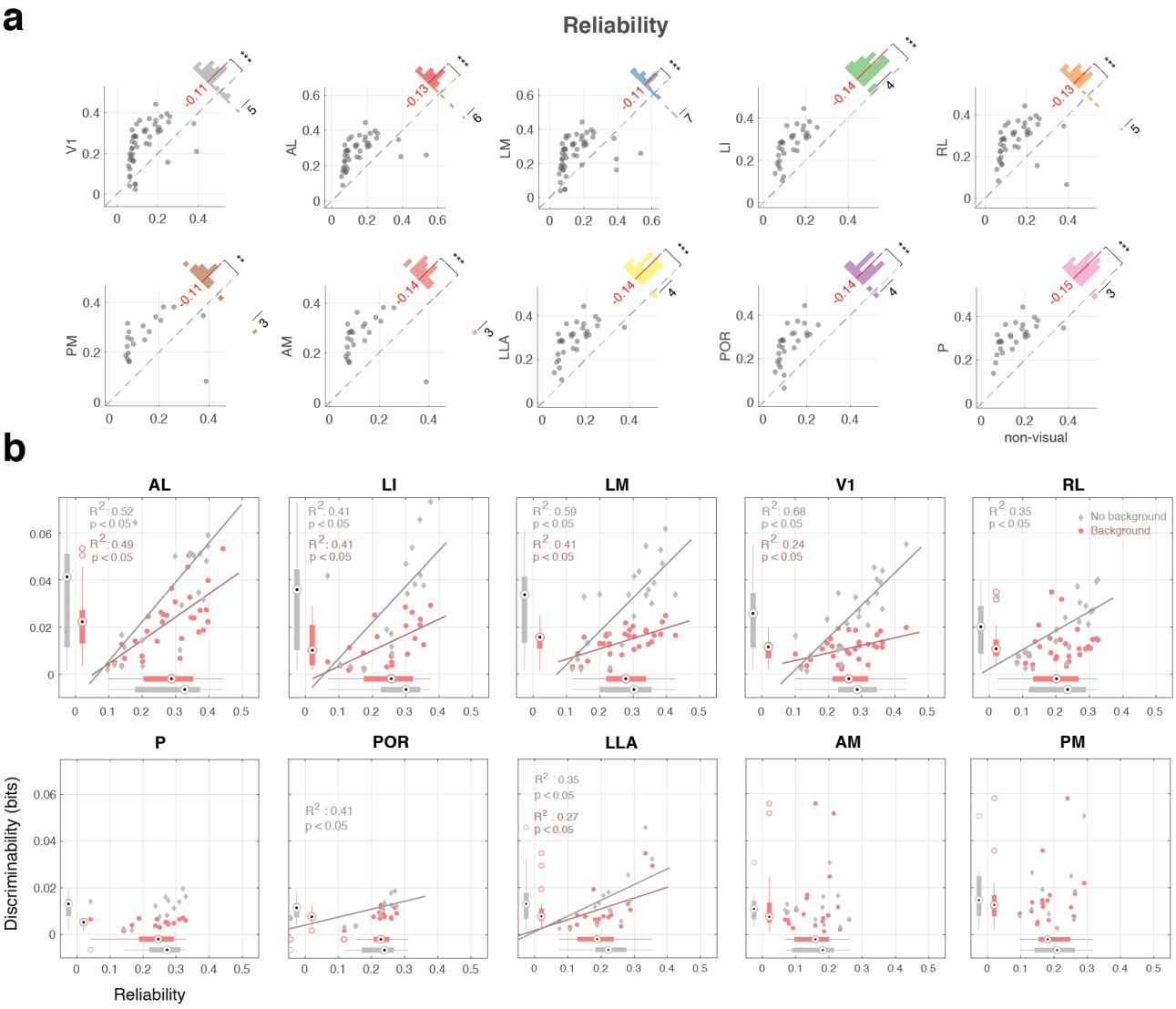
[46] Ponce Carlos et al. "Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences". In: *Cell* 177 (4 May 2019), 999–1009.e10. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.04.005.

[47] Walker E et al. "Inception loops discover what excites neurons most using deep predictive models". In: *Nat. Neurosci* 22 (2019), pp. 2060–2065. DOI: 10.1038/s41593-019-0517-x.

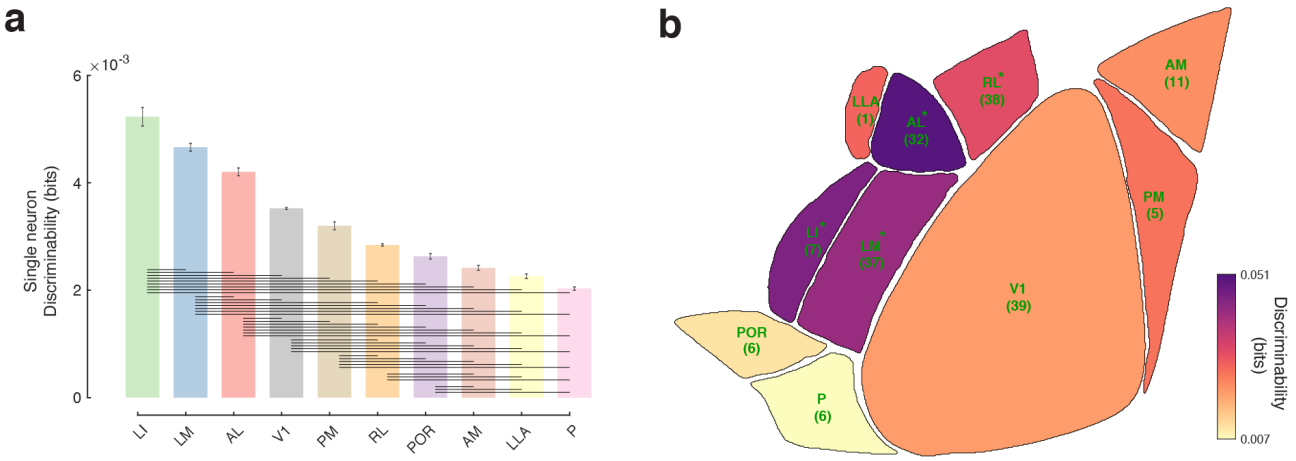
[48] Froudarakis Emmanouil et al. "Population code in mouse V1 facilitates readout of natural scenes through increased sparseness". In: *Nature Neuroscience* 17 (6 Apr. 2014), pp. 851–857. ISSN: 1097-6256. DOI: 10.1038/nn.3707. PMID: 24747577.

[49] Pnevmatikakis Eftychios et al. "Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data". In: *Neuron* 89 (2 Jan. 2016), pp. 285–299. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2015.11.037. PMID: 26774160.

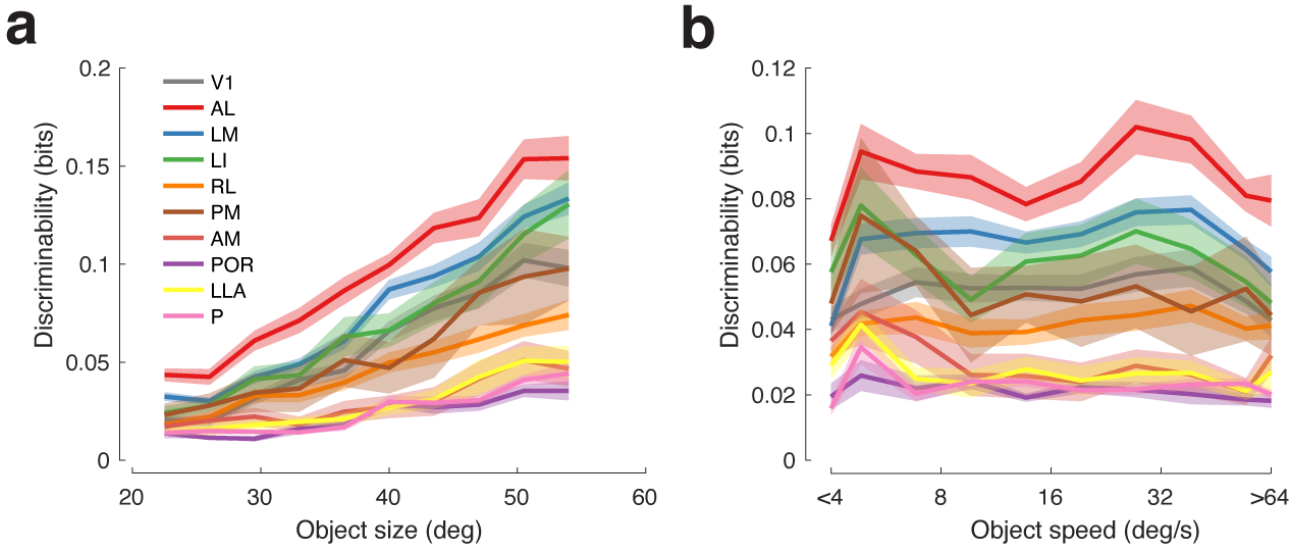
7 Supplementary Figures



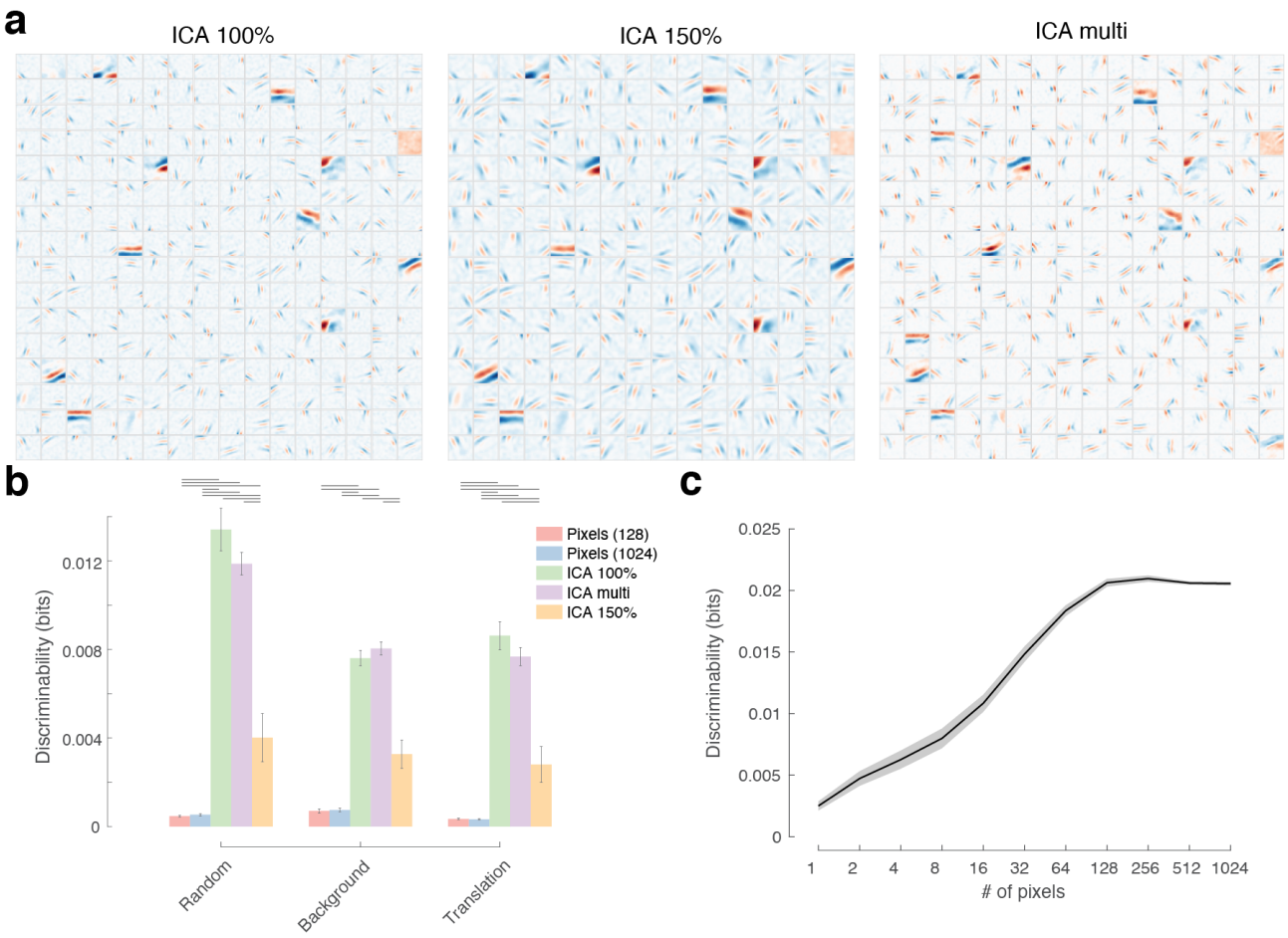
Supplementary Figure 1: Reliability across all visual areas. (a) Comparison of the average reliability of the responses to the object stimuli across neurons of all visual areas (y-axis) and neurons in non-visual areas (x-axis). Insert histogram represents the difference between the average reliability between each visual area and non-visual areas. Red line and number indicate the mean difference across all recording sites. Wilcoxon signed rank test *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. (b) Discriminability vs average reliability for all the cells with each recording. Plotted separately for objects w/ (red) and w/o background (gray). The regression line is indicated for each and the explained variance of the regression is noted on the top left of each plot. At the sides of each axis are the boxplots for each of the datasets.



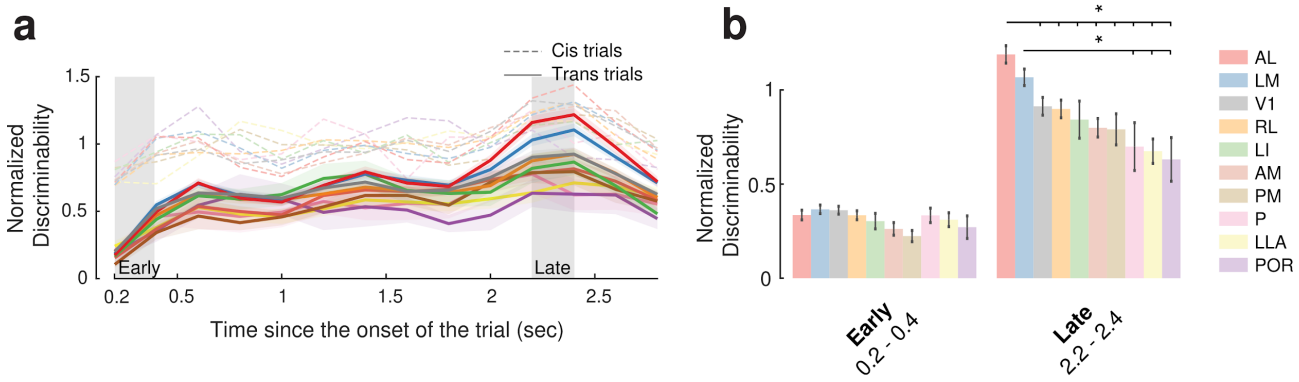
Supplementary Figure 2: Discriminability with single neurons or with RF restriction. (a) Bar graph of the average discriminability when using single neurons to decode the object identity. Horizontal lines indicate $p < 0.01$ Kruskal-Wallis with multiple comparisons test. N is reported in figure 2c. (b) Average discriminability for all visual areas when selecting a population of 20 cells that have their receptive fields centered within the same 20 degrees of visual space. The number below each area represents the recording sites sampled. * $p < 0.05$ Wilcoxon signed rank test when compared to V1.



Supplementary Figure 3: Object size and object speed effect on decoding. (a) Discriminability as a function of the object size for all visual areas. (b) Discriminability as a function of object speed for all visual areas. Shaded areas represent S.E.M.



Supplementary Figure 4: Models of RF and decoding performance. (a) The 256 receptive fields (ICA 100%), their enlarged version (ICA 150%) and their combination (ICA multi) that were used for the computational model. (b) Discriminability of the simulated responses of 128 units to objects and the generalization test on objects with background and translation, from the filters in (a) and also the pixels of the stimuli as input. Horizontal lines indicate $p < 0.05$, Kruskal-wallis multiple comparison test. (c) Discriminability of the simulated responses from the 1024 pixel model as a function of the number of pixels used from the decoder. Shaded areas represent S.E.M.



Supplementary Figure 5: Temporal dynamics when object identity is switching. (a) Discriminability across time for trials where the preserved object identity is preserved (cis-trials) or switched (trans-trials). Discriminability is normalized to the average discriminability of the cis trials. Shaded areas represent S.E.M. (b) Bar plot of the normalized discriminability of the trans trials during the 0.2-0.4 and 2.2-2.4 seconds of the trial. Kruskal-wallis multiple comparison test * $p < 0.05$.