Invariance and selectivity to spatial variables in deep neural networks and the brain

Margaret C Von Ebers[1,3,4] and Xue-Xin Wei[1,2,3,4,5]

1) Department of Neuroscience, The University of Texas at Austin
2) Department of Psychology, The University of Texas at Austin
3) Center for Perceptual Systems, The University of Texas at Austin
4) Center for Theoretical and Computational Neuroscience, The University of Texas at Austin
5) Center for Learning and Memory, The University of Texas at Austin

Abstract

The hippocampus contains cognitive maps of the environment that consist of multiple functional types of spatial cells, such as place cells, grid cells, and head direction cells. Although it is generally thought that the formation of cognitive maps involves path integration processes, recent work challenges this view. In particular, it was reported that model units in deep neural networks trained on vision tasks are selective to locations and other spatial variables, despite the fact that these networks receive only visual inputs. While these studies and most previous studies emphasize the selectivity of individual neurons' responses to a given spatial variable, invariance to other spatial variables has not been systematically studied. We propose a computational framework for understanding neurons' joint tuning properties, which capture both selectivity and invariance. Applying this framework to analyze and compare spatial representations in recently proposed vision-based deep neural networks and the rodent brain, we find substantial differences in these representations. In particular, in the rodent brain, place cells are selective to location and invariant to head direction, whereas head direction cells are selective to head direction and invariant to location. In contrast, in deep networks, model units that are selective to location are not invariant to head direction, leading to intermingled representations of location and head direction. We further apply our approach to study the joint tuning of head direction and movement direction in rodent head direction cells, and find that information about movement direction is almost absent in these neurons. Collectively, our results demonstrate the importance of studying joint tuning properties for understanding the structure of neural codes.

1 Introduction

One basic approach to studying the neural code is to examine the tuning properties of individual neurons [1, 2, 3, 4, 5, 6, 7]. Neuroscientists have characterized the tuning properties of neurons to various behavioral variables in many brain regions (e.g., [3, 4, 5, 6, 8]). In particular, in the hippocampus and related brain areas (such as the entorhinal cortex and parahippocampus), neurons selective to different spatial variables (e.g., location, head direction, speed) have been discovered [2, 9, 3, 6, 10, 8]. Prior research has focused mainly on examining selectivity for one stimulus variable at a time (e.g., visual orientation, location, or head direction). If a neuron is selective to multiple stimulus dimensions individually, it is considered to exhibit conjunctive tuning [11] or mixed selectivity [12, 13, 14].

A large body of work has investigated how the brain constructs spatial maps of the environment. It is generally believed that the brain integrates spatial inputs such as velocity input and head rotation to generate such maps [15] – typically known as the path integration hypothesis [16] – while other types of sensory input, such as visual landmarks and boundaries, may be primarily important for recalibrating the maps [17, 18]. Hand-crafted continuous attractor network models based on this idea have been developed to model place cells, head direction cells, and grid cells [19, 18, 20]. More recently, it was shown that optimizing recurrent neural networks to perform path integration (or angular path integration) based on spatial inputs can model grid cells [21, 22, 23, 24] and head direction cells [25]. However, recent studies have challenged this view. By studying deep neural network models trained to solve purely visual tasks, it was reported that many model units in the networks exhibited spatial selectivity similar to that observed in the rodent hippocampus [26, 27]. Unlike previous models based on path integration, these models received only visual inputs. Importantly, when comparing the tuning properties of individual neurons in the brain and corresponding units in the model, these studies only examined the marginal tuning for individual stimulus variables, one at a time. However, this approach generally assumes that the joint tuning can be factorized into the product of the marginal tuning curves, and may fail to provide an unbiased characterization of the joint tuning properties to multiple stimulus variables.

To address this problem, we propose to directly analyze the *joint tuning* properties of individual neurons to multiple stimulus variables simultaneously. We will show that studying the tuning curve for one variable at a time is insufficient to reveal the joint tuning properties. We develop an analysis framework to analyze the selectivity and invariance of neural responses when multiple stimulus variables are considered. Applications of this framework to study rodent spatial navigation circuits and recently proposed deep neural network models of these circuits reveal a number of insights. First, rodent place cells are selective to the animal's location while being invariant to head direction. Conversely, head direction cells are selective to head direction while being invariant to the animal's location. Second, model units in deep neural network models trained to perform vision tasks [27, 28] exhibit complex and entangled tuning properties for location and head direction, and thus are markedly distinct from place cells and head direction cells in rodents. Third, rodent head direction cells are largely invariant to movement direction, an observation that has important implications for models of spatial navigation circuits. While we focus on spatial navigation here, the proposed approach of analyzing joint tuning is general and can be applied to other

neural systems as well.

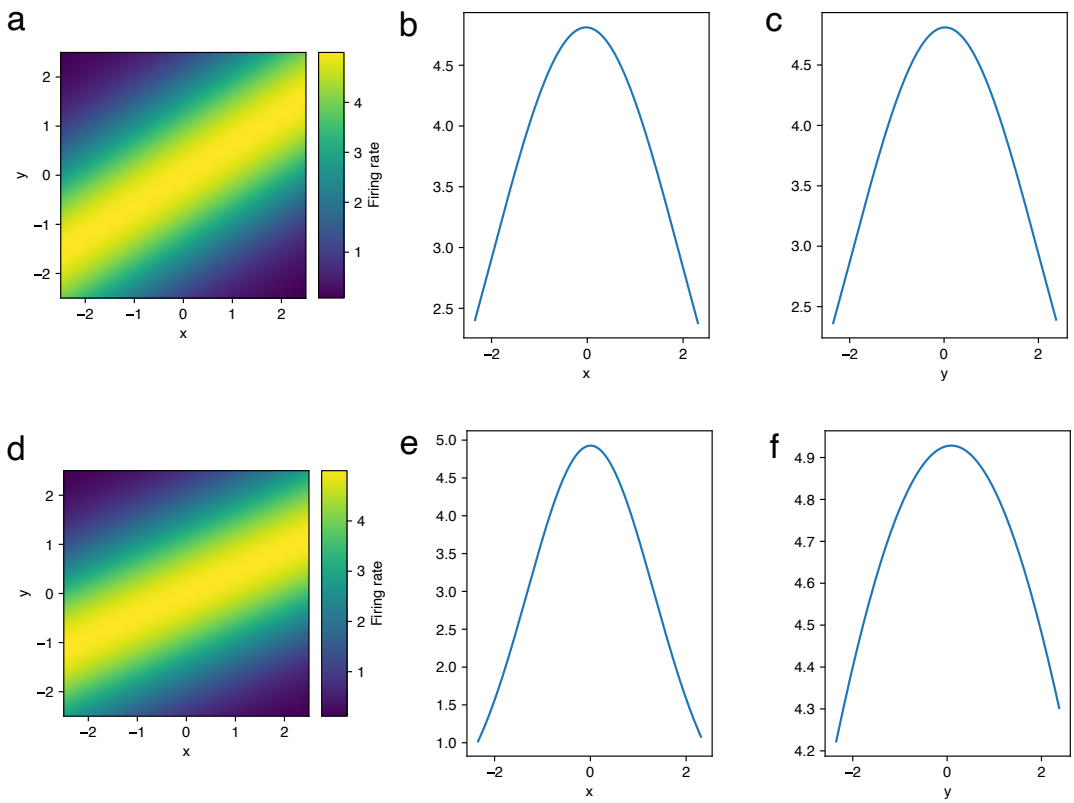

Figure 1: The joint tuning curve may not be captured by the generalized linear model (GLM). (a) The ground truth joint tuning curve for a hypothetical neuron. This neuron exhibits entangled joint tuning to variables x and y. The marginal tuning for x and y are both flat. (b,c) If a GLM framework is applied to analyze simulated responses sampled from this neuron, the extracted marginal tuning using GLM would both be non-flat, thus misleading. (d-f) Similar to (a-c), but for a different hypothetical model neuron.

2 Results

2.1 Marginal tuning curves are insufficient to capture joint tuning properties

When only one stimulus variable (denoted as x, such as the head direction of the animal) is considered, the tuning property of a neuron can be described by its tuning curve $f(x)$. The tuning curve quantifies the firing rate of a neuron for a given stimulus variable. If another stimulus variable y is also of interest (e.g., the movement direction of the animal), the tuning curve $g(y)$ can also be computed separately for y. Suppose that our goal is to understand whether this neuron is selective to both x and y; the standard approach would be to examine tuning curves $f(x)$ and $g(y)$ to see if one or both exhibit strong selectivity. Suppose that $f(x)$ is a narrow Gaussian function and $g(y)$ is completely flat; then it is tempting to claim that this neuron is selective to variable x and is not selective to (or is invariant to) y. However, this interpretation is generally not warranted.

We need to change this section because we ended up using a different example. Fig. 2 illustrates two

such counter-examples. The joint tuning curve $h(x, y)$ for the first example is plotted in Fig. 2. Examining the marginal tuning curves, we see that $f(x)$ is tuned while $f(y)$ is flat. Although the marginal tuning curve for y indicates a lack of tuning, from the joint tuning curve, the response of this neuron is clearly not invariant to y. What happens is that y accounts for the variability of the neural response through its interaction with the encoding of x. For the second example (see Fig. 2b), the marginal tuning curves $f(x)$ and $g(y)$ are both flat. However, it would clearly be unreasonable to claim that this neuron is invariant to x and y because its response clearly contains information about (x, y). That is, when knowing the firing rate of the neuron, one can substantially narrow down the possible values of x and y. However, the tuning of the two stimulus variables is highly entangled and cannot be revealed by marginal tuning curves based on individual stimulus variables.

When joint tuning can be factorized into the product of marginal tuning curves, i.e., $h(x, y) = f(x)g(y)$, marginal tuning curves are sufficient to interpret joint tuning. The popular GLM framework used in neural data analysis makes this assumption [29]. Because in most applications of GLM an exponential link function is assumed, linearly adding responses in the log of firing rates is equivalent to multiplying the tuning curves. However, in general, this assumption needs to be tested experimentally. For the examples discussed above, applying GLM to the simulated data would lead to misleading results regarding selectivity to the stimulus variables (see Fig. 1XX).

The interpretation of marginal tuning curves may be further complicated by correlated sampling of stimuli. Consider the case of studying neural tuning to two correlated stimulus variables. Correlated stimulus variables are fairly common. In the study of spatial navigation systems, one example is the animal's head direction and movement direction during navigation. Another example is time and distance traveled when the animal is running on a linear track. Suppose that a neuron is selective to one variable x while invariant to the other variable y (see Fig. 2 for an example). Due to the correlation in the sampling of stimuli, the marginal tuning for y still suggests substantial selectivity (see Methods for more theoretical results). In this case, if we only rely on marginal tuning curves to interpret the tuning properties, we would incorrectly conclude that the neuron is tuned to y, although its response is completely invariant to y.

2.2 A method for analyzing the invariance and selectivity of neurons

Our results above show that, when analyzing the selectivity and invariance of neural response properties, it is crucial to examine joint tuning properties beyond marginal tuning curves. Relying on marginal tuning curves alone may lead to misinterpretation of the tuning properties. How should we study the joint tuning properties of an individual neuron? Practically, one simple procedure to determine whether the neural response is invariant to a stimulus variable x is to vary x and examine the tuning for the remaining variables. If only two variables are considered, this amounts to examining the tuning curve of $g(y)$ conditioned on different values of x. If the tuning curve for y is invariant when changing x, one may conclude that the response of this neuron is invariant to x. Otherwise, the response of this neuron depends on variable x. In other words, variable x contributes to the response variability of this neuron.

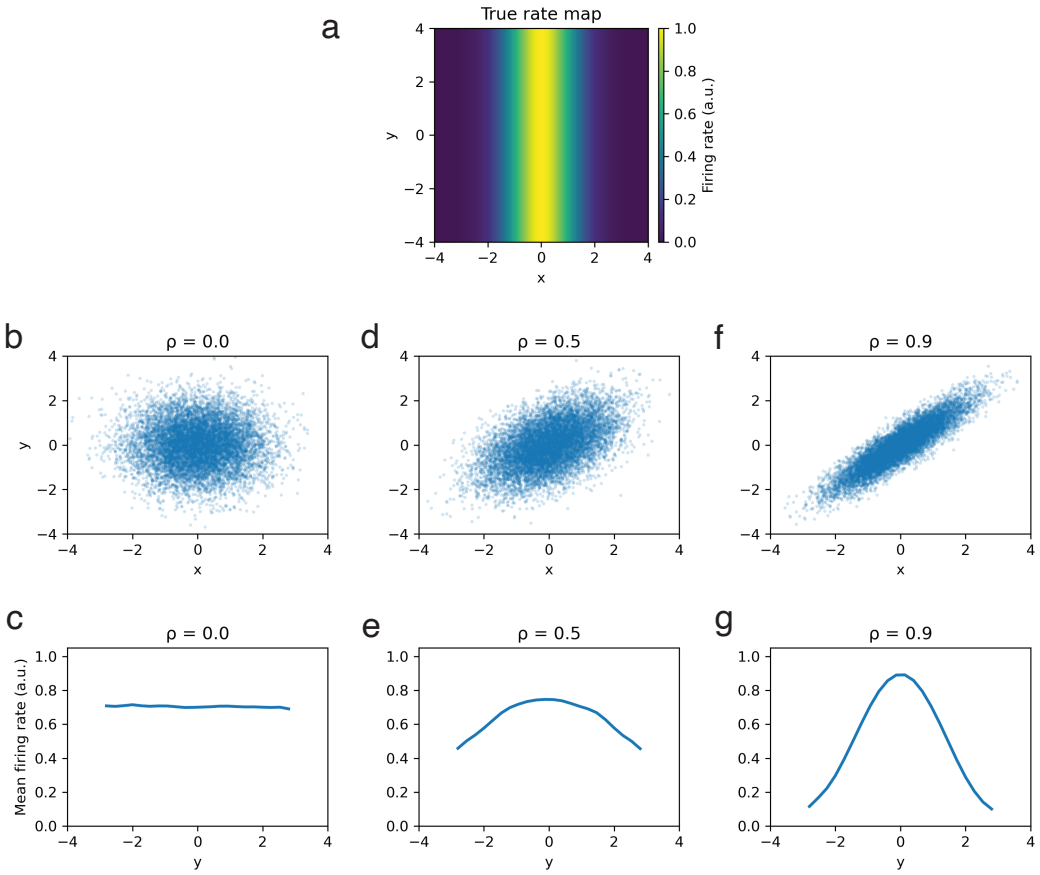

Figure 2: The joint tuning curve of a neuron cannot be adequately described by marginal tuning. (a) The joint tuning curves of a hypothetical neuron for variables x and y. This neuron is selective to x but is invariant to y. (b,c) When samples of x and y are uncorrelated, empirically computed marginal tuning for y accurately recovers the ground truth tuning. (d,e) When samples of x and y are correlated, empirically computed marginal tuning for y shows selectivity. (f,g) Similar to (d,e), but for a case where x and y are even more correlated.

To quantify joint tuning and understand the contribution of individual variables as well as their nonlinear interactions, we develop a simple analysis approach based on the intuition described above. To quantify the contribution of variable x, we calculate the variability of the joint tuning that is not explained by all other variables. In the case of two variables, this amounts to the residual variability that is not explained by y. Similarly, one can calculate the contribution of y via the variability that is not explained by x. Unlike analysis of variance (i.e., ANOVA) for linear additive models, summing up the contributions of y and x may not equal the total variance of the data.

Practically, one complication in estimating the contribution of individual variables is measurement noise. Consider the case mentioned above, in which the neural response is selective to x while completely invariant to variable y. Due to measurement noise (which could come from many sources), there will be residual variability in the estimated joint tuning maps that cannot be explained by x. If one simply uses the residual variance to estimate the contribution of variable y, the true contribution will be overestimated. To address this problem, we developed a procedure to estimate the variability induced by noise by quantifying the variability across joint tuning maps estimated from two equal splits of the data. Our final

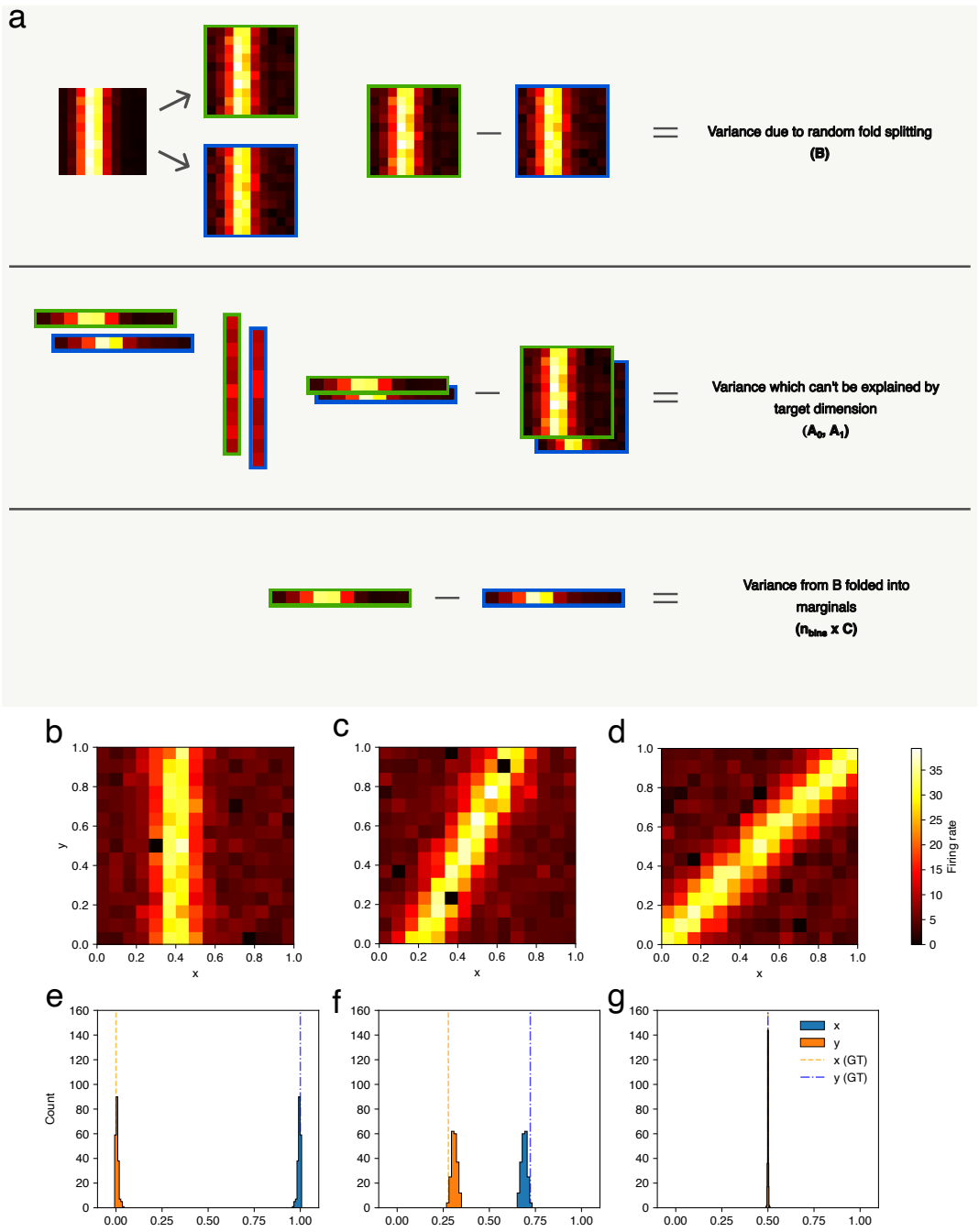

Figure 3: The explained variance of a neuron can be accurately quantified with the new method. (a) A schematic of the estimation procedure. Our method first splits the responses into two halves. Two joint tuning maps are estimated. The response variability of one variable conditioned on the other is calculated, and then the variability across the two tuning maps is subtracted to estimate the variance explained by the target variable. For mathematical details, see Methods. (b,c) Simulation results showing that the method can accurately recover the fraction of variance explained by the two variables. In this case, the ground truth is that one variable fully explains the variance and the other variable explains zero variance. (d-g) Similar to (b-c), but for two other scenarios.

estimate for the contribution of individual stimulus variables removes such noise-induced variability.

A schematic of our proposed analysis procedure is shown in Fig. 3a. Our mathematical analysis shows that the proposed estimator (details in the Methods Section) is unbiased under the assumption that mea-

surement noise is independent across the two splits. We performed systematic simulations to validate and benchmark our estimator. We find that the estimation procedure can accurately recover the relative contribution of each variable (for an example, see Fig. 3d,e).

2.3 Distinct selectivity and invariance of place cells and head direction cells in the rodent brain

We investigated the tuning properties of neurons recorded in the rodent spatial navigation system. We analyzed two datasets. The first is a hippocampal place cell dataset recorded while rats navigated in 2-D open environments [30]. The second contains head direction cells from multiple brain regions [31]. In addition to their selectivity, we are interested in the invariance of these distinct functional cell types. For example, we examined whether place cells are invariant to head direction. We used our analysis framework developed for understanding joint tuning together with population decoding techniques.

Place cells in the rat hippocampus are invariant to head direction. For the place cell dataset collected in [30], we first calculated the 2-D spatial firing rate maps for individual neurons. Consistent with previous reports on place fields, many neurons were selective to spatial location in the 2-D open environment with clear place fields visible upon inspection. Computing the marginal tuning curve for head direction, we find that most place cells do not exhibit strong head direction tuning, consistent with previous reports. However, it remains unclear whether these neurons exhibit strong invariance to head direction. Recall that our theoretical results above suggest that even when the tuning curve for a stimulus is flat, neural responses may still depend on the stimulus (Fig. 2). Might need to be changed here as well since fig 1 has changed. To further investigate this question, we computed the joint 3-D tuning map for 2-D location and HD (see Fig. 5a for one example). Indeed, visual inspection suggests that neural responses are invariant to HD. However, due to limited data, some combinations of HD and location have few or no data points.

To address this question, we performed two additional visualizations. First, we plotted the joint tuning for HD and the x-axis of the 2-D environment. This analysis projects the 3D data onto the 2-D subspace defined by HD and the x-axis of location. If the neuronal response is invariant to HD, the heatmap should be dominated by vertical bars/stripes. This is indeed true for most neurons that were classified as place cells according to the standard procedures used to analyze place cells. Figs. 4a,b show a couple of example place cells from this dataset. The vertical bar pattern is apparent when we examine the joint tuning of HD and the x-axis in the 2-D environment. Similar results hold when examining the joint tuning for HD and the y-axis of the 2-D environment.

We quantified the contribution of HD and 2-D location using the procedure we developed (see Fig. 3). In particular, we are interested in understanding whether it is indeed true that place cells exhibit invariant responses to the animal's head direction. Fig. XX shows that neural response variability is mostly explained by the location of the animal. Although head direction (HD) also accounts for a small fraction of the variance, its contribution is much smaller. In general, we found that the ratio of the contribution of

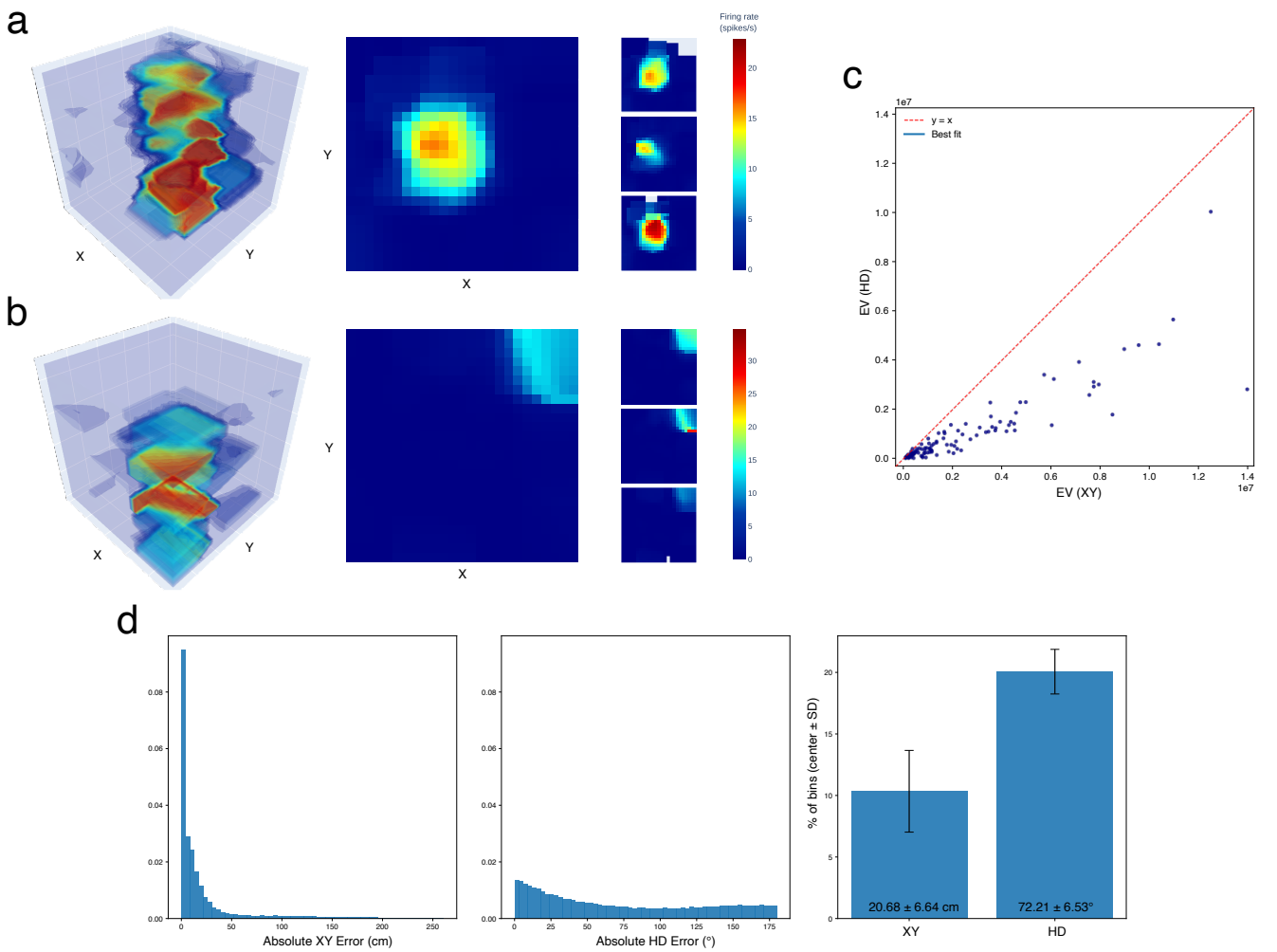

Figure 4: Place cells in the rat hippocampus are selective to location and largely invariant to head direction. (a) An example place cell. The three panels show its joint tuning for x and head direction, y and head direction, and 2-D location tuning, respectively. For x and head direction joint tuning, a vertical band pattern is apparent, indicating invariance to head direction. (b) Similar to (a), but for a different place cell. (c) Quantification of the variance explained by location and head direction. (d) Decoding results for location suggest that location can be accurately decoded. (e) Decoding results for head direction suggest that head direction cannot be accurately decoded from the place cell population. (f) The number of neurons used in the decoding analysis.

location to HD is XXX.

To further corroborate these results, we performed additional decoding analysis based on simultaneously recorded place cell populations. We found that the decoding error for location in the 2-D environment is relatively small (MAE = 20.68 cm, about 10% of the environment). In contrast, the decoding error for head direction is high. Fig. 4d,e shows the decoding results of four experimental sessions. These results are consistent with the low HD contribution based on partitioning the contribution of neural response variability in Fig. 4c.

Rodent head direction cells are invariant to the animal's location. Next, we analyzed head direction cell data recorded from rats [32]. This dataset contains simultaneously recorded neurons in the postsubiculum. We first examined the tuning properties of individual neurons by plotting the 3-D joint

tuning for head direction and location in the 2-D environment. Visual inspection suggests that these head direction cells are largely invariant to the animal's location.

Using our method to partition the variability, we find that head direction explains most of the variability, with location accounting for only a minor portion (Fig. 5). These results suggest that head direction cells are largely invariant to location. Decoding head direction and location from simultaneously recorded populations (n=30 sessions) reveals a large decoding error for location (MAE = 40.43 cm; size of the environment is 80 cm) and a small decoding error for HD (MAE = $16.95°$).

Together, these results suggest that there is a double dissociation of the response properties of place cells and head direction cells in rodents. The response of a place cell strongly depends on the animal's location and is invariant to the animal's head direction. In contrast, head direction cells are strongly tuned to head direction and are only minimally modulated by the animal's location. These findings have implications for the computations that lead to these response properties. They suggest that it is unlikely that location selectivity in place cells originates from certain visual features in the environment, in which case place cells would be expected to depend on head direction as well.

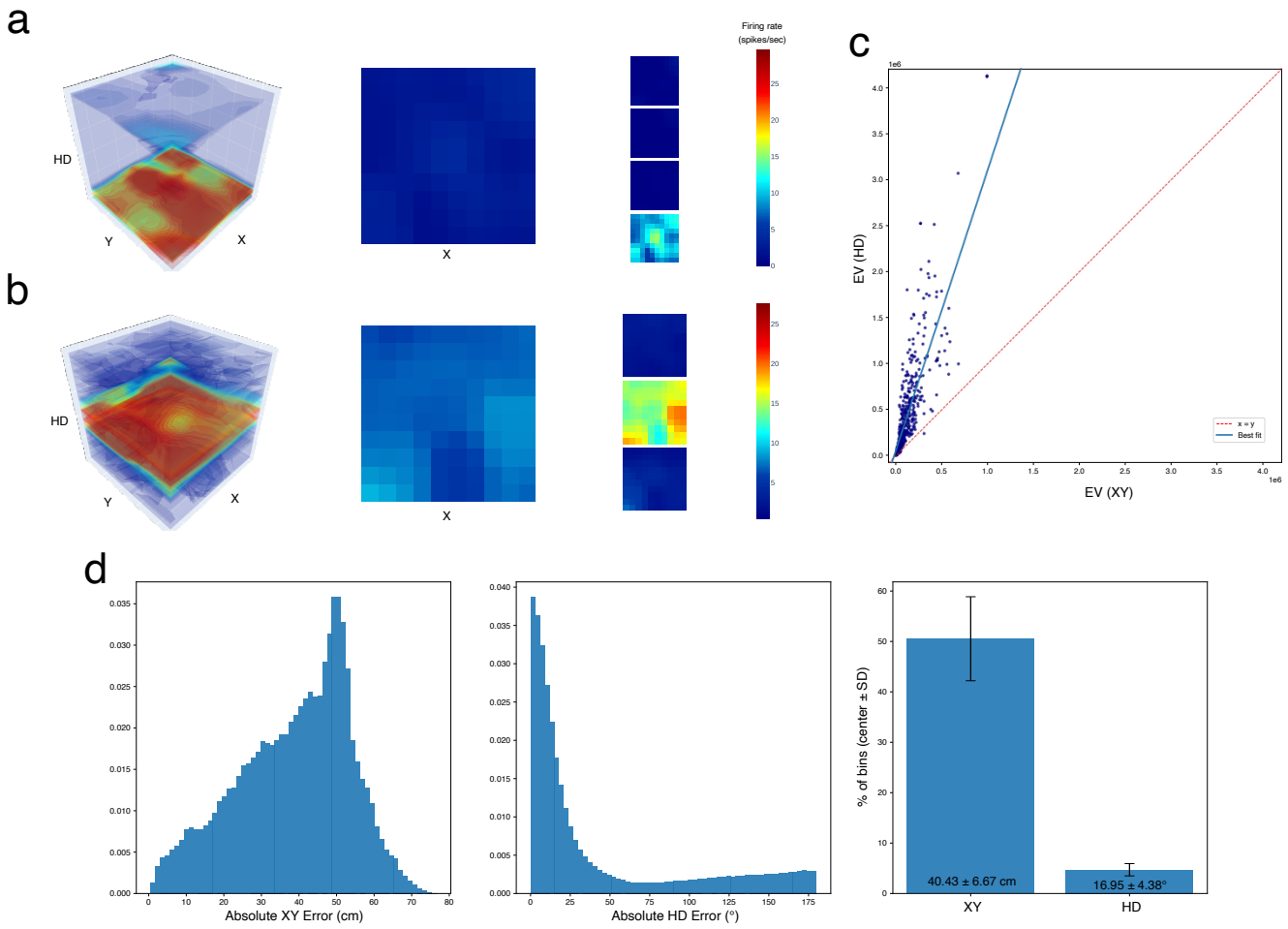

Figure 5: Head direction cells in the rat are selective to head direction and largely invariant to location. (a,c,e) Example HD tuning curves of three example neurons. (b,d,f) The corresponding joint location and HD tuning maps for these three neurons. (g) Quantification of the response variance explained by HD and location.

2.4 Model units in deep network models based on vision exhibit entangled patterns of spatial tuning

While many previous studies suggest that path integration is important for the formation of cognitive maps in the brain [33, 20, 21], recent studies reported place cell-like responses in deep neural networks trained or pretrained to perform vision tasks [27, 26]. In [26], it was reported that a virtual navigation agent trained to predict the next image frame leads to the formation of cognitive maps. In particular, neurons in intermediate layers exhibit spatial selectivity similar to place cells. Meanwhile, Luo et al. analyzed the responses of individual units in ResNet (pretrained on image classification tasks, or untrained) in a spatial environment [27], and reported that a large fraction of model units throughout the network layers exhibit spatial selectivity (to location or head direction). Importantly, these studies [27, 26] only examined marginal tuning for these spatial variables such as location or head direction without examining joint tuning properties. Thus, it remains unclear whether the joint tuning for place and HD would be similar to that of hippocampal place cells. We sought to address this question by examining the joint tuning of units in the types of deep network models studied in [27, 26].

Model units in networks trained on predictive coding exhibit entangled spatial tuning. Inspired by work on predictive coding [34, 35], Gornet et al. [26] trained a virtual navigation agent to predict the next image frame in image sequences (Fig. 6a). They then analyzed the response properties of model neurons in the intermediate layers and found that these neurons show place-cell-like responses.

We re-implemented the training settings used in [26] using a Minecraft environment similar to what they developed. After training the simulated agent to solve the next-frame image prediction task, we analyzed the response properties of model units in different layers of the model (see SI Fig. XX for the performance of the agent in this prediction task). Because the head direction of the simulated agent was allowed to vary, we could obtain model responses corresponding to different head direction angles when the agent was at the same location in the environment. Examining the response properties of the models using established metrics in the place field literature [36], we found that a fraction of them can be classified as place cells (see Fig. 6b). These results suggest that there are putative "place cells" in the network, which is consistent with the original reports in [26]. We next examined the joint tuning of place and head direction. Strikingly, model units generally exhibit complex and highly entangled tuning for location and head direction (see Fig. 6c for example model units). When conditioned on different ranges of head direction, the location tuning function clearly changes.

Quantification of variance explained by location and head direction in individual units suggests that the responses of most model units were roughly equally explained by head direction and location. We selectively analyzed the model units identified as putative place cells and found similar results. This entangled tuning property is clearly different from place cells in rodents (and is also different from head direction cells). This result is highly consistent across different layers of the network.

CNNs also exhibit an entangled pattern of spatial tuning. We next investigated another recent model based on networks trained to perform image classification tasks. Over the past decade, convolutional

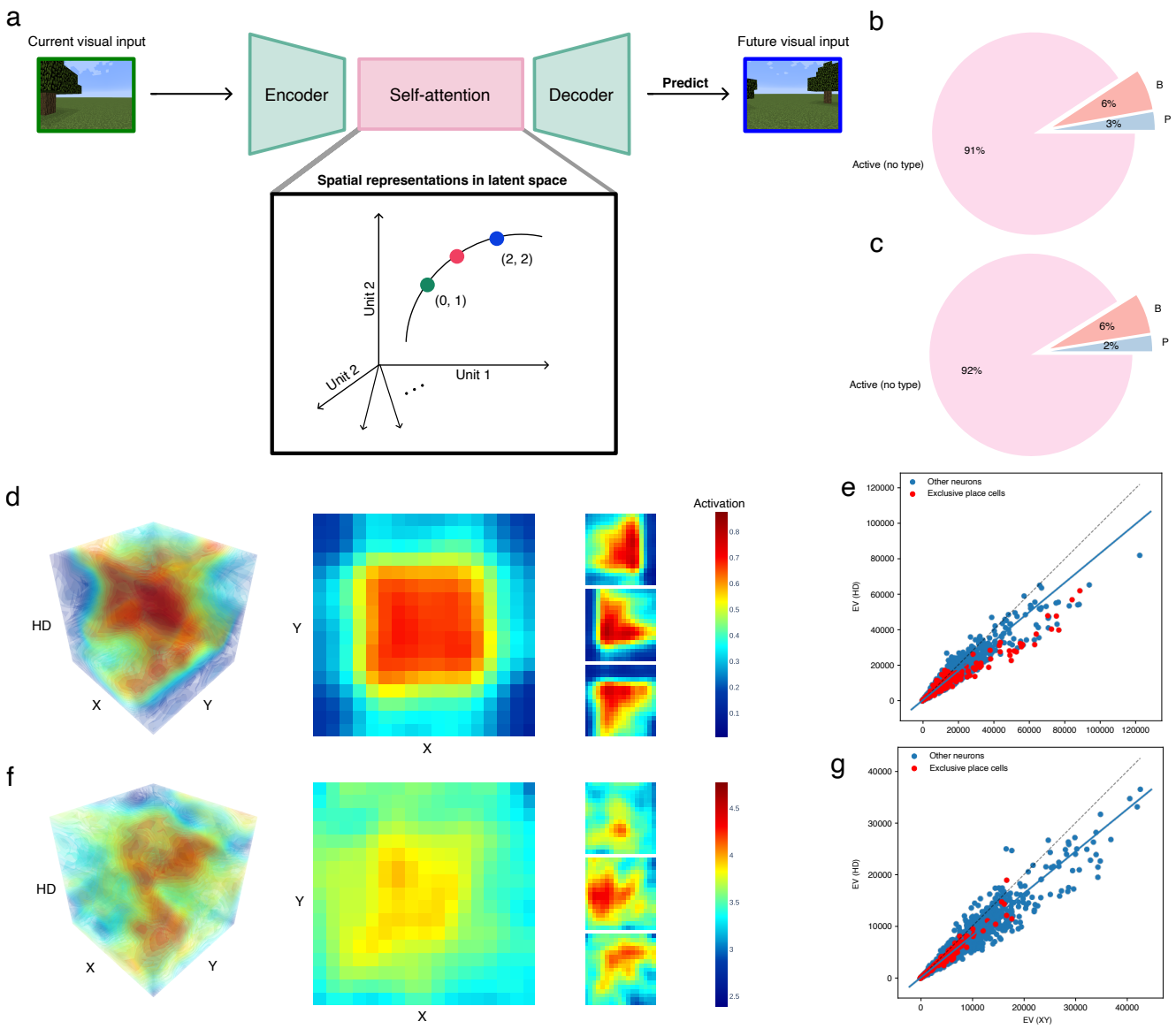

Figure 6: Model neurons in deep networks trained to predict the next image frame also exhibit highly entangled tuning properties for location and head direction. (a) A schematic of the modeling framework used in [26]. Adapted from [37]. (b) Pie charts showing the fraction of model neurons classified as place cells and other cell types in three different layers of the model. (c) One example model neuron that was classified as a place cell. While the 2-D spatial map exhibits a place field, the 3-D joint tuning (location and HD) clearly indicates that the response is not invariant to HD. (d) Quantification of the variance explained by location and HD by individual neurons in the network model.

neural network models (CNNs) have become a popular modeling framework for understanding how the brain processes visual information [38, 39, 40]. Luo et al. [27] used this framework to study the navigation system in the brain. Specifically, they modeled the visual inputs of a virtual agent navigating in a 2-D open area similar to the typical environment for experiments done in rodents. Importantly, the CNNs are pretrained on visual tasks and receive only visual inputs with no input about the agent's velocity. Naively, one would expect that model units based on pure visual inputs should not lead to spatial selectivity. Surprisingly, they reported that many neurons in the network throughout the layers exhibit place cell-like responses, as well as head direction tuning.

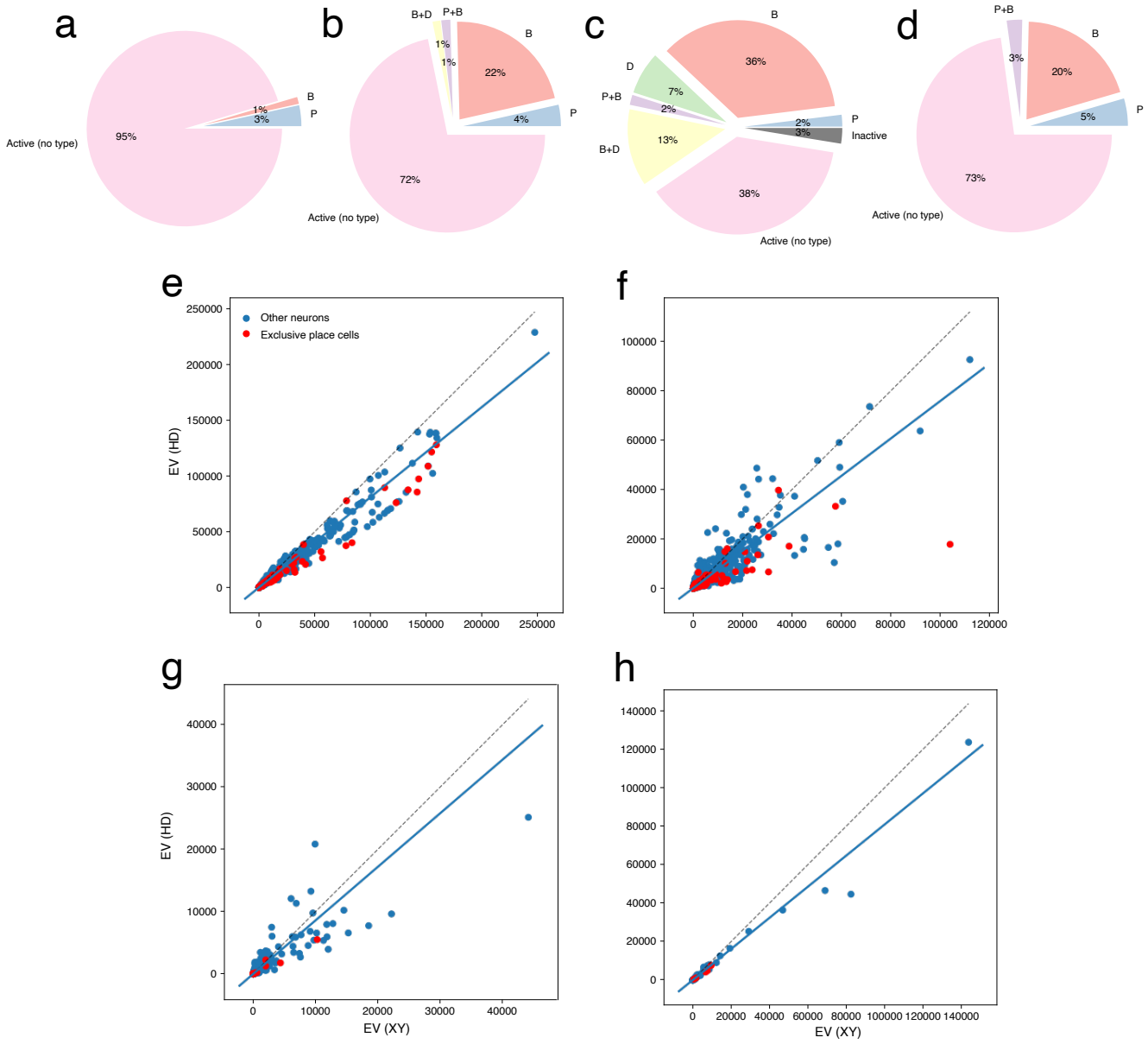

Figure 7: Model neurons in CNNs trained on image recognition tasks exhibit highly entangled tuning properties for location and head direction. (a-d) Quantification of the variance explained by location and HD by individual model neurons in the VGG-19 network. Each panel shows the result from one model layer. The results clearly show that the two variables contribute equally to response variability, unlike place cells and head direction cells in the brain.

We replicated their results using a similar setting based on the Minecraft environment, except that we did not find a large fraction of model units selective to head direction. We then took a CNN (VGG-19) pretrained to perform image classification tasks and collected the responses of this CNN in the virtual environment. We first analyzed the marginal tuning for location and HD in the CNN. Indeed, consistent with the results in [27], we found that some neurons exhibit place fields in 2D space. [27] also reported that many model units exhibit strong head direction tuning and a subset of neurons exhibit conjunctive place and HD tuning. Our own analysis suggests that few model units are tuned to HD and almost none of them are conjunctive place and HD cells. At face value, this seems to suggest that the model units

exhibit properties similar to place cells in rodents. Again, as noted earlier, according to our results above it is necessary to examine the joint tuning of place and HD for a better comparison.

To investigate this question, we visualized the 3D place and HD maps for model units that had been classified as tentative place cells. The results suggest that the firing fields are highly scattered in 3D space. Few neurons exhibit invariance to HD. Typically, the model unit only generated high activity with particular combinations of place and HD. This observation is similar to what we found in the network trained on the prediction task. We next quantified the variance of individual neurons explained by location and head direction in different layers of the network. We found that across all layers, location and head direction explain approximately equal amounts of variance for individual model units (see Fig. 7 for results from four example layers from a CNN).

Together, our results replicated the key findings reported in recent studies that investigated spatial tuning properties in deep neural networks trained on purely visual tasks, without vestibular inputs. Indeed, a fraction of model neurons exhibit selectivity for key variables during navigation, such as location and head direction, based on previous analysis methods that analyze marginal tuning to these variables. However, examining joint tuning properties for head direction and location reveals a large discrepancy between the spatial tuning properties of model neurons and spatial cells in the brain. Thus, these recently proposed models based on visual inputs are insufficient to explain the key properties of spatial cells discovered in the brain. In addition, our results highlight the importance of examining joint tuning to multiple variables together.

2.5 Head direction cells are only weakly tuned to movement direction

Our framework is not limited to the comparison of deep network models and the brain. It can also be used to study the structure of conjunctive tuning properties that may inform computational theories of brain function. To demonstrate this, we use our framework to address a question that is essential to models of grid cells. Grid cells exhibit hexagonal firing fields that tile space [3, 6]. A popular class of models assumes that grid cells perform path integration [20, 33, 41, 21]. Crucially, this class of models requires grid cells to receive inputs about the animal's movement direction and speed. So far, neurons encoding the animal's movement speed have been identified in EC and other areas [10]. However, neurons that selectively encode movement direction have not been explicitly identified. In previous models, it was often assumed that head direction and movement direction are equivalent. As head direction and movement direction are highly correlated, it is tempting to consider that an accurate movement direction signal is encoded in head direction cells. One previous study [42] examined these issues. They found that head direction was more strongly encoded than movement direction. They also found that head direction and movement direction are sufficiently distinct that using measured head direction as the path integration input was insufficient to generate grid cells in the path integration model of grid cells. However, it remains unclear whether head direction (HD) cells are tuned to movement direction (MD) at all, and whether the head direction population encodes a sufficiently strong movement direction signal to allow accurate path integration.

To address this question, we re-analyzed the head direction dataset from [32]. We first examined the marginal tuning curves for head direction and movement direction, respectively. A couple of example neurons are shown in Fig. 8a-d. Indeed, a substantial fraction of neurons in the dataset are also tuned to MD. Despite this observation, it remains unclear whether these neurons are truly selective to MD or whether they exhibit selectivity simply because MD and HD are correlated. As shown earlier, correlated stimulus sampling may complicate the interpretation of marginal tuning curves (see Fig. 2f,g). To address this confound, we examined the 2-D joint tuning of HD and MD. Fig. 8f shows three example neurons. We found that MD tuning when conditioned on HD is rather weak. We further used our variance partitioning method to quantify the neural response variability explained by MD and HD, respectively, for individual neurons. We found that, across the entire population, MD explains only a minor portion of response variability.

We further performed decoding analysis based on the population response to assess how well we can decode MD. We found that, while MD can be decoded at a level higher than chance, the overall error is quite large (see Fig. 8g for results from two example sessions). Overall, the MAE for decoded MD is XXX. These results suggest that the MD signal in head direction cells is weak and unlikely to be sufficient for accurate path integration. These results are consistent with the evidence reported in [42] and suggest that the sources encoding movement direction that would be needed for attractor network models of path integration remain to be identified. If a signal for MD exists in the spatial navigation system, it is unlikely to be encoded in head direction cells.

3 Discussion

We propose that it is important to study the selectivity and invariance of the neural code simultaneously. Crucially, invariance to a certain dimension generally cannot be assessed by examining the marginal tuning curve for that dimension. A completely flat marginal tuning curve may not indicate invariance to a certain dimension. There are cases where flatness is neither sufficient nor necessary for invariance. We developed metrics to quantify the contribution of individual stimulus variables to the response variability of individual neurons. If the contribution of a variable is zero, it means that the neuron is invariant to this variable. The analysis of joint tuning of a neuron to multiple stimulus variables provides a more comprehensive characterization of neural response properties that may be missed in the analysis of tuning curves for individual stimulus variables. Joint tuning is more informative for comparing the properties of different models and brains.

The joint tuning properties of individual neurons are often studied using GLM models [29, 14] to fit a regression model to partition the contribution of individual stimulus variables. The advantage of GLM is that it allows partitioning of response variance among many different variables, and the model is relatively easy to fit given limited data. However, GLM assumes that joint tuning can be factorized into the product of tuning along individual dimensions—an important assumption that may be violated in neural response analysis. For cases with a large number of behavioral variables, GLM seems to be a viable approach as

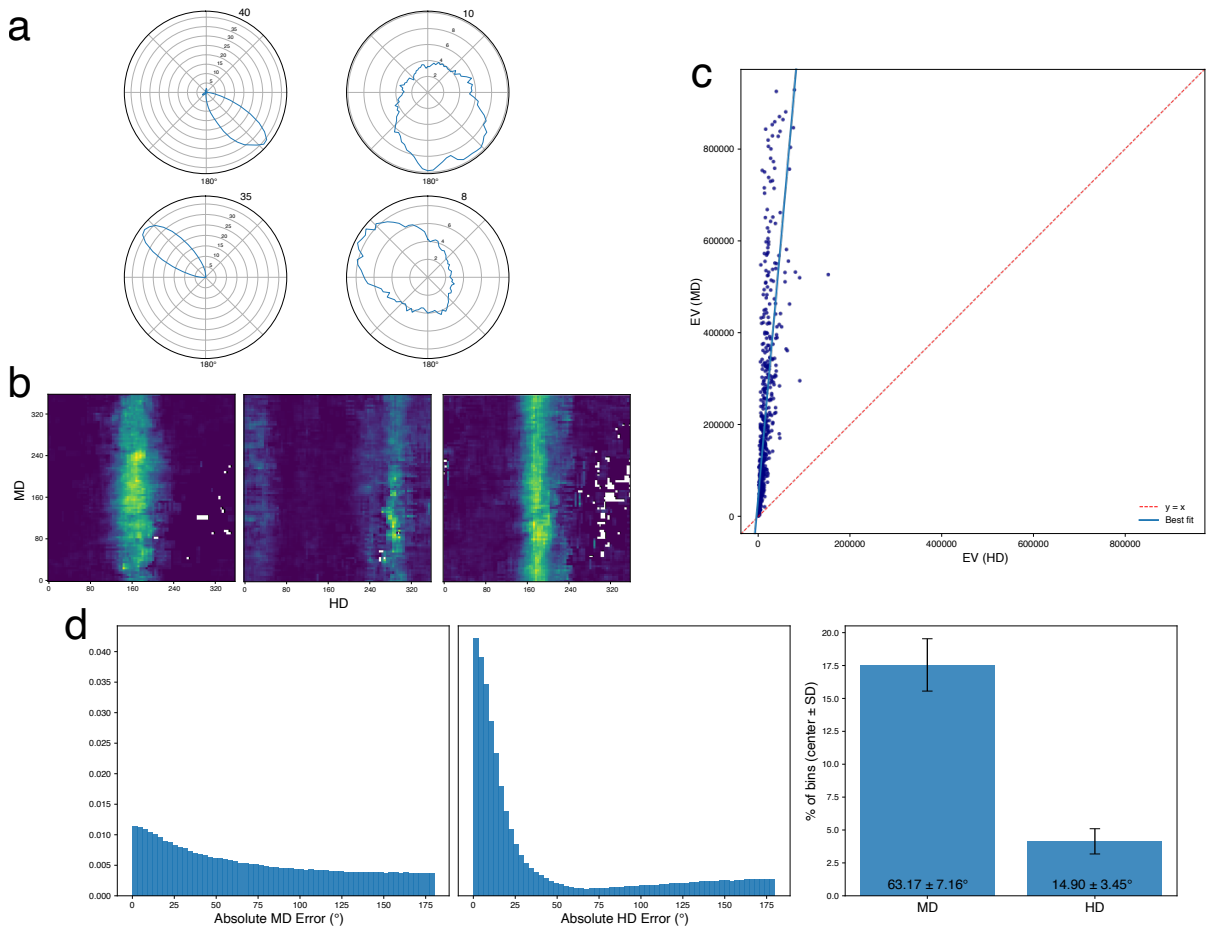

Figure 8: The MD signal encoded in the head direction population is extremely weak. (a,b) The HD and MD tuning curves of one example neuron. (c,d) Similar to (a,b), but for a different neuron. (e) Quantification of the variance explained by HD and MD clearly shows that HD contributes to most of the response variance. (f) Example joint tuning curves of HD and MD from three neurons. (g) Decoding results for MD based on two example sessions.

more complex models may not fit the data due to limited sample sizes. However, when only a handful of variables are considered, our proposed approach provides a powerful alternative that requires minimal assumptions and thus may potentially reveal unique insights about the structure of the code that are not possible with GLM.

Applying this framework to study rodent place cells, we find that place cells are not only selective to location but, at the same time, are largely invariant to the animal's head direction. In addition, we find that rodent head direction cells are strongly selective to head direction while being invariant to location. Thus, in rodents, two key spatial variables (location and head direction) are represented in distinct neural populations, which supports the notion of distinct functional cell types. The invariance of place cells to head direction (and vice versa) suggests that the code for space is "abstract" and is unlikely to be inherited from selectivity to sensory features.

When studying neural network models developed recently based on vision tasks [27, 26], we found that the tuning properties of individual neurons differ drastically from those of real place cells or head direction cells in rodents. Specifically, while a fraction of model units exhibit spatial selectivity and thus

can be classified as putative "place cells," these units in the deep network models do not exhibit invariant responses to HD. Instead, these model units encode strongly entangled representations of head direction and location. In addition, most model units do not exhibit clear marginal tuning for head direction, yet their response is strongly modulated by head direction via the interaction with location. The results are qualitatively similar in the two types of deep network models we examined [27, 28]. One interpretation of these results is that the highly entangled HD and location representations in deep networks arise because these neurons are selective to visual features. Because both HD and location are correlated with visual features, some kind of selectivity to these variables emerges in the network. For example, even a neuron in an early layer of the network that essentially behaves as an edge detector can exhibit location and HD tuning. By revealing a fundamental divergence in the response properties of model units and real neurons, these results help resolve a recent debate regarding whether deep neural network models trained on vision tasks are sufficient to explain tuning properties in the spatial navigation system, particularly in rodents.

Our approach can be broadly applied to study the joint tuning of individual neurons to multiple behavioral variables (ideally two or three), without making additional assumptions regarding how tuning of different variables interacts (such as multiplicative interactions assumed in GLM). We demonstrate this point by studying joint MD and HD tuning in head direction cells to show that head direction cells are largely invariant to movement direction. Thus, it is unlikely that these neurons encode the movement direction input that is important for path integration. Similar techniques may be used to understand the selectivity of place cells on linear tracks with respect to location, time, or distance traveled. It has been proposed that place cells on linear tracks may encode location, time, or distance traveled, or a combination of these [refs]. Although some progress has been made, it remains challenging to disentangle the contributions of these different factors because they are often highly correlated in experiments. Our approach provides a unique opportunity to estimate the contributions of these different factors for future research.

Our results have implications for mixed selectivity [ref], which has been reported in many different brain areas across different tasks. First, from a methodological point of view, our proposed techniques provide new ways to quantify mixed selectivity through examination of joint tuning and quantification of the contribution of individual stimulus variables. Second, our results based on analyzing the spatial response properties of deep network models suggest that, in some cases, mixed selectivity may arise due to difficulty in identifying the correct feature set. For example, in the deep network models we examined, many neurons in the first layer are oriented filters, thus encoding edge orientation. Analysis of orientation selectivity of these units would lead to clear interpretation of their response properties. However, when analyzing the spatial tuning properties of these units, they exhibit complex nonlinear mixed selectivity for these spatial variables. Nonlinear mixed selectivity arises in this case due to the highly nonlinear mapping between visual features and spatial variables (location and head direction). Only for particular combinations of location and head direction of the agent would an edge of a certain orientation appear in the receptive field of the model unit. Thus, extra care is needed when interpreting response patterns that appear to be mixed selective.

In this paper, we have focused entirely on the properties of individual neurons. One interesting question

is the implication of selectivity and invariance for population-level properties, such as neural manifold geometry. Previous theoretical work suggests that the response properties of individual neurons are informative about population-level properties as they define the projection of the neural manifold onto individual neuronal axes [7]. However, prior work is often done in the context of encoding single stimulus variables. Understanding the relationship between the properties of joint tuning to multivariate stimuli and neural manifold structure represents an important direction for future research.

4 Methods

4.1 Marginal tuning curve under correlated stimulus sampling

Assuming uniform sampling over x, the marginal tuning curve for x is estimated to be

$$\hat{f}(x) = \int p(y|x)h(x,y)dy = \int p(y|x)f(x)g(y)dy = f(x)\int p(y|x)g(y)dy$$

If random variables x and y are independent, the procedure recovers $f(x)\int p(y)g(y)dy$, which is proportional to $f(x)$. Otherwise, due to the interaction in the $p(y|x)$ term, the estimated marginal tuning curve for x will be affected by $g(y)$.

Consider a special case where the neuron is invariant to x, i.e., $h(x,y) = h(y)$ is a narrow Gaussian function. We further assume that only $y = x$ is sampled:

$$\hat{f}(x) = \int p(y|x)h(y)dy = \int \delta(y-x)h(y)dy = h(x).$$

In this case, if we rely on computing the marginal tuning curves to infer tuning properties, we would incorrectly conclude that the neuron is strongly tuned to x.

4.2 Estimating the contribution of individual stimulus variables to neural response variability

We first split the data into two sets: a training set and a test set. Suppose that we compute the firing rate maps for variables x and y using the training and test sets, denoted as $f(x,y)$ and $f_t(x,y)$, respectively. Further suppose that we discretize x and y into m and n bins, respectively. We denote $f_{i,j} = f(x_i, y_j)$ and $f^t_{i,j} = f_t(x_i, y_j)$. Further denote the marginalized (averaged) firing rate for y when conditioning on x_i as $\overline{f}_{x_i}$, and the marginalized (averaged) firing rate for x when conditioning on y_j as $\overline{f}_{y_j}$.

We can quantify the contribution of x by considering the residual variance that is not explained by y alone. This portion of variance contains the contribution of both y alone and the contribution of y and x through interactions. This quantity is useful because there are cases where the (marginal) tuning of a variable is flat, but nonetheless it is important for explaining the variability of the response due to its joint tuning with other variables. In the absence of noise, the contribution of y can be estimated by

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(f_{i,j} - \overline{f}_{y_i})^2.$$

This expression captures the variability in the firing rate map that is not explained by y (as $\overline{f}_{y_i}$ is conditioned on individual x), providing a quantification of all contributions of the variable x.

Practically, due to the presence of noise, proper cross-validation is needed. In particular, in the presence of measurement noise, Eq. 4.2 can lead to an overestimation of the contribution of a given variable. Even

if a variable y does not have any explanatory power for $f_{i,j}$, due to noise, $\sum_{i=1}^{m}\sum_{j=1}^{n}(f_{i,j}-\overline{f}_{y_i})^2$ will be non-zero. To address this problem, we developed cross-validated statistics to estimate the contribution of individual stimulus variables. In particular, we find that the following statistics provide an unbiased estimate of the variability of firing rates explained by variable x, under mild assumptions.

$$C_y = \frac{1}{2}\sum_{i=1}^{m_i}\sum_{j=1}^{n_j}[(f_{i,j}-\overline{f}_{y_j})^2 + (f_{i,j}^t-\overline{f}_{y_j}^t)^2 - (f_{i,j}^t-f_{i,j})^2 + (\overline{f}_{y_j}-\overline{f}_{y_j}^t)^2]. \quad (1)$$

Practically, we evenly partition the data into training and test sets for each bin of the joint firing rate map of (x, y). The statistics above are computed based on a given split. By partitioning the data multiple times and averaging these statistics over different partitions, the variance of the resulting estimator, denoted as $\overline{C_y}$, can be reduced. Here, m_i could be different for different i to capture the different number of bins after filtering out certain bins. Similarly, n_j could be different for different j.

We can also compute the contribution of the other variable x using a similar method. A ratio that summarizes the contributions of the two variables can also be computed. The method can be generalized to more than two variables. In this case, to estimate the contribution of a given stimulus variable, one would need to condition on all other variables. Practically, sufficient data would be needed to accurately estimate the contribution to response variability.

4.3 Neural network models

To investigate whether deep neural networks trained on vision tasks develop brain-like spatial representations, we evaluated models based on two distinct research frameworks. The first involves pretrained convolutional neural networks (CNNs) on image classification tasks, following the approach proposed in [27]. The second involves a "predictive coding" network trained to predict the next image frame based on previous image sequences during navigation [28]. For these two models, we developed a common implementation using the Minecraft environment.

CNNs

Luo et al. [27] investigated representations in ResNet-50, VGG-16, and ViT models. These deep neural networks are promising candidate models for the primate ventral visual stream [39, 38, 43, 44, 45]. The analysis in [27] aimed to determine whether processing egocentric visual information within these complex, non-spatial models could lead to allocentric spatial representations akin to those observed in the hippocampus. As in [27], we examined each model with weights both pretrained on ImageNet and randomly initialized. None of the models were further trained on the testing environment.

The experimental setup in [27] uses a three-dimensional virtual environment reminiscent of laboratory settings in neuroscience studies built in Unity. Instead of using the exact environment described in Luo et al. [27], we constructed a Minecraft environment with similar dimensions and found that this did not significantly impact findings. The same Minecraft environment was used for the predictive coding network

(described in the next section). This environment is a 17×17 block area, designed to minimize visual aliasing by adding distinctive features along the borders. An agent explored this environment using the Malmo platform, a Python package for AI agents to interact natively with Minecraft, collecting first-person perspective images for each two-dimensional location and head direction pair.

Activations from these CNN models were analyzed for spatial coding properties, using standard criteria to classify place, direction, and border cells [36, 22]. Notably, similar to previous work on place cells, the method used in [27] for identifying place cells did not require invariance to head direction. We adopted their criteria to classify units into putative "place cells" for further analysis.

In the Appendix of [27], it was stated that they adopted the place cell classification method of Tanni et al. [36], yet their implementation omits several components of the original procedure in [36]. Because our reanalysis indicated that many units labeled as place cells in Luo et al. [27] were low quality and should be excluded, we applied all applicable criteria from Tanni et al. [36]. One criterion from [36] does not apply to the analysis of model units. That is, in [36] place cells are required to exhibit pyramidal-like extracellular waveforms, spatial stability, at least one place field, and a low mean firing rate. The pyramidal-like waveform criterion pertains exclusively to spiking neurophysiology and has no analogue for artificial neural network (ANN) units; accordingly, it is excluded from our analysis. We note that [27] also discarded this criterion.

For spatial stability, Tanni et al. [36] compute Pearson correlations between rate maps constructed from odd versus even minutes ($r > 0.5$) and between the first and last halves of the recording ($r > 0.25$) in more than one recording. To translate the odd–even stability test to ANNs, we partition head-direction samples into odd and even segments, treat these as independent visits to the same spatial bins, and require $r > 0.5$ between the corresponding rate maps.

For spatial selectivity, Tanni et al. [36] define a place field as a contiguous (four-neighbor, non-diagonal) region comprising ≥ 10 spatial bins whose firing rate rises to a single prominent peak. Because standard 1 Hz thresholding can merge adjacent fields after smoothing, they employ, and we replicate, a custom iterative thresholding procedure: threshold the rate map at 1 Hz to identify candidate fields; then raise the threshold in fixed increments, re-identifying candidates at each step under the same validity rules. This yields a nested hierarchy in which higher-threshold candidates are contained within lower-threshold parents. Candidates are then examined from the smallest (highest-threshold) upward, discarding regions that are too large (occupying more than half of all bins in the map) or insufficiently stable. Within any overlapping stack, the lowest-threshold valid region is retained as the field; if a parent region contains multiple valid children, the parent is discarded and the children are retained as distinct fields. By contrast, the Luo et al. [27] procedure requires only the presence of any place field, defined as any contiguous set of > 10 bins that covers less than half of the map (for a 17×17 grid, < 145 bins; that is, $289/2 \approx 144$).

Networks trained on predictive coding

We also examined a recently proposed model for cognitive maps in the hippocampus by Gornet and Thomson [26]. This model uses a UNet-style encoder-decoder framework where the encoder and de-

coder are both ResNet models with skip connections, and it also includes several self-attention blocks in the latent space between the two. In contrast to the deep network models tested in [27], which are only either pretrained on ImageNet or untrained and produce features in response to a single image, the predictive coding model receives a sequence of first-person image observations from a trajectory of an agent moving around in the Minecraft environment and is trained to predict the next image in the sequence. This model is trained in the same environment across 200 epochs with gradient descent optimization with Nesterov momentum [46], a weight decay of $5 \cdot 10^{-6}$, and a learning rate of 10^{1} adjusted by OneCycle learning rate scheduling [47], with which it achieves an MSE of (To-do: add). Representations are taken after the last attention block and classified with the same criteria as in Luo et al. [27] to identify place cells.

An agent traversed the same walled Minecraft environment used in the Luo et al. [27] study and collected sequences of observations of length 10. In the original Gornet and Thomson paper [28], data were collected by selecting two random locations on a map and generating a traversal path between them using the A* algorithm. To promote more naturalistic movement and to expose the model to viewpoints not constrained to grid centers, our agent instead initialized at a random position within the environment and updated its movement direction by randomly sampling small changes to its angular velocity. The agent additionally executed minimal turns as needed to avoid obstacles.

We also evaluated a variant of the model in which head direction was decoupled from movement direction and oscillated across a 180° range. This contrasts with the Gornet and Thomson implementation, in which head and movement directions are coupled. Because this modification did not produce a substantial effect in our experiments, we do not include this variant in the present analysis.

4.4 Neural data

Place cells

In order to compare model "place cells" with their biological counterparts, we analyzed recordings from Pfeiffer et al. [48]. Data came from four adult male Long-Evans rats performing an open-field spatial memory task in a $2m \times 2m$ black arena enclosed by 30 cm walls with 36 identical, evenly spaced reward wells flush with the floor. Animals were pretrained on trials where, after consuming a reward at a randomly selected well, they were required to navigate back to a fixed "home" well to receive a reward. The location of the "home" well changed after each day. After pretraining, sessions were collected where the animal performed approximately 30 trials per session, ensuring broad sampling of the arena. No explicit cues signaled which well was filled; filling/emptying was performed silently via tubing beneath the floor.

Each rat was implanted with a lightweight microdrive array targeted to area CA1 of the dorsal hippocampus. Sessions commonly yielded 100–250 simultaneously recorded hippocampal units with spatially selective activity. Spike clusters were identified by manual clustering based on spike waveform peak amplitudes, with only well-isolated units included in the analysis, and putative interneurons were excluded by spike width and mean rate. Position and head direction were tracked continuously at 60 Hz using two

differently colored, head-mounted LEDs recorded by an overhead camera.

Our analysis performs a similar identification of place fields as in the original paper [48]. Position was binned at 8 cm, head direction was binned at 4°, and spikes were binned into 60 Hz time intervals. 3D place fields (X-Y-HD) were calculated as the histogram of firing activity normalized by the number of time bins that intersected with the location bin. Time bins where the animal was moving slower than 5 cm^{-1} were excluded. Visits to each positional bin were shuffled before splitting into two folds for use in the downstream method, and bins with fewer than 10 visits were discarded. We additionally required two folds of each rate map and calculated these by randomly shuffling the time-bin visits to each positional bin, splitting them into equal halves, and taking each half's average firing rate. Each rate map was smoothed with a Gaussian kernel (s.d. 4 cm, 4°) before analysis.

We additionally performed a decoding analysis on excitatory units with a similar pipeline as described in Pfeiffer et al. [48], extended to decode either 2D position or head direction. For this analysis, we use smaller positional bins of 2 cm and 1°. We used a memoryless Bayesian decoder with independent Poisson neurons and a uniform prior over position bins. For each discretized state s (either head-direction bin or 2D position bin), each neuron's tuning $f_i(s)$ (Hz) was estimated from occupancy-normalized firing rates, then smoothed (HD: circular Gaussian, s.d. 4 cm, 2°). Given spike counts $n_t = (n_{t,1}, \ldots, n_{t,N})$ in a non-overlapping window of duration $\tau = 250$ ms, we find the probability of the animal's position or head direction as

$$\Pr(\text{pos} \mid \text{spikes}) \;=\; \bigcup \Big/ \sum_{j=1}^{M} \bigcup$$

where

$$\bigcup \;=\; \left(\prod_{i=1}^{N} f_i(\text{pos})^{n_i}\right) \exp\left(-\tau \sum_{i=1}^{N} f_i(\text{pos})\right)$$

A time window of 250 ms was used to estimate the rat's position on a behavioral timescale. We performed 4-fold cross-validation by splitting up the time bins into four folds, creating rate maps from the spikes in one fold, and performing decoding over the remaining three.

Head direction cells

We analyzed the response properties of head-direction cells using the dataset reported in Duszkiewicz et al. [32]. In that study, mice were first habituated over several days to forage freely for small food rewards in an open field. During recordings, animals explored a modular arena built on a plastic platform (80×80 cm floor, 50 cm high walls) mounted in a larger metal frame ($90 \times 90 \times 180$ cm). The walls could be configured either as a square or a triangular enclosure; we only consider sessions recorded in the former. A white rectangular cue card placed on one wall served as the main distal visual landmark in the standard open-field sessions.

Neural activity was recorded using high-density silicon probes while animals moved freely. For postsubiculum (PoSub) recordings, a 64-channel linear array was implanted either vertically (yielding 931 units from 14 mice, 46–101 units per recording) or parallel to the PoSub cell layers (1,999 units from 18 mice, 42–185 units per recording). Recorded units were classified as putative excitatory cells or fast-spiking (FS) inhibitory cells based on waveform shape and mean firing rate. A large fraction of excitatory cells were head-direction cells: those whose HD information exceeded the 99th percentile of a time-reversed control distribution were labeled PoSub-HD cells ($n = 1,602$; 87% of excitatory units). Additional cells were recorded in area ADN of the thalamus; these were excluded from our analysis due to the small number of head direction cells recorded in square environments. During all recording sessions, the animals' position and head orientation were tracked at 100 Hz using seven infrared cameras plus an overhead camera and synchronized to the electrophysiology via digital pulses recorded on the acquisition board.

We applied the same general process for rate map calculation as in the place cell analysis, on cells classified as both excitatory and head-direction-coded, with bins of size 4 cm for XY and 4° for head direction as well as movement direction, and spikes binned into 100 ms time bins. In addition to calculating 3D rate maps across XY and HD, we also calculated 2D rate maps for head direction and movement direction. While head direction is provided as part of the positional data, we estimated instantaneous movement direction from the 2D positional trace by first computing frame-to-frame displacements in x and y and converting these into angles using the $\mathrm{atan2}$ function. The resulting angles were expressed in degrees and wrapped to the range [0, 360). To reduce frame-to-frame noise while respecting the circular nature of angles, we then applied a circular boxcar filter with a window of 20 samples to obtain the smoothed movement direction time series.

We performed the same Bayesian decoding process as above on XY, HD, and MD. Positional bin sizes were 2 cm, 1°, and 1°, respectively. The decoding was performed with a 50 ms window, and time bins with a velocity slower than 2 cm/s were excluded. Spikes were also smoothed before decoding to match their analysis with a Gaussian kernel (σ = 100 ms).

References

[1] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591, 1959.

[2] John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.

[3] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

[4] Gregory C DeAngelis, Izumi Ohzawa, and Ralph D Freeman. Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature*, 352(6331):156–159, 1991.

[5] R Christopher Decharms and Anthony Zador. Neural representation and the cortical code. *Annual review of neuroscience*, 23(1):613–647, 2000.

[6] Hanne Stensola, Tor Stensola, Trygve Solstad, Kristian Frøland, May-Britt Moser, and Edvard I Moser. The entorhinal grid map is discretized. *Nature*, 492(7427):72–78, 2012.

[7] Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, 2021.

[8] Jeffrey S Taube, Robert U Muller, and James B Ranck. Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435, 1990.

[9] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford: Clarendon Press, 1978.

[10] Emilio Kropff, James E Carmichael, May-Britt Moser, and Edvard I Moser. Speed cells in the medial entorhinal cortex. *Nature*, 523(7561):419–424, 2015.

[11] Francesca Sargolini, Marianne Fyhn, Torkel Hafting, Bruce L McNaughton, Menno P Witter, May-Britt Moser, and Edvard I Moser. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312(5774):758–762, 2006.

[12] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[13] Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*, 37:66–74, 2016.

[14] Kiah Hardcastle, Niru Maheswaranathan, Surya Ganguli, and Lisa M Giocomo. A multiplexed, heterogeneous, and adaptive code for navigation in medial entorhinal cortex. *Neuron*, 94(2):375–387, 2017.

[15] Bruce L McNaughton, Francesco P Battaglia, Ole Jensen, Edvard I Moser, and May-Britt Moser. Path integration and the neural basis of the'cognitive map'. *Nature Reviews Neuroscience*, 7(8):663–678, 2006.

[16] Charles Darwin. Origin of certain instincts. *Nature*, 7:417–418, 1873.

[17] Kiah Hardcastle, Surya Ganguli, and Lisa M Giocomo. Environmental boundaries as an error correction mechanism for grid cells. *Neuron*, 86(3):827–839, 2015.

[18] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

[19] Alexei Samsonovich and Bruce L McNaughton. Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, 17(15):5900–5920, 1997.

[20] Yoram Burak and Ila R Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS computational biology*, 5(2):e1000291, 2009.

[21] Christopher J Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *the International Conference on Learning Representations*, 2018.

[22] Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, Greg Wayne, Hubert Soyer, Fabio Viola, Brian Zhang, Ross Goroshin, Neil Rabinowitz, Razvan Pascanu, Charlie Beattie, Stig Petersen, Amir Sadik, Stephen Gaffney, Helen King, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, and Dharshan Kumaran. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

[23] Ben Sorscher, Gabriel C Mel, Samuel A Ocko, Lisa Giocomo, and Surya Ganguli. A unified theory for the computational and mechanistic origins of grid cells. *bioRxiv*, 2020.

[24] Ruiqi Gao, Jianwen Xie, Xue-Xin Wei, Song-Chun Zhu, and Ying Nian Wu. On path integration of grid cells: Group representation and isotropic scaling. *Advances in Neural Information Processing Systems*, 34:28623–28635, 2021.

[25] Christopher J Cueva, Peter Y Wang, Matthew Chin, and Xue-Xin Wei. Emergence of functional and structural properties of the head direction system by optimization of recurrent neural networks. In *International Conference on Learning Representations*, 2020.

[26] James Gornet and Matt Thomson. Automated construction of cognitive maps with visual predictive coding. *Nature Machine Intelligence*, 6(7):820–833, 2024.

[27] Xiaoliang Luo, Robert M Mok, and Bradley C Love. The inevitability and superfluousness of cell types in spatial cognition. *bioRxiv*, 2024.

[28] James Gornet and Matt Thomson. Automated construction of cognitive maps with visual predictive coding. *Nature Machine Intelligence*, 6:820–833, 2024.

[29] Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

[30] Brad E Pfeiffer and David J Foster. Autoassociative dynamics in the generation of sequences of hippocampal place cells. *Science*, 349(6244):180–183, 2015.

[31] Adrian J Duszkiewicz, Pierre Orhan, Sofia Skromne Carrasco, Eleanor H Brown, Eliott Owczarek, Gilberto R Vite, Emma R Wood, and Adrien Peyrache. Local origin of excitatory–inhibitory tuning equivalence in a cortical network. *Nature Neuroscience*, 27(4):782–792, 2024.

[32] Adrian J. Duszkiewicz, Pierre Orhan, Sofia Skromne Carrasco, Eleanor H. Brown, Eliott Owczarek, Gilberto R. Vite, Emma R. Wood, and Adrien Peyrache. Local origin of excitatory–inhibitory tuning equivalence in a cortical network. *Nature Neuroscience*, 27:782–792, 2024.

[33] Bruce L McNaughton, Francesco P Battaglia, Ole Jensen, Edvard I Moser, and May-Britt Moser. Path integration and the neural basis of the'cognitive map'. *Nature Reviews Neuroscience*, 7(8):663–678, 2006.

[34] Mandyam Veerambudi Srinivasan, Simon Barry Laughlin, and Andreas Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1205):427–459, 1982.

[35] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[36] Sander Tanni, William de Cothi, and Caswell Barry. State transitions in the statistically stable place cell population correspond to rate of perceptual change. *Current Biology*, 32(16):3505–3514.e7, 2022.

[37] Margaret C von Ebers and Xue-Xin Wei. Cognitive maps from predictive vision. *Nature Machine Intelligence*, 6(8):850–851, 2024.

[38] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[39] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[40] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016.

[41] Jonathan J Couey, Aree Witoelar, Sheng-Jia Zhang, Kang Zheng, Jing Ye, Benjamin Dunn, Rafal Czajkowski, May-Britt Moser, Edvard I Moser, and Yasser Roudi. Recurrent inhibitory circuitry as a mechanism for grid formation. *Nature neuroscience*, 16(3):318–324, 2013.

[42] Florian Raudies, Mark P Brandon, G William Chapman, and Michael E Hasselmo. Head direction is coded more strongly than movement direction in a population of entorhinal neurons. *Brain research*, 1621:355–367, 2015.

[43] Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[44] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

[45] Astrid A. Zeman, J. Brendan Ritchie, Stefania Bracci, and Hans Op de Beeck. Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Scientific Reports*, 10(1):2453, 2020.

[46] Ilya Sutskever, James Martens, George Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *ICML*, pages 1139–1147. PMLR, 2013.

[47] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006 of *Proceedings of SPIE*, pages 369–386. SPIE, 2019.

[48] Brad E. Pfeiffer and David J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79, 2013.