# EMERGENCE OF GRID-LIKE REPRESENTATIONS BY TRAINING RECURRENT NEURAL NETWORKS TO PERFORM SPATIAL LOCALIZATION

**Christopher J. Cueva**[*], **Xue-Xin Wei**[*]
Columbia University
New York, NY 10027, USA
{ccueva,weixxpku}@gmail.com

## ABSTRACT

Decades of research on the neural code underlying spatial navigation have revealed a diverse set of neural response properties. The Entorhinal Cortex (EC) of the mammalian brain contains a rich set of spatial correlates, including grid cells which encode space using tessellating patterns. However, the mechanisms and functional significance of these spatial representations remain largely mysterious. As a new way to understand these neural representations, we trained recurrent neural networks (RNNs) to perform navigation tasks in 2D arenas based on velocity inputs. Surprisingly, we find that grid-like spatial response patterns emerge in trained networks, along with units that exhibit other spatial correlates, including border cells and band-like cells. All these different functional types of neurons have been observed experimentally. The order of the emergence of grid-like and border cells is also consistent with observations from developmental studies. Together, our results suggest that grid cells, border cells and others as observed in EC may be a natural solution for representing space efficiently given the predominant recurrent connections in the neural circuits.

## 1 INTRODUCTION

Understanding the neural code in the brain has long been driven by studying feed-forward architectures, starting from Hubel and Wiesel's famous proposal on the origin of orientation selectivity in primary visual cortex (Hubel & Wiesel, 1962). Inspired by the recent development in deep learning (Krizhevsky et al., 2012; LeCun et al., 2015; Hochreiter & Schmidhuber, 1997; Mnih et al., 2015), there has been a burst of interest in applying deep feedforward models, in particular convolutional neural networks (CNN) (LeCun et al., 1998), to study the sensory systems, which hierarchically extract useful features from sensory inputs (see *e.g.*, Yamins et al. (2014); Kriegeskorte (2015); Kietzmann et al. (2017); Yamins & DiCarlo (2016)).

For more cognitive tasks, neural systems often need to maintain certain internal representations of relevant variables in the absence of external stimuli- a process that requires more than feature extraction. We will focus on spatial navigation, which typically requires the brain to maintain a representation of self-location and update it according to the animal's movements and landmarks of the environment. Physiological studies done in rodents and other mammals (including humans, non-human primates and bats) have revealed a variety of neural correlates of space in Hippocampus and Entorhinal Cortex (EC), including place cells (O'Keefe, 1976), grid cells (Fyhn et al., 2004; Hafting et al., 2005; Fyhn et al., 2008; Yartsev et al., 2011; Killian et al., 2012; Jacobs et al., 2013), along with border cells (Solstad et al., 2008), band-like cells (Krupic et al., 2012) and others (see Figure 1a). In particular, each grid cell only fires when the animal occupies a distinct set of physical locations, and strikingly these locations lie on a lattice. The study of the neural underpinning of spatial cognition has provided an important window into how high-level cognitive functions are supported in the brain (Moser et al., 2008; Aronov et al., 2017).

[*] equal contribution

How might the spatial navigation task be solved using a network of neurons? Recurrent neural networks (RNNs) (Hochreiter & Schmidhuber, 1997; Graves et al., 2013; Oord et al., 2016; Theis & Bethge, 2015; Gregor et al., 2015; Sussillo et al., 2015) seem particularly useful for these tasks. Indeed, recurrent-based continuous attractor networks have been one popular type of models proposed for the formation of grid cells (McNaughton et al., 2006; Burak & Fiete, 2009; Couey et al., 2013) and place cells (Samsonovich & McNaughton, 1997). Such models have provided valuable insights into one set of possible mechanisms that could support the formation of the grids. However, these models typically rely on fine-tuned connectivity patterns, in particular the models need a subtle yet systematic asymmetry in the connectivity pattern to move the attractor state according to the animal's own movement. The existence of such a specific 2D connectivity in rodent EC remains unclear. Additionally, previous models have mainly focused on grid cells, while other types of responses that co-exist in the Entorhinal Cortex have been largely ignored. It would be useful to have a unified model that can simultaneously explain different types of neural responses in EC.

Motivated by these considerations, here we present an alternative modeling approach for understanding the representation of space in the neural system. Specifically, we trained a RNN to perform some spatial navigation tasks. By leveraging the recent development in RNN training and knowledge of the navigation system in the brain, we show that training a RNN with biologically relevant constraints naturally gives rise to a variety of spatial response profiles as observed in EC, including grid-like responses. To our knowledge, this is the first study to show that grid-like responses could emerge from training a RNN to perform navigation.

Our result implies that the neural representation in EC may be seen as a natural way for the brain to solve the navigation task efficiently (Wei et al., 2015). More generally, it suggests that RNNs can be a powerful tool for understanding the neural mechanisms of certain high-level cognitive functions.

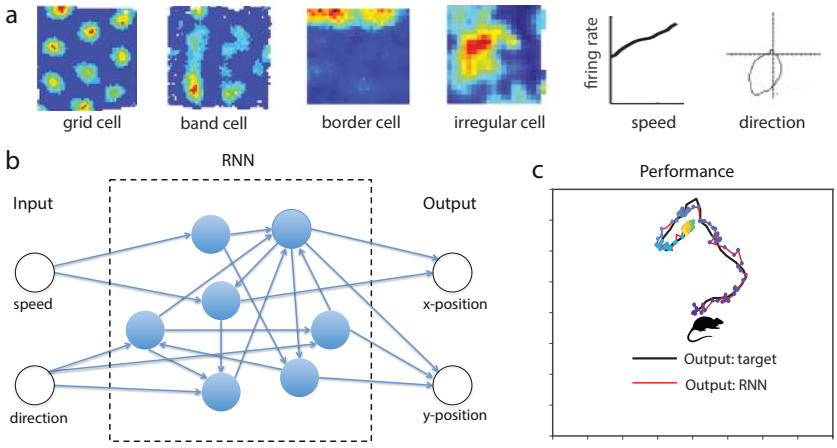



Figure 1: **a)** Example neural data showing different kinds of neural correlates underlying spatial navigation in EC. All figures are replotted from previous publications. From left to right: a "grid cell" recorded when an animal navigates in a square environment, replotted from Krupic et al. (2012), with the heat map representing the firing rate of this neuron as a function of the animal's location (red corresponds to high firing rate); a "band-like" cell from Krupic et al. (2012); a border cell from Solstad et al. (2008); an irregular spatially tuned cell from Diehl et al. (2017); a "speed cell" from Kropff et al. (2015), which exhibits roughly linear dependence on the rodent's running speed; a "heading direction cell" from Sargolini et al. (2006), which shows systematic change of firing rate depending on animal's heading direction. **b)** The network consists of $N = 100$ recurrently connected units (or neurons) which receive two external inputs, representing the animal's speed and heading direction. The two outputs linearly weight the neurons in the RNN. The goal of training is to make the responses of the two output neurons accurately represent the animal's physical location. **c)** Typical trajectory after training. As shown, the output of the RNN can accurately, though not perfectly, track the animal's location during navigation.

# 2 MODEL

## 2.1 MODEL DESCRIPTION

Our network model consists of a set of recurrently connected units ($N = 100$). The dynamics of each unit in the network $u_i(t)$ is governed by the standard continuous-time RNN equation:

$$\tau \frac{dx_i(t)}{dt} = -x_i(t) + \sum_{j=1}^{N} W_{ij}^{\text{rec}} u_j(t) + \sum_{k=1}^{N_{\text{in}}} W_{ik}^{\text{in}} I_k(t) + b_i + \xi_i(t) \tag{1}$$

for $i = 1, \ldots, N$. The activity of each unit, $u_i(t)$, is related to the activation of that unit, $x_i(t)$, through a nonlinearity which in this study we take to be $u_i(t) = \tanh(x_i(t))$. Each unit receives input from other units through the recurrent weight matrix $W^{\text{rec}}$ and also receives external input, $I(t)$, that enters the network through the weight matrix $W^{\text{in}}$. Each unit has two sources of bias, $b_i$ which is learned and $\xi_i(t)$ which represents noise intrinsic to the network and is taken to be Gaussian with zero mean and constant variance. The network was simulated using the Euler method for $T = 500$ timesteps of duration $\tau/10$.

To perform a 2D navigation task with the RNN, we linearly combine the firing rates of units in the network to estimate the current location of the animal. The responses of the two linear readout neurons, $y_1(t)$ and $y_2(t)$, are given by the following equation:

$$y_j(t) = \sum_{i=1}^{N} W_{ji}^{\text{out}} u_i(t) \tag{2}$$

## 2.2 INPUT TO THE NETWORK

The network inputs and outputs were inspired by simple spatial navigation tasks in 2D open environments. The task resembles dead-reckoning (sometimes referred to as path integration), which is ethologically relevant for many animal species (Darwin, 1873; Mittelstaedt & Mittelstaedt, 1980; Etienne & Jeffery, 2004; McNaughton et al., 2006). To be more specific, the inputs to the network were the animal's speed and direction at each time step. Experimentally, it has been shown that the velocity signals exist in EC (Sargolini et al., 2006; Kropff et al., 2015; Hinman et al., 2016), and there is also evidence that such signals are necessary for grid formation (Winter et al., 2015a;b).

Throughout the paper, we adopt the common assumption that the head direction of the animal coincides with the actual moving direction. The outputs were the x- and y-coordinates of the integrated position. The direction of the animal is modeled by modified Brownian motion to increase the probability of straight-runs, in order to be consistent with the typical rodent's behavior in an open environment. The usage of such simple movement statistics has the advantage of having full control of the simulated trajectories. However, for future work it would be very interesting to test the model using different animals' real movement trajectories to see how the results might change.

Special care is taken when the animal is close to the boundary. The boundary of the environment will affect the statistics of the movement, as the animal cannot cross the boundary. This fact was reflected in the model by re-sampling the angular input variable until the input angle did not lead the animal outside the boundary. In the simulations shown below, the animal always starts from the center of the arena, but we verified that the results are insensitive to the starting locations.

## 2.3 TRAINING

We optimized the network parameters $W^{\text{rec}}$, $W^{\text{in}}$, $b$ and $W^{\text{out}}$ to minimize the squared error in equation (3) between target x- and y-coordinates from a two dimensional navigation task (performed in rectangular, hexagonal, and triangular arenas) and the network outputs generated according to equation (2).

$$E = \frac{1}{MTN_{\text{out}}} \sum_{m,t,j=1}^{M,T,N_{\text{out}}} (y_j(t,m) - y_j^{\text{target}}(t,m))^2 \tag{3}$$

Parameters were updated with the Hessian-free algorithm (Martens & Sutskever, 2011) using mini-batches of size $M = 500$ trials. In addition to minimizing the error function in equation (3) we

regularized the input and output weights according to equation (4) and the squared firing rates of the units (referred to as metabolic cost) according to equation (5). In sum, the training aims to minimize a loss function, that consists of the error of the animal, the metabolic cost, and a penalty for large network parameters.

$$R_{L2} = \frac{1}{NN_{\text{in}}} \sum_{i,j=1}^{N,N_{\text{in}}} (W_{ij}^{\text{in}})^2 + \frac{1}{NN_{\text{out}}} \sum_{i,j=1}^{N_{\text{out}},N} (W_{ij}^{\text{out}})^2 \tag{4}$$

$$R_{FR} = \frac{1}{NTM} \sum_{i,t,m=1}^{N,T,M} u_i(t,m)^2 \tag{5}$$

We find that the results are qualitatively insensitive to the initialization schemes used for the recurrent weight matrix $W^{\text{rec}}$. For the results presented in this paper, simulations in the hexagonal environment were obtained by initializing the elements of $W^{\text{rec}}$ to be zero mean Gaussian random variables with variance $1.5^2/N$, and simulations in the square and triangular environments were initialized with an orthogonal $W^{\text{rec}}$ (Saxe et al., 2014). We initialized the bias $b$ and output weights $W^{\text{out}}$ to be zero. The elements of $W^{\text{in}}$ were zero mean Gaussian variables with variance $1/N_{\text{in}}$.

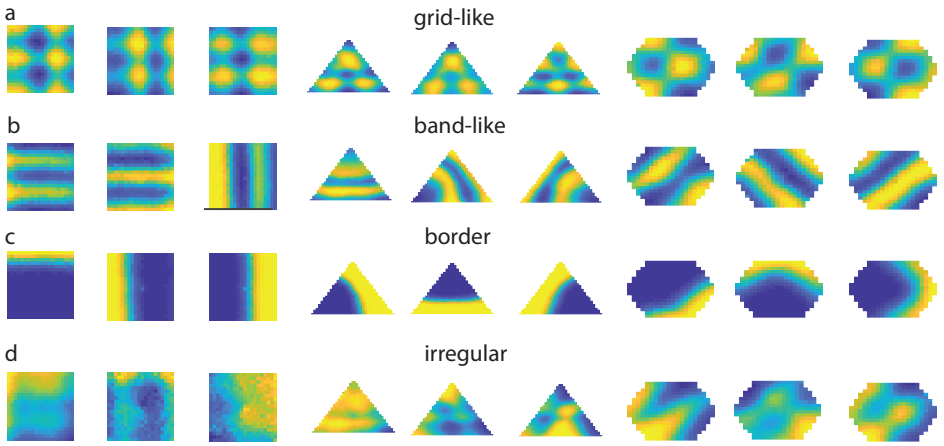



Figure 2: Different types of spatial selective responses of units in the trained RNN. Example simulation results for three different environments (square, triangular, hexagon) are presented. Blue (yellow) represents low (high) activity. **a)** Grid-like responses. **b)** Band-like responses; **c)** Border-related responses; **d)** Spatially irregular responses. These responses can be spatially selective but they do not form a regular pattern defined in the conventional sense.

## 3 RESULTS

We run simulation experiments in arenas with different boundary shapes, including square, triangular and hexagonal. Figure 1c shows a typical example of the model performance after training; the network (red trace) accurately tracks the animal's actual path (black).

### 3.1 TUNING PROPERTIES OF THE MODEL NEURONS

We are mostly interested in what kind of representation the RNN has learned to solve this navigation task, and whether such a representation resembles the response properties of neurons in EC (Moser et al., 2008).

#### 3.1.1 SPATIAL TUNING

To test whether the trained RNN developed location-selective representations, we plot individual neurons' mean activity level as a function of the animal's location during spatial exploration. Note

that these average response profiles should not be confused with the linear filters typically shown in feedforward networks. Surprisingly, we find neurons in the trained RNN show a range of interesting spatial response profiles. Examination of these response profiles suggests they can be classified into distinct functional types. Importantly, as we will show, these distinct spatial response profiles can be mapped naturally to known physiology in EC. The spatial responses of all units in trained networks are shown in the Appendix.

**Grid-like responses** Most interestingly, we find some of the units in the RNN exhibit clear grid-like responses (Figure 2a). These firing patterns typically exhibit multiple firing fields, with each firing field exhibiting roughly circular symmetric or ellipse shape. Furthermore, the firing fields are highly structured, *i.e.*, when combined, are arranged on a regular lattice. Furthermore, the structure of the response lattice depends on the shape of the boundary. In particular, training the network to perform self-localization in a square environment tends to give rectangular grids. In hexagonal and triangular environments, the grids are closer to triangular.

Experimentally, it is shown that (medial) EC contains so-called grid cells which exhibit multiple firing fields that lie on a regular grid (Fyhn et al., 2004; Hafting et al., 2005). The grid-like firing patterns in our simulation are reminiscent of the grid cells in rodents and other mammals. However, we also notice that the the grid-like model responses typically exhibit few periods, not as many as experimental data (see Figure 1a). It is possible that using a larger network might reveal finer grid-patterns in our model. Nonetheless, it is surprising that the gird-like spatial representations can develop in our model, given there is no periodicity in the input. Another potential concern is that, experimentally it is reported that the grids are often on the corners of a triangular lattice (Hafting et al., 2005) even in square environments (see Figure 1a), though the grids are somewhat influenced by the shape of the environment. However, the rats in these experiments presumable had spatial experience in other environments with various boundary shapes. Experimentally, it would be interesting to see if grid cells would lie on a square lattice instead if the rats are raised in a single square environment - a situation we are simulating here.

**Border responses** Many neurons in the RNN exhibit selectivity to the boundary (Figure 2c). Typically, they only encode a portion of the boundary, e.g. one piece of wall in a square shaped environment. Such properties are similar to the border cells discovered in rodent EC (Solstad et al., 2008; Savelli et al., 2008; Lever et al., 2009). Experimentally, border cells mainly fire along one piece of wall, although some have been observed to fire along multiple borders or along the whole boundary of the environment; interestingly, these multi-border responses were also observed in some RNN models. Currently, it is unclear how the boundary-like response profiles emerge (Solstad et al., 2008; Savelli et al., 2008; Lever et al., 2009). Our model points to the possibility that the border cells may emerge without the presence of tactile cues. Furthermore, it suggests that border cell formation may be related to the movement statistics of the animals, i.e. due to the asymmetry of the movement statistics along the boundary.

**Band-like responses** Interestingly, some neurons in the RNN exhibit band-like responses (Figure 2b). In most of our simulations, these bands tend to be parallel to one of the boundaries. For some of the units, one of the bands overlaps the boundary, but for others, that is not the case. Experimentally, neurons with periodic-like firing patterns have been recently reported in rodent EC. In one study, it has been reported that a substantial portion of cells in EC exhibit band-like firing characteristics (Krupic et al., 2012). However, we note that based on the reported data in Krupic et al. (2012), the band pattern is not as clear as in our model.

**Spatially-stable but non-regular responses** Besides the units described above, most of the remaining units also exhibit stable spatial responses, but they do not belong to the above categories. These response profiles can exhibit either one large irregular firing field; or multiple circular firing fields, but these firing fields do not show a regular pattern. Experimentally these types of cells have also been observed. In fact, it is recently reported that the non-grid spatial cells constitute a large portion of the neurons in Layer II and III of rodent EC (Diehl et al., 2017).

#### 3.1.2 SPEED TUNING AND HEAD DIRECTION TUNING

**Speed tuning** We next ask how neurons in the RNN are tuned to the inputs. Many of the model neurons exhibit linear responses to the running speed of the animal, while some neurons show no selectivity to speed, as suggested by the near-flat response functions. Example response profiles are

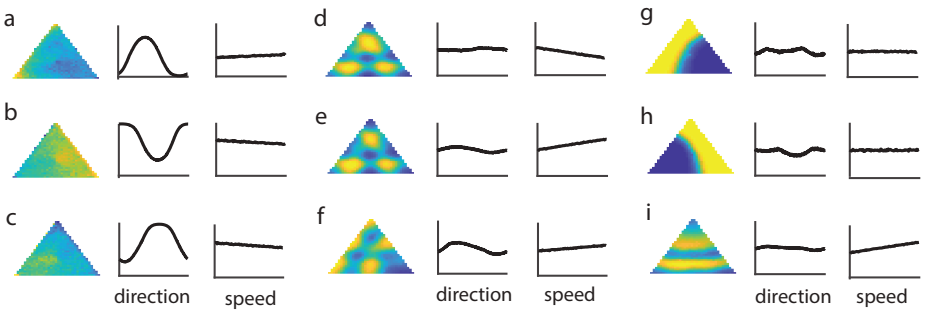



Figure 3: Direction tuning and speed tuning for nine example units in an RNN trained in a triangular arena. For each unit, we show the spatial tuning, (head) directional tuning, speed tuning respectively, from left to right. **a,b,c)** The three model neurons show strong directional tuning, but the spatial tuning is weak and irregular. The three neurons also exhibit linear speed tuning. **d,e,f)** The three neurons exhibit grid-like firing patterns, and clear speed tuning. The strength of their direction tuning differ. **g,h)** Border cells exhibit weak and a bit complex directional tuning and almost no speed tuning. **i)** This band cell shows weak directional tuning, but strong speed tuning.

shown in Figure 3. Interestingly, we observe that the model border cells tend to have almost zero speed-tuning (e.g., see Figure 3g,h).

**Head direction tuning** A substantial portion of the model neurons show direction tuning. There are a diversity of direction tuning profiles, both in terms of the strength of the tuning and their preferred direction. Example tuning curves are shown in Figure 3, and the direction tuning curves of a complete population are shown in the Appendix. Interestingly, in general model neurons which show the strongest head direction tuning do not show a clear spatial firing pattern (see Figure 3a,b,c). This suggests that there are a group of neurons which are mostly responsible for encoding the direction. We also notice that neurons with clear grid-like firing can exhibit a variety of direction tuning strengths, from weak to strong (Figure 3d,e,f). In the Appendix, we quantify the relation between these different tuning properties at the whole population level, which show somewhat complex dependence.

Experimentally, the heading direction tuning in EC is well-known, *e.g.*, Sargolini et al. (2006). Both the grid and non-grid cells in EC exhibit head direction tuning (Sargolini et al., 2006). Furthermore, the linear speed dependence of the model neurons is similar to the properties of speed cells reported recently in EC (Kropff et al., 2015). Our result is also consistent with another recent study reporting that the majority of neurons in EC exhibit some amount of speed tuning (Hinman et al., 2016). It remains an open question experimentally, at a population level, how different types of tuning characteristics in EC relate to each other.

### 3.1.3 Development of the tuning properties

We next investigate how the spatial response profiles evolve as learning/training progresses. We report two main observations. First, neurons that fire selectively along the boundary typically emerge first. Second, the grid-like responses with finer spatial tuning patterns only emerge later in training. For visualization, we perform dimensionality reduction using the t-SNE algorithm (Maaten & Hinton, 2008). This algorithm embeds 100 model neurons during three phases of training (early, intermediate, and late) into a two-dimensional space according to the similarity of their temporal responses. Here the similarity metric is taken to be firing rate correlation. In this 2D space as shown in Figure 4a, border cell representations appear early and stably persist through the end of training. Furthermore, early during training all responses are similar to the border related responses. In contrast, grid-like cells typically undergo a substantial change in firing pattern during training before settling into their final grid-like representation (Figure 4b).

The developmental time line of the grid-like cells and border cells is roughly consistent with developmental studies in rodents. Experimentally, it is known that border cells emerge earlier in development, and they exist at about 2 weeks after the rat is born (Bjerknes et al., 2014). The grid

cells mature only at about 4 weeks after birth (Langston et al., 2010; Wills et al., 2010; Bjerknes et al., 2014). Furthermore, our simulations suggest the reason why border cells emerge earlier in development may be that computationally it is easier to wire-up a network that gives rise to border cell responses.

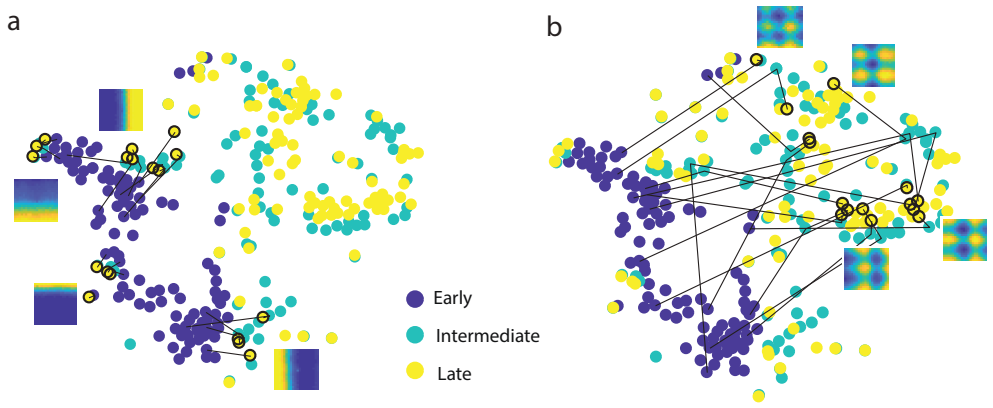



Figure 4: Development of border cells and grid-like cells. Early during training all responses are similar to the border related responses, and only as training continues do the grid-like cells emerge. We perform dimensionality reduction using the t-SNE algorithm on the firing rates of the neurons. Each dot represents one neuron ($N = 100$), and the color represents different training stages (early/intermediate/late shown in blue/cyan/yellow). Each line shows the trajectory of a single highlighted neuron as its firing responses evolve during training. In panel **a)**, we highlight the border representation. It appears there are four clusters of border cells, each responding to one wall of a square environment (spatial responses from four of these border cells are inset). These cells' response profiles appear early and stably persist through training, illustrated by the short distance they travel in this space. In **b)**, we show that the neurons which eventually become grid cells initially have tuning profiles similar to the border cells but then change their tuning substantially during learning. As a natural consequence, they need to travel a long distance in this space between the early and late phase of the training. Spatial responses are shown for four of these grid-like cells during the late phase of training.

### 3.2 The importance of regularization

We find appropriate regularizations of the RNN to be crucial for the emergence of grid-like representations. We only observed grid-like representations when the network was encouraged to store information while perturbed by noise. This was accomplished by setting the speed input to zero, e.g. zero speed 90% of the time, and adding Gaussian noise to the network ($\xi_i(t)$ in equation (1)); the precise method for setting the speed input to zero and the value of the noise variance is not crucial for our simulations to develop grid-like representations. The cost function which aims to capture the penalization on the metabolic cost of the neural activity also acts as an important regularization. Our simulations show that the grid-like representation did not emerge without this metabolic cost. In Figure 5, we show typical simulation results for a square environment, with and without proper metabolic regularization. In the Appendix, we illustrate the effect of regularization further, in particular the role of injecting noise into the RNN units.

Our results are consistent with the general notion on the importance of incorporating proper constraint for learning useful representations in neural networks (Bengio et al., 2013). Furthermore, it suggests that, to learn a model with response properties similar to neural systems it may be necessary to incorporate the relevant constraints, *e.g.*, noise and metabolic cost.

### 3.3 Error correction around the boundary

One natural question is whether the trained RNNs are able to perform localization when the path length exceeds the typical length used during training (500 steps), in particular given that noise in

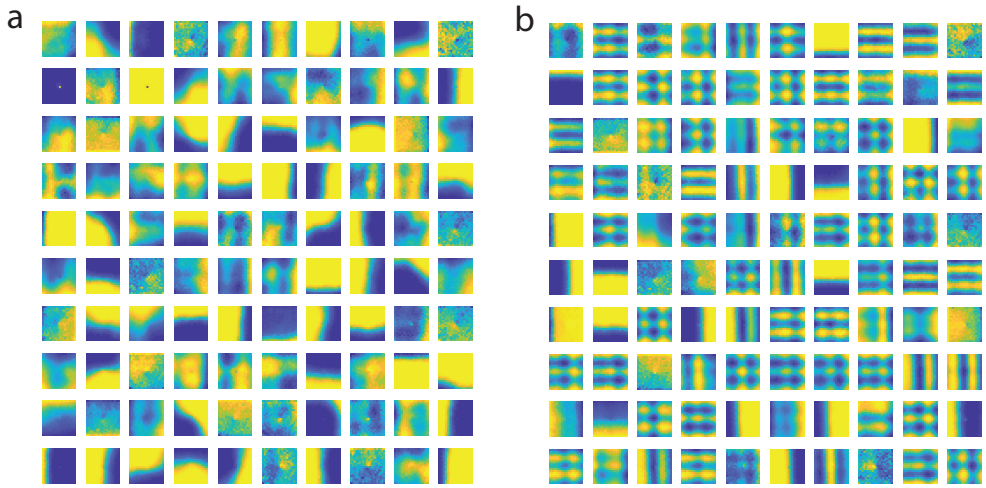



Figure 5: Complete set of spatial response profiles for 100 neurons in a RNN trained in a square environment. **a)** Without proper regularization, complex and periodic spatial response patterns do not emerge. **b)** With proper regularization, a rich set of periodic response patterns emerge, including grid-like responses. Regularization can also be adjusted to achieve spatial profiles intermediate between these two examples.

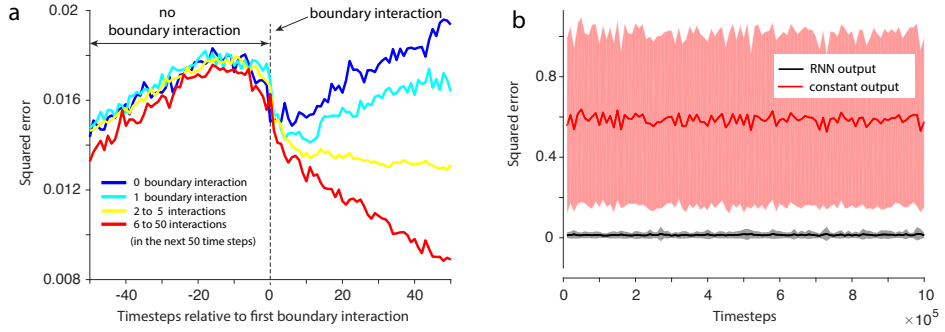



Figure 6: Error-correction happens at the boundary and the error is stable over time. At the boundary, the direction is resampled to avoid input velocities that lead to a path extending beyond the boundary of the environment. These changing input statistics at the boundary, termed a boundary interaction, are the only cue the RNN receives about the boundary. We find that the RNN uses the boundary interactions to correct the accumulated error between the true integrated input and its prediction based on the linear readout of equation (2). Panel **a)**, the mean squared error increases when there are no boundary interactions, but then decreases after a boundary interaction, with more boundary interactions leading to greater error reduction. In the absence of further boundary interaction, the squared error would gradually increase again (blue curve) at roughly a constant rate. **b)** The network was trained using mini-batches of 500 timesteps but has stable error over a duration at least four orders of magnitude larger. The error of the RNN output (mean and standard deviation shown in black, computed based on 10000 timesteps) is compared to the error that would be achieved by an RNN outputting the best constant values (red).

the network would gradually accumulate, leading to a decrease in localization performance. We test this by simulating paths that are several orders of magnitude longer. Somewhat surprisingly, we find the RNNs still perform well (Figure 6b). In fact, the squared error (averaged over every 10000 steps) is stable. The spatial response profiles of individual units also remain stable. This implies that the RNNs have acquired intrinsic error-correction mechanisms during training.

As shown earlier, during training some of the RNN units develop boundary-related firing (Figure 2c), presumably by exploiting the change of input statistics around the boundary. We hypothesize that

boundary interactions may enable error-correction through signals based on these boundary-related activities. Indeed, we find that boundary interactions can dramatically reduce the accumulated error (Figure 6a). Figure 6a shows that, without boundary interactions, on average the squared error grows roughly linearly as expected, however, interactions with the boundaries substantially reduce the error, and more frequent boundary interactions can reduce the error further. Error-correction on grid cells via boundary interactions has been proposed (Hardcastle et al., 2015; Pollock et al., 2017), however, we emphasize that the model proposed here develops the grid-like responses, boundary responses and the error-correction mechanisms all within the same neural network, thus potentially providing a unifying account of a diverse set of phenomena.

## 4 DISCUSSION

In this paper, we trained RNNs to perform path integration (dead-reckoning) in 2D arenas. We found that after training RNNs with appropriate regularization, the model neurons exhibit a variety of spatial and velocity tuning profiles that match neurophysiology in EC. What's more, there is also similarity in terms of when these distinct neuron types emerge during training/development. The EC has long been thought to be involved in path integration and localization of the animal's location (Moser et al., 2008). The general agreement between the different response properties in our model and the neurophysiology provide strong evidence supporting the hypothesis that the neural population in EC may provide an efficient code for representation self-locations based on the velocity input.

Recently, there has been increased interest in using complex neural network models to understand the neural code. But the focus has been on using feedforward architectures, in particular CNNs (LeCun et al., 1998). Given the abundant recurrent connections in the brain, it seems a particularly fruitful avenue to take advantage of the recent development in RNNs to help with neuroscience questions (Mante et al., 2013; Song et al., 2016; Miconi, 2017; Sussillo et al., 2015). Here, we only show one instance following this approach. However, the insight from this work could be general, and potentially useful for other cognitive functions as well.

The finding that metabolic constraints lead to the emergence of grid-like responses may be seen as conceptually related to the efficient coding hypothesis in visual processing (Barlow, 1961), in particular the seminal work on the emergence of the V1-like Gabor filters in a sparse coding model by Olshausen & Field (1996). Indeed, our work is partly inspired by these results. While there are conceptual similarities, however, we should also note there are differences between the sparse coding work and ours. First, the sparsity constraint in sparse coding can be naturally viewed as a particular prior while in the context of the recurrent network, it is difficult to interpret that way. Second, the grid-like responses are not the most sparse solution one could imagine. In fact, they are still quite dense compared to a more spatially localized representation. Third, the grid-like patterns that emerged in our network are not filters based on the raw input, rather the velocity inputs need to be integrated first in order to encode spatial locations. Our work is also inspired by recent work using the efficient coding idea to explain the functional architecture of the grid cells (Wei et al., 2015). It has been shown that efficient coding considerations could explain the particular set of grid scales observed in rodents (Stensola et al., 2012). However, in that work, the firing patterns of the neurons are assumed to have a lattice structure to start with. Furthermore, our work is related to the study by Sussillo and others (Sussillo et al., 2015), in which they show that regularization of RNN models are important for generating solutions that are similar to the neural activity observed in motor cortex. In Sussillo et al., a *smoothness* constraint together with others lead to simple oscillatory neural dynamics that well matches the neural data. We have not incorporated a smoothness constraint into our network.

Additionally, we note that there are a few recent studies which use place cells as the input to generate grid cells (Dordek et al., 2016; Stachenfeld et al., 2016), which are fundamentally different from our work. In these feedforward network models, the grid cells essentially perform dimensionality reduction based on the spatial input from place cells. However, the main issue with these models is that, it is unclear how place cells acquire spatial tuning in the first place. To the contrary, our model takes the animal's velocity as the input, and addresses the question of how the spatial tuning can be generated from such input, which are known to exist in EC (Sargolini et al., 2006; Kropff et al., 2015). In another related study (Kanitscheider & Fiete, 2016), the authors train a RNN with

LSTM units (Hochreiter & Schmidhuber, 1997) to perform different navigation tasks. However, no grid-like spatial firing patterns are reported.

Although our model shows a qualitative match to the neural responses observed in the EC, nonetheless it has several major limitations, with each offering interesting future research directions. First, the learning rule we use seems to be biologically implausible. We are interested in exploring how a more biologically plausible learning rule could give rise to similar results (Lillicrap et al., 2016; Miconi, 2017; Guerguiev et al., 2017). Second, the simulation results do not show a variety of spatial scales in grid-like cells. Experimentally, it is known that grid cells have multiple spatial scales, that scale geometrically with a ratio $1.4$ (Stensola et al., 2012), and this particular scale ratio is predicted by efficient coding of space (Wei et al., 2015). We are investigating how to modify the model to get a hierarchy of spatial scales, perhaps by incorporating more neurons or modifying the regularization. Last but not least, we have focused on the representation produced by the trained RNN. An equally important set of questions concern how the networks actually support the generation of such a representation. As a preliminary effort, we have examined the connectivity patterns of the trained network, and they do not seem to resemble the connectivity patterns required by standard attractor network models. Maybe this should not be seen as too surprising. After all, the trained networks can produce a diverse set of neural responses, while the previous models only led to grid responses. It would be interesting for future work to systematically examine the questions related to the underlying mechanisms.

## 5 Acknowledgements

We thank members of the Center for Theoretical Neuroscience at Columbia University for useful discussions and three anonymous reviewers for constructive feedback. Research supported by NSF NeuroNex Award DBI-1707398 and NIH training grant 5T32NS064929 (CJC).

## References

Dmitriy Aronov, Rhino Nevers, and David W Tank. Mapping of a non-spatial dimension by the hippocampalentorhinal circuit. *Nature*, 2017.

Horace B Barlow. Possible principles underlying the transformation of sensory messages. *Sensory communication*, pp. 217–234, 1961.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Tale L Bjerknes, Edvard I Moser, and May-Britt Moser. Representation of geometric borders in the developing rat. *Neuron*, 82(1):71–78, 2014.

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

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

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

Geoffrey W Diehl, Olivia J Hon, Stefan Leutgeb, and Jill K Leutgeb. Grid and nongrid cells in medial entorhinal cortex represent spatial location and environmental features with complementary coding schemes. *Neuron*, 94(1):83–92, 2017.

Yedidyah Dordek, Daniel Soudry, Ron Meir, and Dori Derdikman. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife*, 5:e10094, 2016.

Ariane S Etienne and Kathryn J Jeffery. Path integration in mammals. *Hippocampus*, 14(2):180–192, 2004.

Marianne Fyhn, Sturla Molden, Menno P Witter, Edvard I Moser, and May-Britt Moser. Spatial representation in the entorhinal cortex. *Science*, 305(5688):1258–1264, 2004.

Marianne Fyhn, Torkel Hafting, Menno P Witter, Edvard I Moser, and May-Britt Moser. Grid cells in mice. *Hippocampus*, 18(12):1230–1238, 2008.

Alex Graves, Abdel-Rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pp. 6645–6649. IEEE, 2013.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

Jordan Guerguiev, Timothy P Lillicrap, and Blake A Richards. Towards deep learning with segregated dendrites. *eLife*, 6, 2017.

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.

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

James R Hinman, Mark P Brandon, Jason R Climer, G William Chapman, and Michael E Hasselmo. Multiple running speed signals in medial entorhinal cortex. *Neuron*, 91(3):666–679, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

Joshua Jacobs, Christoph T Weidemann, Jonathan F Miller, Alec Solway, John F Burke, Xue-Xin Wei, Nanthia Suthana, Michael R Sperling, Ashwini D Sharan, Itzhak Fried, et al. Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature neuroscience*, 16(9):1188–1190, 2013.

Ingmar Kanitscheider and Ila Fiete. Training recurrent networks to generate hypotheses about how the brain solves hard navigation problems. *arXiv preprint arXiv:1609.09059*, 2016.

Tim Christian Kietzmann, Patrick McClure, and Nikolaus Kriegeskorte. Deep neural networks in computational neuroscience. *bioRxiv*, pp. 133504, 2017.

Nathaniel J Killian, Michael J Jutras, and Elizabeth A Buffalo. A map of visual space in the primate entorhinal cortex. *Nature*, 491(7426):761–764, 2012.

Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

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.

Julija Krupic, Neil Burgess, and John O?Keefe. Neural representations of location composed of spatially periodic bands. *Science*, 337(6096):853–857, 2012.

Rosamund F Langston, James A Ainge, Jonathan J Couey, Cathrin B Canto, Tale L Bjerknes, Menno P Witter, Edvard I Moser, and May-Britt Moser. Development of the spatial representation system in the rat. *Science*, 328(5985):1576–1580, 2010.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Colin Lever, Stephen Burton, Ali Jeewajee, John O'Keefe, and Neil Burgess. Boundary vector cells in the subiculum of the hippocampal formation. *The journal of neuroscience*, 29(31):9771–9777, 2009.

Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7, 2016.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013.

James Martens and Ilya Sutskever. Learning recurrent neural networks with hessian-free optimization. pp. 10331040, 2011.

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.

Thomas Miconi. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife*, 6:e20899, 2017.

M-L Mittelstaedt and H Mittelstaedt. Homing by path integration in a mammal. *Naturwissenschaften*, 67(11):566–567, 1980.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31:69–89, 2008.

John O'Keefe. Place units in the hippocampus of the freely moving rat. *Experimental neurology*, 51 (1):78–109, 1976.

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

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

Eli Pollock, Niral Desai, Xue-Xin Wei, and Vijay B Balasubramanian. A mechanism for self-organized error-correction of grid cells by border cells. *CoSyNe abstract*, 2017.

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.

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.

Francesco Savelli, D Yoganarasimha, and James J Knierim. Influence of boundary removal on the spatial representations of the medial entorhinal cortex. *Hippocampus*, 18(12):1270, 2008.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. 2014.

Trygve Solstad, Charlotte N Boccara, Emilio Kropff, May-Britt Moser, and Edvard I Moser. Representation of geometric borders in the entorhinal cortex. *Science*, 322(5909):1865–1868, 2008.

H Francis Song, Guangyu R Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLoS Comput Biol*, 12(2): e1004792, 2016.

Kimberly Lauren Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. The hippocampus as a predictive map. *bioRxiv*, pp. 097170, 2016.

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.

David Sussillo, Mark M Churchland, Matthew T Kaufman, and Krishna V Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, 18(7): 1025–1033, 2015.

Lucas Theis and Matthias Bethge. Generative image modeling using spatial lstms. In *Advances in Neural Information Processing Systems*, pp. 1927–1935, 2015.

Xue-Xin Wei, Jason Prentice, and Vijay Balasubramanian. A principle of economy predicts the functional architecture of grid cells. *Elife*, 4:e08362, 2015.

B Willmore and DJ Tolhurst. Characterizing the sparseness of neural codes. *Network*, 12:255–270, 2001.

Tom J Wills, Francesca Cacucci, Neil Burgess, and John O'keefe. Development of the hippocampal cognitive map in preweanling rats. *Science*, 328(5985):1573–1576, 2010.

Shawn S. Winter, Benjamin J. Clark, and Jeffrey S. Taube. Disruption of the head direction cell network impairs the parahippocampal grid cell signal. *Science*, 347(6224):870–874, 2015a.

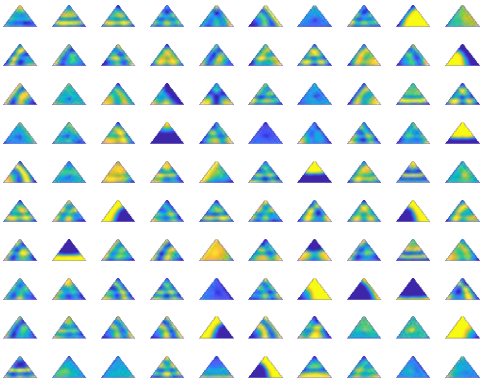
Shawn S. Winter, Max L. Mehlman, Benjamin J. Clark, and Jeffrey S. Taube. Passive transport disrupts grid signals in the parahippocampal cortex. *Current Biology*, 25:2493–2502, 2015b.

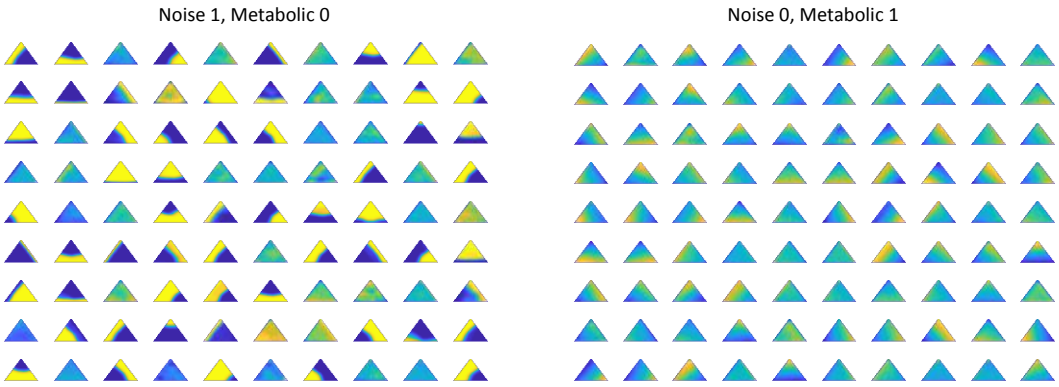
Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

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.

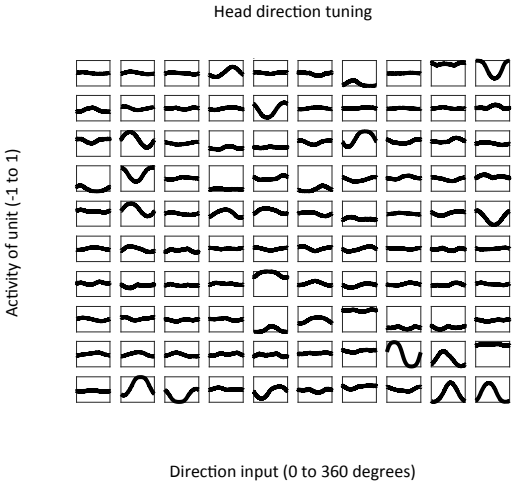
Michael M Yartsev, Menno P Witter, and Nachum Ulanovsky. Grid cells without theta oscillations in the entorhinal cortex of bats. *Nature*, 479(7371):103–107, 2011.

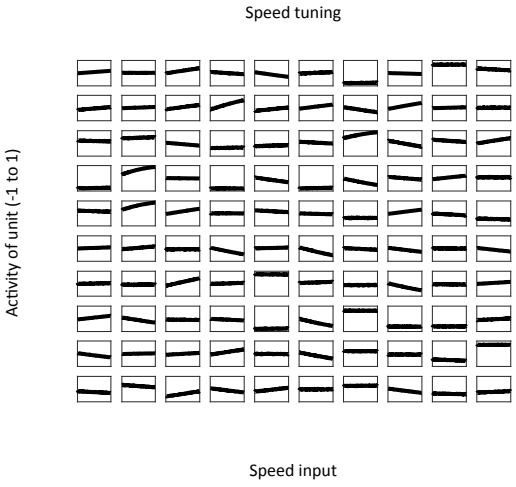
## A TRIANGULAR ENVIRONMENT



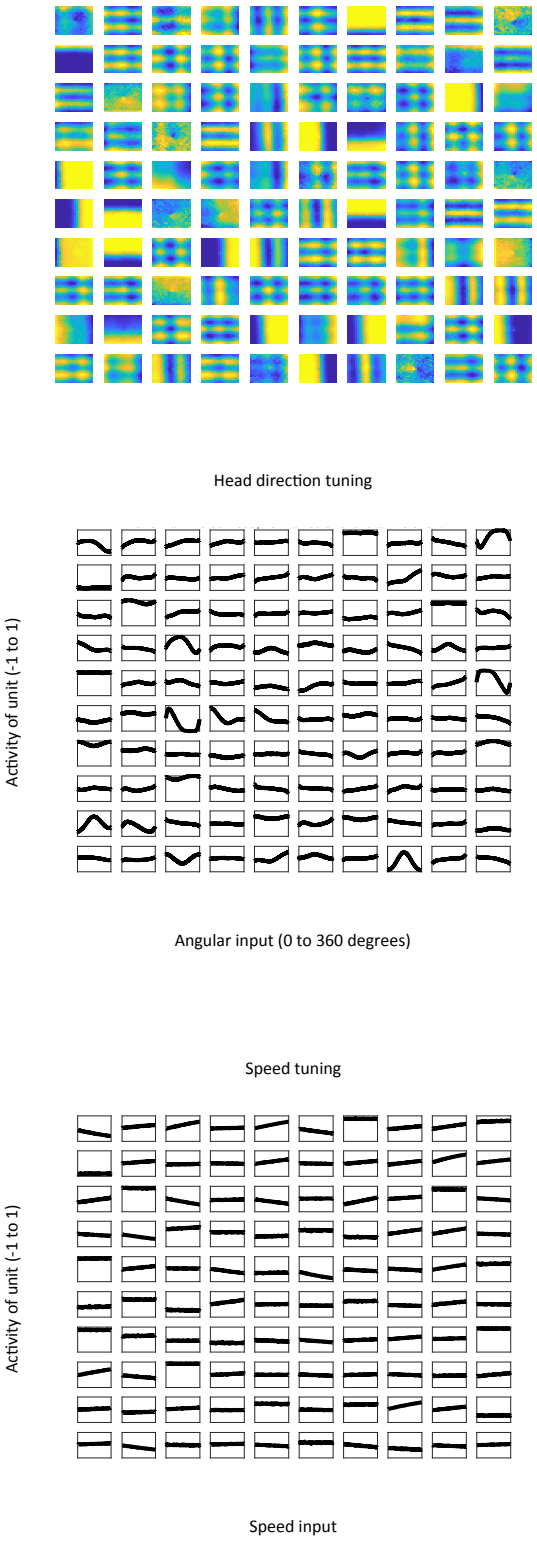



Noise and metabolic cost are important for grid-like representations. The figure on the left shows the spatial responses for a network trained with noise and no metabolic cost. The figure on the right shows the spatial responses for a network trained with no noise and the metabolic cost.

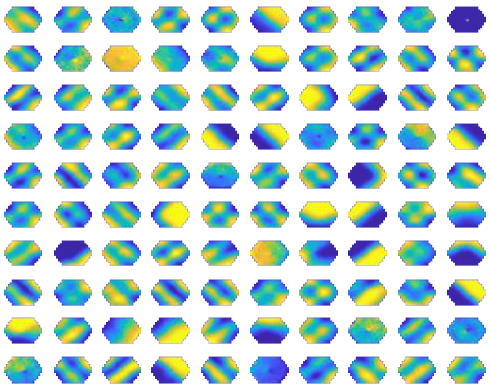

Head direction tuning
Activity of unit (-1 to 1)
Direction input (0 to 360 degrees)


Speed tuning
Activity of unit (-1 to 1)
Speed input

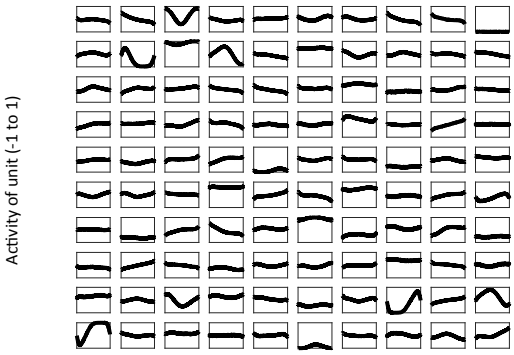
## B RECTANGULAR ENVIRONMENT



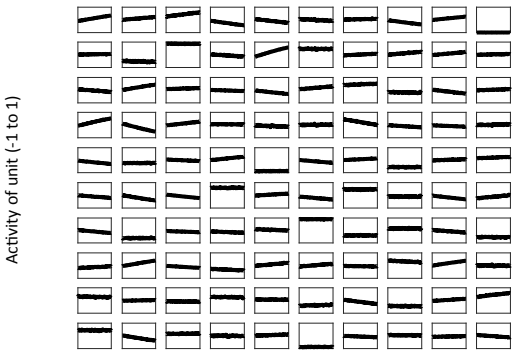
## C HEXAGONAL ENVIRONMENT



Head direction tuning

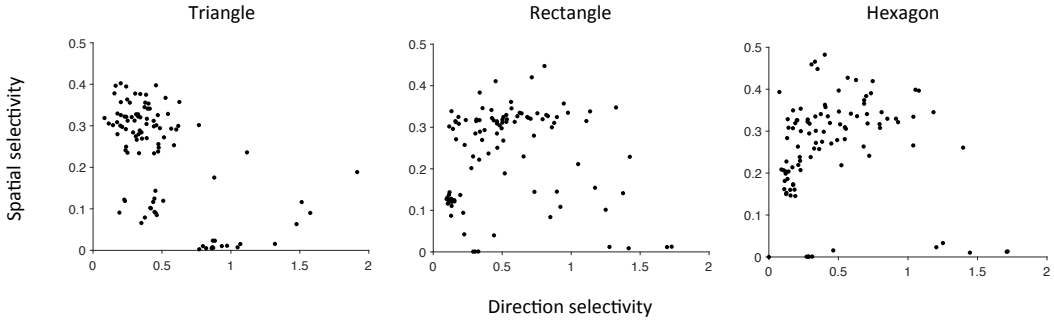


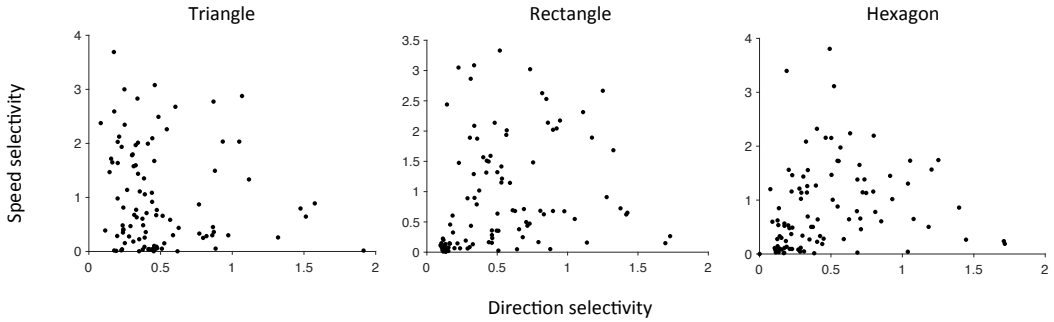

Speed tuning

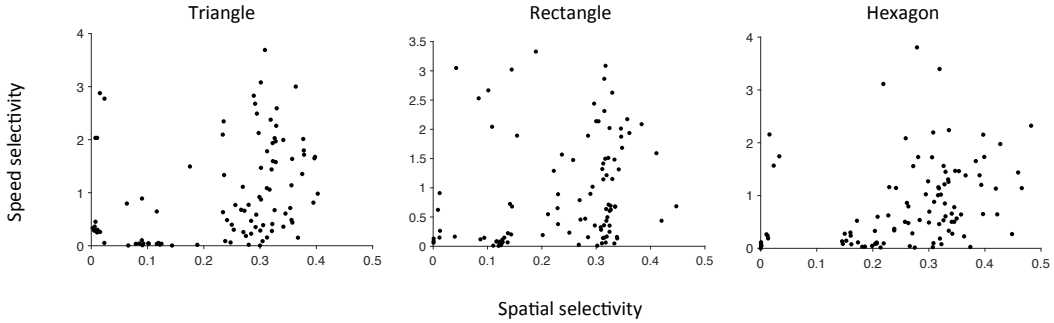


## D RELATION BETWEEN SPEED, DIRECTION, AND SPATIAL SELECTIVITY

To quantify the speed selectivity of each unit we first fit a line to the tuning curve of unit activity as a function of speed. The speed selectivity is the absolute value of the slope. If the unit activity is not modulated by speed then the speed selectivity is 0. To quantify the direction selectivity of each unit we calculated the average unit activity as a function of direction input and then took the maximum minus minimum of this tuning curve. If the unit activity is not modulated by direction then the direction selectivity is 0. To quantify the spatial selectivity we used lifetime sparseness (Willmore & Tolhurst, 2001). If the unit activity is not modulated by spatial location then the spatial selectivity is 0. Each dot in the figures below show the selectivity for a single unit.



## E ADDITIONAL TRAINING DETAILS

During training we tried to balance all three terms we were minimizing ($E$, $R_{L2}$, and $R_{FR}$) so no single term was neglected or dominated. At the beginning of training we weighted the regularization term $R_{L2}$ to be equal to the error function $E$ and then decreased the weighting on $R_{L2}$ according to the schedule used by Martens & Sutskever (2011). We adaptively adjusted the weighting on $R_{FR}$, starting from an initial value of $E/10$ and enforcing an upper bound of $E/3$ as training progressed. We found this training procedure improved training performance and led to more interesting representations.