# A computational model for grid maps in neural populations

Fabio Anselmi[1,2,3] · Micah M. Murray[4,5,6,7] · Benedetta Franceschiello[4,5]

**Abstract**

Grid cells in the entorhinal cortex, together with head direction, place, speed and border cells, are major contributors to the organization of spatial representations in the brain. In this work we introduce a novel theoretical and algorithmic framework able to explain the optimality of hexagonal grid-like response patterns. We show that this pattern is a result of minimal variance encoding of neurons together with maximal robustness to neurons' noise and minimal number of encoding neurons. The novelty lies in the formulation of the encoding problem considering neurons as an overcomplete basis (a frame) where the position information is encoded. Through the modern Frame Theory language, specifically that of tight and equiangular frames, we provide new insights about the optimality of hexagonal grid receptive fields. The proposed model is based on the well-accepted and tested hypothesis of Hebbian learning, providing a simplified cortical-based framework that does not require the presence of velocity-driven oscillations (oscillatory model) or translational symmetries in the synaptic connections (attractor model). We moreover demonstrate that the proposed encoding mechanism naturally explains axis alignment of neighbor grid cells and maps shifts, rotations and scaling of the stimuli onto the shape of grid cells' receptive fields, giving a straightforward explanation of the experimental evidence of grid cells remapping under transformations of environmental cues.

**Keywords** Hippocampus · Grid cells · Computational model

## 1 Introduction

Grid cells in the entorhinal cortex efficiently represent an animal's spatial position using a hexagonal symmetric code (Hafting et al. 2005; Burak and Fiete 2009). Mathematical models have been developed to explain the emergence of such surprisingly regular firing activity (McNaughton et al. 2006; Fuhs and Touretzky 2006; Burgess et al. 2007; Hasselmo et al. 2007; Blair et al. 2007; Kropff and Treves 2008). However, the problem is far from being solved, and many questions remain open (Renart et al. 2003; Yartsev et al. 2011; Schmidt-Hieber and Häusser 2013; Heys et al. 2013). From the modelling point of view, two main mechanisms have been proposed to generate the hexagonal periodic activity: oscillatory interference (Burgess et al. 2007; Orchard et al. 2013) and continuous attractor dynamics (McNaughton et al. 2006; Fuhs and Touretzky 2006). First, we address briefly the main ideas underlying these models.

In oscillatory models, grid cells' patterns emerge from the interference between oscillations of velocity-modulated cells (Burgess et al. 2007; Hasselmo et al. 2007). Experimental results in Krupic et al. (2012) have identified a class of cells, named band cells, that fire at specific spatial periodicity; the interference of three cells of this kind, whose wave vectors' orientations differ by multiples of 120 degrees, leads to hexagonal grid-type interference patterns.

The core idea of continuous attractor models explains the regularity of the grid firing activity as an attractor state generated by symmetrical recurrent interactions between grid cells (McNaughton et al. 2006; Fuhs and Touretzky 2006). A major weakness of this class of models is that it requires an unrealistically high degree of translational symmetry in the strength of the connections among neurons: neurons at equal distance should connect with equal strength. However, real neuronal populations are affected by noise and randomness and therefore break this symmetry and the grid regularity (Renart et al. 2003). Alternative models based on single-cell firing, adaptation, slowly varying spatial inputs, or, more recently, on deep reinforcement learning have been proposed in Kropff and Treves (2008),

✉ Fabio Anselmi
anselmi@mit.edu

Extended author information available on the last page of the article.

Franzius et al. (2007), Banino et al. (2018), and Botvinick et al. (2017).

The model we propose has a number of advantages with respect to those mentioned above. For clarity we list the novel contributions of our work:

- The model is based on the well-accepted and tested hypothesis of Hebbian learning, (Hebb 1949), is much simpler than interference and attractor models, and it does not require the presence of velocity driven oscillations or translational symmetries in the synaptic connections.
- We explain the experimental phenomenon of the alignment of the axes of neighboring grid cells.
- We show how shifted, rotated and scaled grid cells' receptive fields naturally remap, given transformed visual landmarks (Sargolini et al. 2006).
- We sketch a theoretical framework for the otherwise puzzling experimental findings in Constantinescu et al. (2016) where the authors show how grid cells may play a role in the organization of "conceptual" spaces.

## 2 Results

### 2.1 Model description and predictions

The model is based on three assumptions. By analogy with the Hubel and Wiesel simple-complex cells computation in the primary visual cortex (Hubel and Wiesel 1965; 1968), we propose grid cells to emerge from a linear sum of "simple cells" whose receptive fields (RFs) are learned from a collection of neuronal inputs with *stationary* second-order stimulus statistics (**H**1). In other words, we assume that the encoding of the objects' movements at the level of the entorhinal cortex obeys a statistic that does not differ significantly from that of natural images (which is approximately stationary (Field 1999)). Indeed, deep connections between visual recognition tasks and entorhinal cortex has been suggested in Bicanski and Burgess (2019). We also assume that each neuron computes a response that is the scalar product between the input and its synaptic weights i.e.

$$r_i(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w}_i \rangle \tag{1}$$

with $\mathbf{x}$ the input image function and $\mathbf{w}_i$ the synaptic weights function of neuron $i$. In the following, we will fix $\mathbf{x}$ and omit to write the dependence on $\mathbf{x}$.

Furthermore, we assume that the synaptic weights are updated following Oja's rule, derived as a the first order expansion of a normalized Hebbian rule, (Oja 1992), (**H**2). The normalization assumption is plausible, because normalization mechanisms are widespread in the brain (Carandini and Heeger 2011). The original paper of Oja (1982) showed that the weights of a neuron updated according to this rule will converge to the top principal component (PC) of the neuron's past input, i.e. to an eigenfunction of the input's covariance. Plausible modifications of the rule, involving added noise or inhibitory connections with similar neurons, yield additional eigenfunctions (Oja 1992). Thus, this generalized Oja's rule can be regarded as an online algorithm to compute the principal components of incoming streams of input; in our case, the stationary neuronal responses of simple cells. Our last and most important assumption is that the neural population's goal is to encode a variation of its input, in this case the position, with maximal precision (Deneve et al. 1999). Neuronal responses are noisy, and thus repeated, equal stimuli can produce different outputs. The hypothesis tells us that the population coding aims to minimize the variance of the responses. We further assume that the neuronal encoding of the self-position is achieved with minimal number of neurons (**H**3).

The first important consequence of hypotheses (**H**1, **H**2) is that the synaptic weights of the neuronal population are tuned to Fourier functions i.e.

$$\mathbf{w}(\mathbf{k}, \xi) = e^{I\mathbf{k}^T\xi}, \; \mathbf{k}, \xi \in \mathbb{R}^2 \tag{2}$$

where $I$ is the imaginary number, $\mathbf{k}$ is the two-dimensional frequency vector and $\xi$ is the vector of the spatial Cartesian coordinates. This follows from the stationarity of neuronal inputs i.e. the fact that the associated covariance matrix is diagonalized by Fourier functions. A consequence of Oja's rule is that those are also the learned neuronal weights. The relative change of position of the objects in the scene (due to the animal navigating in the environment) is modelled in a first approximation as covariant translations at the level of the highly processed input of the enthorinal neurons i.e. :

$$T_{\mathbf{y}}\mathbf{x}(\xi) = \mathbf{x}(\xi - \mathbf{y}), \mathbf{y} \in \mathbb{R}^2; \tag{3}$$

where $T_{\mathbf{y}}$ is the translation operator. The response of a $N$ neurons population encoding the position of stimulus $\mathbf{y}$ will be:

$$\mathbf{r}(\mathbf{y}) = (r_1(\mathbf{y}), \cdots, r_N(\mathbf{y})) \tag{4}$$

with $r_i(\mathbf{y}) = \left\langle T_{\mathbf{y}}\mathbf{x}, e^{I\mathbf{k}_i^T\xi} \right\rangle$. Upon a change in the observer position the simple cells responses change as:

$$r_i(\mathbf{y}) = \left\langle T_{\mathbf{y}}\mathbf{x}, e^{I\mathbf{k}_i^T\xi} \right\rangle = e^{I\mathbf{k}_i^T\mathbf{y}} c_i(\mathbf{x})$$

where $c_i(\mathbf{x})$ are the Fourier coefficients of $\mathbf{x}$ with respect to the vectors of frequencies $\mathbf{k}_i$. Thus, the position information is encoded in the phase factors $e^{I\mathbf{k}_i^T\mathbf{y}}$, due to the translation covariance of the Fourier transform. Simple cells could be identified with band cells in Krupic et al. (2012), although their origin in our work is of a completely different nature.

In the following, we will focus on the emergence of grid cells' receptive fields, in particular on how they can be derived from optimality of the position information contained in the phase factors of the simple cells' responses.

The simplest model of a *"complex" grid cell* aggregates the responses of simple cells by summation:

$$r(\mathbf{y}) = \sum_{i=1}^{N} r_i(\mathbf{y}) = \left\langle T_{\mathbf{y}}\mathbf{x}, \sum_{i=1}^{N} \mathbf{w}_i \right\rangle = \sum_{i=1}^{N} c_i(\mathbf{x}) e^{I\mathbf{k}_i^T \mathbf{y}} \qquad (5)$$

The phases, as in Orchard et al. (2013), encode the information about the observer position.

In general, each single simple cell's response can be considered as a random variable subject to noise. In this work, we analyze the case of constant pairwise noise i.e. the noise covariance matrix has the form

$$C = \sigma_1^2 \mathbf{I} + \sigma_2^2 \mathbf{1}\mathbf{1}^T, \quad \sigma_{1,2} > 0 \qquad (6)$$

with $\mathbf{1}$ the vector of all ones.

Assuming (**H**2) and (**H**3), the question that follows is: which set of frequencies $\{\mathbf{k}_i\}$ are best to encode the animal's position $\mathbf{y}$ with maximal precision, given the noise and a fixed, $N$, number of encoding neurons?

A lower bound on any possible unbiased estimator of the random variable $\mathbf{y}$ is given by the Cramer-Rao bound ((Kay 1993), see Materials and Methods 4.1 for more details). The bound reads:

$$\|Cov(\mathbf{y})\| \geq \left\| \mathbf{F}^{-1}(\mathbf{y}) \right\| \qquad (7)$$

where $\mathbf{F}$ is the so-called *Fisher information* matrix, $Cov$ is the covariance matrix of the neuronal responses, and $\|\cdot\|$ is a matrix norm. Intuitively, $\mathbf{F}$ measures the amount of information that the encoding population carries about the random variable $\mathbf{y}$. The main theoretical result of the paper is that, in dimension two, the lower bound for the right hand side of Eq. (7) is achieved, for any fixed $N$, when the frequency vectors are a *tight frame*. These are frames that maximize the angle among each pair of frame elements and have been proven to have various beneficial properties in signal processing, including robustness to noise. In particular, we prove that if we require a minimal number of neurons, the only frame which is robust to neuronal pairwise constant noise and which has minimal associated covariance is the so-called Mercedes-Benz frame, composed by vectors, $\{\mathbf{k}_i\}$, whose orientations differ by 120° degrees (see Fig. 1). The set of frequency vectors form a so-called *Equiangular frame*:

$$\mathbf{k}_p^T \mathbf{k}_q = \cos(\alpha), \ \forall\, p, q; \ \alpha \in [0, 2\pi], \alpha \text{ constant.} \qquad (8)$$

It is quite biologically implausible that grid cells receive input from so few simple cells. Thus, we consider the activity of one simple cell unit in our model as summarizing that of a whole population of cells with the same preferential orientation $\mathbf{k_i}$. Summing over multiple simple cells' responses sensitive to the same orientation maintains the value of Eq. (5) unchanged up to an overall constant factor.

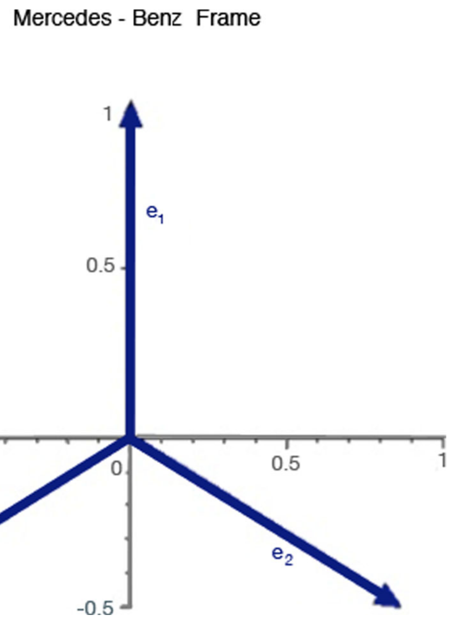

**Fig. 1** Mercedes-Benz frame in dimension two. Note how the vectors are separated by 120° one from the other

More formally we can prove (see Materials and methods):

**Theorem 1** *Given the hypotheses $H(1, 2, 3)$, the minimum variance position encoded by a set of $N$ neurons is achieved when the frequency vectors form a tight frame. If we further require the encoding to be done with minimum number of neurons and to be maximally robust to constant pairwise correlated noise, we have a unique solution for the set of frequency vectors:*

$$f = \{\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3\} = \{(\cos(2\pi j/3), \sin(2\pi j/3)), j = 1, 2, 3\}$$

*the Mercedes Benz frame.*

Our result states that the best encoding of position robust to noise, with minimal number of neurons, is achieved when the "complex" grid-like cell is aggregating the responses of three neurons (or similarly-tuned neural populations) whose neuronal weights are Fourier functions with *equiangular frequencies* in the frequency space. Suppose now the neurons' weights have been learned. The response in Eq. (5) produces output in terms of an interference pattern of three planar waves that is consistent with a hexagonal grid (see Fig. 2). Before proceeding, we discuss the novelty of our contribution with regard to the existing literature in the following remarks.

– Our model may resemble the interference model of Burgess et al. (2007). However, it is important to point

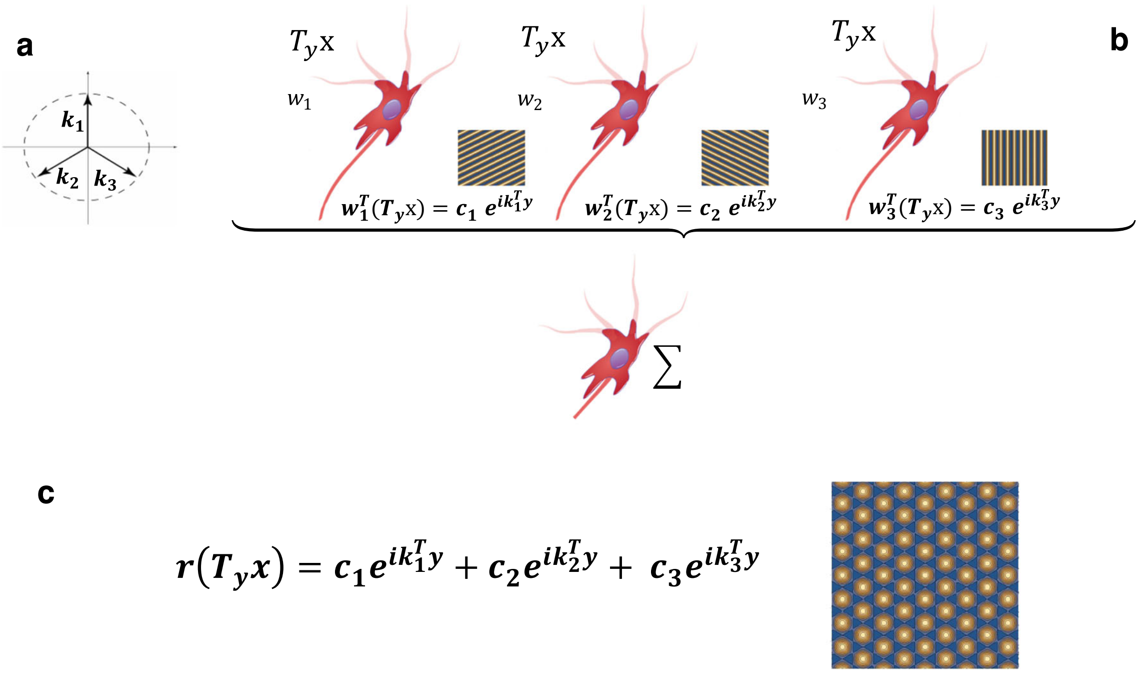
a
$k_1$
$k_2$
$k_3$
b
$T_y x$
$w_1$
$w_2$
$w_3$
$w_1^T(T_y x) = c_1\, e^{ik_1^T y}$
$w_2^T(T_y x) = c_2\, e^{ik_2^T y}$
$w_3^T(T_y x) = c_3\, e^{ik_3^T y}$
$\Sigma$
c
$r(T_y x) = c_1 e^{ik_1^T y} + c_2 e^{ik_2^T y} + c_3 e^{ik_3^T y}$

**Fig. 2** The image shows how a grid-like cell's pattern arises from the interference of planar waves responses. **a** The Mercedes-Benz frame is constituted by the equiangular vectors $k_1, k_2, k_3$, whose directions are along the angles $\theta = \pi/2, -\pi/6, -5\pi/6$. **b** Three neurons with input stimulus $x$, translated by a vector $y$, $T_y x$ and their receptive fields $w_1, w_2, w_3$ i.e. three planar waves in equiangular directions. **c** Linear sum of the three neurons' responses (indicated in B as $\Sigma$) resulting in the grid-like hexagonal pattern

out one crucial difference. The oscillations interference pattern in our model is due to the shape of the learned receptive fields of the simple cells and not to the oscillations in the hippocampal circuit. Therefore, our simple cells can be identified with band cells in Krupic et al. (2012), although their emergence is explained in our work without the need of those oscillations.

- In the same vein, although PCA is used in our model to derive the shape of the simple cells' receptive fields, its role is completely different from the one described in Dordek et al. (2016) or Castro and Aguiar (2014) or Stachenfeld et al. (2017), where PCA is used to derive the shape of the grid cells' responses from place cells.
- Optimality of grid cells' hexagonal receptive field is here derived in a novel way with regard to that in Fiete et al. (2008), Mathis et al. (2012), Vágó and Ujfalussy (2018), and Domínguez and Caplan (2018). In fact, we use concepts and properties from Frame theory; in particular those of the so-called Mercedes-Benz frame (see e.g. Kovacevic and Chebira (2007)).
- Strip cells in Mhatre et al. (2012) can be identified with the phase of our simple cells that give information about the animal's movement in a particular direction. However, the principle we use to derive the grid cells' receptive fields is different. We minimize the position response covariance using the Cramer Rao bound and Fisher Information. The authors in Mhatre et al. (2012) give a geometrical reasoning for the maximization of the stripe cells response. In our derivation of the Fisher information, the phase factors cancel (see Materials and Methods) and only the frequency vectors' directions play a role in the minimization.

## 2.2 Two-phase application of Oja's learning rule

In this section we explain how hexagonal grid receptive fields emerge from a two-phase learning process.

### 2.2.1 First phase:simple cells learning

A collection of cyclical translations in the cartesian directions of natural images is used as input stimuli to compute the "simple" cell profile of activation. Next, the principal components of these activation profiles are extracted, diagonalizing the covariance matrix of the input data. The second order statistic of the input, i.e. the covariance matrix, is clearly stationary. Note that the stationarity is independent from the nature of the stimulus and it crucially depends on the more abstract notion of transformation (in this case translations). Under the assumption of Oja's rule, this mechanism simulates the learning phase of a simple cell.

An example of the learned receptive fields is shown in Fig. 3a. As expected, they are Fourier components.

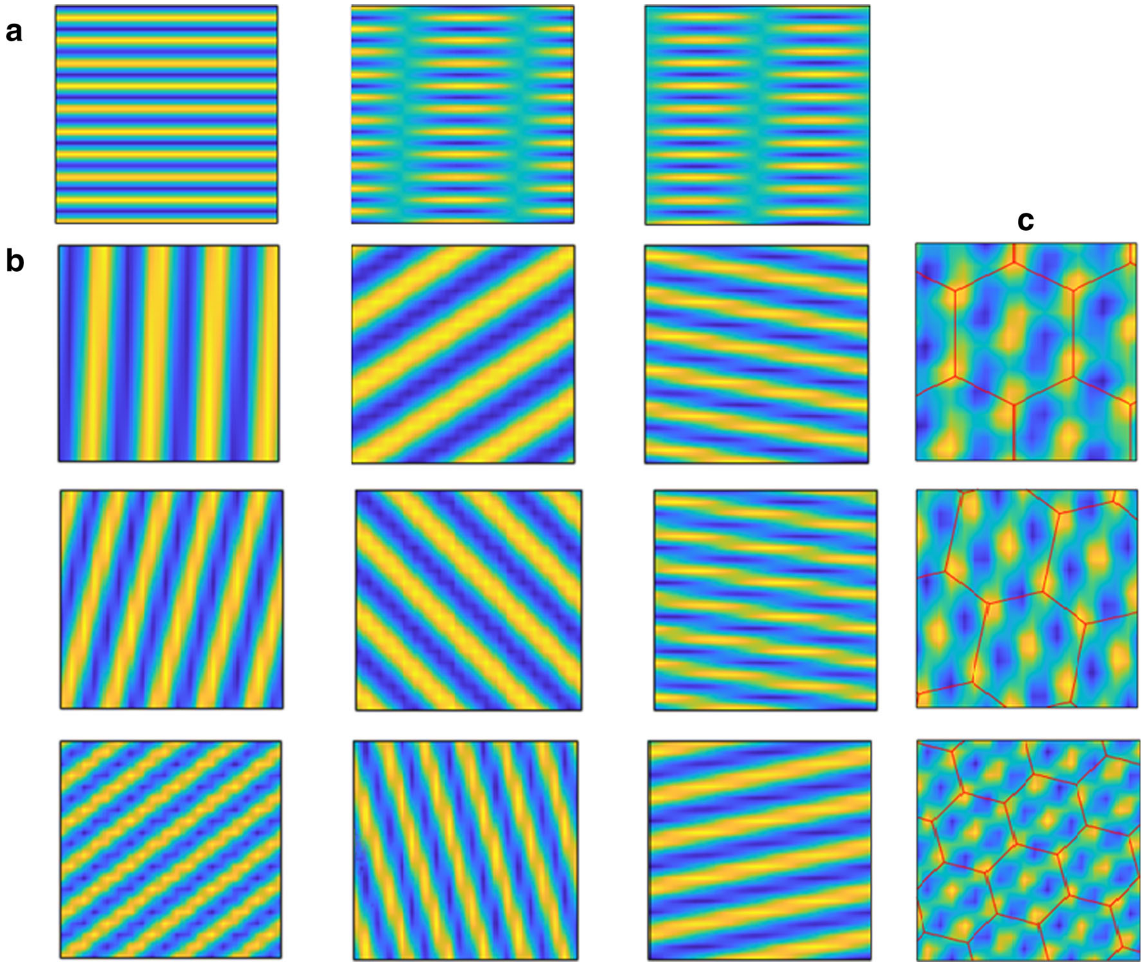

**Fig. 3** **a** Example of simple cells' receptive fields, obtained through the first phase learning, from stationary stimuli. **b** Simple cells' receptive fields, selected through variance minimization of the estimated position together with minimum number of neurons constraint. **c** Superposition of equiangular patterns selected from figure (b)

### 2.2.2 Second phase: "complex" grid cell learning.

The second step entails aggregating the responses of simple cells. A collection of cyclical translations of a test image is used to calculate the aggregation vector $J$ according to the minimization problem (see Materials and Methods, *Algorithmic formulation* for the algorithm details). The results displayed in Fig. 3 show that, as a result of variance minimization of the position estimate, waves with 120° angular distance are selected (b). Their superposition have a grid-like shape (c). Although the algorithm provides, approximately, the angular directions predicted by the theorem, not all superpositions produce grid-like patterns. This is due to the different frequencies of the Fourier receptive fields. For a fixed frequency, the selected receptive fields sum to produce a grid-like interference pattern.

The mechanism underpinning the combination of receptive fields of the same frequency relies on the nature of principal component decomposition: the first eigen-component is an oscillating wave whose frequency depends on the strongest oscillating component in the stimuli (see Aapo et al. (2009), pg 120). Thus, since neighborhood cells have approximately the same dendritic extension, the RF will be tuned, to similar frequencies. Moreover, cells in the same spatial neighbourhood, receiving the same input, will be tuned to similar wave orientations vectors. This explains a salient aspect of grid cells phenomenological behavior: neighboring grids cells have aligned orientations of their axes (i.e. the same orientations of the hexagonal axes).

### 2.2.3 Grid remapping by changing environmental cues

Experimental evidence shows that changes in environmental cues are matched by a transformation in the animal's grid cell responses (Sargolini et al. 2006). For example, a rotation of the main visual cues in the environment results in a rotation of the grid cells' orientation fields.

This aspect can be readily explained by our model. For example, if an environmental cue is rotated by an angle $\theta$ the grid rotates accordingly, since

$$r_i(\mathbf{R}_\theta \mathbf{x}) = \langle \mathbf{R}_\theta \mathbf{x}, \mathbf{w}_i \rangle = \langle \mathbf{x}, \mathbf{R}_{-\theta} \mathbf{w}_i \rangle$$

where $\mathbf{x}$ is the input of the simple cells. In this case, the frequency vectors $\{\mathbf{k}_i\}$ will be all rotated by the opposite angle $R_{-\theta}\mathbf{k}_i$, with a resulting rotation of the hexagonal grid. Similarly, for a scale transformation, the frequency vectors will be rescaled by the scaling factor.

## 3 Discussion

We successfully showed how hexagonal receptive fields, resembling those of grid cells, emerge naturally in the spatial encoding framework by requiring minimal variance (maximal precision) of the population encoding together with Oja's learning rule and minimal number of neurons involved in the encoding.

The assumption of a 2-phase simple-complex cell type learning adapts properties typically found in the early visual cortex (V1) to those characterizing the entorhinal cortex. We contend that similarity in the types of learning is plausible, given that the entorhinal cortex integrates visual information while also determining the relative position of the observer navigating the spatial environment.

Importantly, the presented model provides a theoretical framework capable of explaining the experimental evidence that grid cells encode an abstract notion of space, decoupled from the specificity of the sensory inputs. The notion indeed emerges from the mathematical group properties of the objects' transformations, rather than the objects themselves. More generally, our model would indicate that grid-like coding should manifest whenever the statistics of the neuronal inputs is stationary. Indeed, the model detailed here provides a mathematical framework able to mimic the emergence of grid-like patterns not only in a spatial encoding scheme (where the considered transformations of the space are translations), but also in a more conceptual encoding scheme (where the transformations are dilations, e.g. Constantinescu et al. (2016)).

### 3.1 "Conceptual" encoding schemes

The idea that grid-like cells could provide a model to understand "cognitive", in addition to sensory-related, brain functions is not new (Moser and Moser 2013; Moser et al. 2014). However, it was not before the work in Constantinescu et al. (2016) that the first experimental evidence was provided. Interestingly, our findings can be applied to outline a theoretical framework for investigating a possible computational model of their experimental evidence.

The stimuli in Constantinescu et al. (2016) are described in a two-dimensional *conceptual bird space*, where the position coordinates are the lengths of both the neck and legs of the bird. In Constantinescu et al. (2016) the authors show an hexagonal grid-like pattern, while testing conceptual associations with functional Magnetic Resonance Imaging (fMRI). For simplicity, we model the input space by using the shear group in $2D$ (composed of transformations dilating an image in the $x$, $y$ directions). Instead of the ratio between the legs of the birds and their necks we can think about the ratio between the base and height of a rectangle that scale in the directions $x$ and $y$, respectively, according to the parameters $(l_1, l_2) = \mathbf{l} \in \mathbb{R}^2$ (see Fig. 4c).

The transformation corresponds to Eq. (3) where instead of the translation operator the shear operator was used:

$$D_{\mathbf{l}}\mathbf{x}(\xi) = \mathbf{x}\left(\frac{\xi}{\mathbf{l}}\right), \tag{9}$$

where $D_{\mathbf{l}}$ indicates the shear operator. The main idea is to apply our spatial encoding to assess whether the model allows to represent the grid-like conceptual patterns observed in Constantinescu et al. (2016). We stress that, similarly to the bird space (where the direction of motion in the abstract "bird space" was determined by the ratio between the neck and the leg lengths), the direction of motion (in our abstract "scale space") is determined by the ratio between the base and height of the rectangle. It is simple to demonstrate that also in this case the second order statistic of the input is stationary, since it again depends only on the nature of the transformation. Therefore, the synaptic weights of the neuronal population are tuned to the eigenfunctions of the shear operator as they were before in the translation case. These eigenfunctions are a generalization of Fourier components to the shear group and have the same form, but in the log-scale coordinates $\log(\mathbf{l}) = (log(l_1), log(l_2))$, where $l_1, l_2$ are the scaling factors in the $x$ and $y$ directions (see e.g. Eagleson (1992)) :

$$s(\mathbf{l}, \mathbf{k}) = e^{I\mathbf{k}^T \log(\mathbf{l})}. \tag{10}$$

The key observation is that in this new coordinate frame (provided by the response of the "simple cells"), the shear transformations reduce to translations in the $\log(\mathbf{l})$-space since:

$$\log(\mathbf{l}\mathbf{l}') = \log(\mathbf{l}) + \log(\mathbf{l}'). \tag{11}$$

In this space the eigenfunctions are planar waves as in Fig. 2 A, applying Theorem 1. We can then prove that also in this case the set of frequency vectors $\{\mathbf{k}_i\}$ is the Mercedes-Benz frame. This will produce a square or hexagonal grid in the *shear space*, where the coordinates, instead of the spatial ones, are the scale coordinates $\mathbf{l} = (l_1, l_2)$ as in Fig. 4, (E).

A couple of remarks are in order. The results in Theorem 1 can be generalized to any Abelian group; the

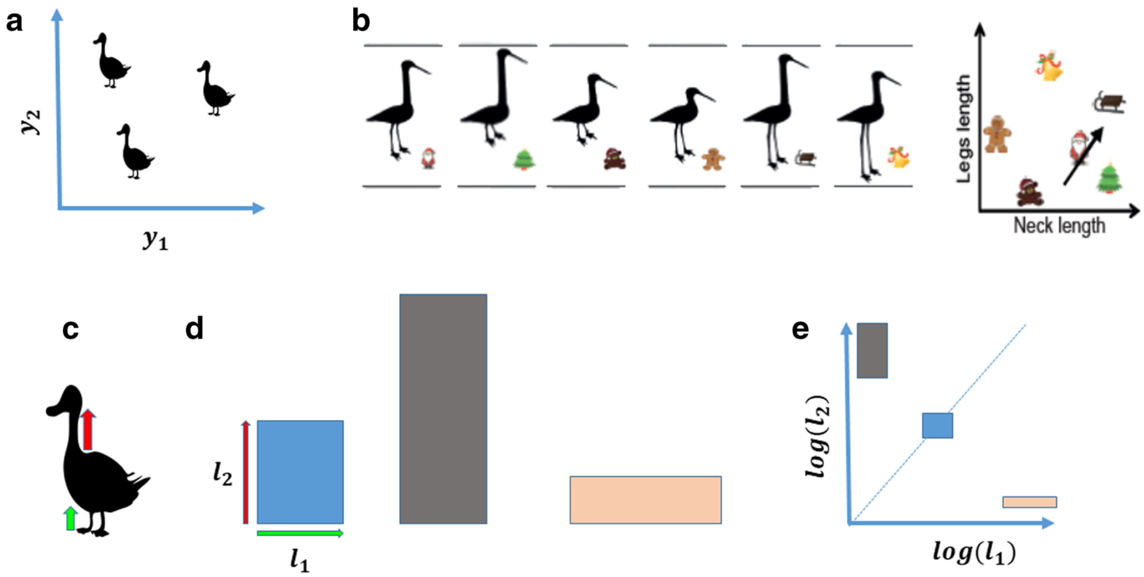

**Fig. 4** Translations of a bird in space (**a**). Transformations of a bird (**b**) and associated points in the 'bird space"(from Constantinescu et al. (2016)). Transformations of a rectangle (**d**) that simplify the bird transformations in (**c**). The associated points in the "rectangle space" (**e**)

eigenfunctions of the group transformations are the group characters (Eagleson 1992). It can be used to predict grid-like cell geometries in higher dimensions.

In dimension three, a possible solution corresponds to the vectors associated to the vertices of a tetrahedron. More generally, in dimension $d$ we will end up with the vectors associated to the vertices of a Platonic Solid. However, it should be noted that many other solution configurations might exist that are distinct from the case of $d = 2$ analyzed in this paper.

### 3.2 Conclusions and outlook

We detailed a computational model able to account for the emergence of hexagonal grid-like response patterns that derives from simple cells' responses and neural sensitivity to the statistics of the input stimuli (i.e. minimal variance encoding). Using results from Frame Theory, we provided a novel formulation of the encoding problem within the framework of tight and equiangular frames. We were able to demonstrate that grid-like receptive field patterns persist despite transformations of the environmental cues as well as when more "conceptual" features are considered as input stimuli. Further work will be required to extend our findings to reproduce the experimental evidence showing that the regular pattern of the grid receptive field adapts to different geometries of the environment, distorting its hexagonal regularity (Krupic et al. 2014; Urdapilleta et al. 2015; Keinath et al. 2018). Our main result in Theorem 1 predicts the same hexagonal grid for the 3D space of rotations of an object, leading to a series of experiments in the same spirit of Cheng (2018), Kim (2019), and Jacobs (2013) (for spatial encoding) and of Constantinescu et al. (2016) for conceptual encoding possibly tested using electroencephalography or magnetoencephalography (Staudigl et al. 2018).

## 4 Materials and methods

### 4.1 Fisher information

The Cramer-Rao bound (CRB) sets a lower bound on the norm of the covariance operator of any random variable $\mathbf{y}$ unbiased estimator. It says:

$$\|Cov(\mathbf{y})\| \leq \left\| F^{-1}(\mathbf{y}) \right\| \tag{12}$$

where $F$ is the Fisher information defined as:

$$(F(\mathbf{y}))_{k,l} = -\mathbb{E}\left(\frac{\partial^2 \log L(\mathbf{y})}{\partial y_k \partial y_l}\right) \tag{13}$$

and $L(\mathbf{y})$ is the likelihood function and the average is over of the measurements of $\mathbf{y}$. In our case the likelihood function is, Yoon and Sompolinsky (1998):

$$L(\mathbf{y}) = \mathcal{N} exp\left(-\frac{1}{2}\sum_{i,j=1}^{N}(y_i - r_i(\mathbf{y}))C_{ij}^{-1}(y_j - r_j(\mathbf{y}))\right) \tag{14}$$

where $\mathcal{N}$ is a normalization constant, $C$ is the correlation matrix of the noise and $r_i$ is the response of the $i^{th}$ neuron (also dependent on $\mathbf{x}$, that we omit for simplicity). Under the hypothesis of uncorrelated gaussian equal noise, i.e. $\mathbf{C}^{-1} = (1/\sigma^2)\mathbf{I}$, a direct calculation of Eq. (14) gives:

$$\mathbf{F}(\mathbf{y}) = -\mathbb{E}\left\{\left(\frac{\partial \mathbf{r}(\mathbf{y})}{\partial \mathbf{y}}\right)^{\dagger} \mathbf{C}^{-1}\frac{\partial \mathbf{r}(\mathbf{y})}{\partial \mathbf{y}}\right\}$$
$$= -\frac{1}{\sigma^2}\mathbb{E}\left\{\left(\frac{\partial \mathcal{F}(T_{\mathbf{y}}\mathbf{x})}{\partial \mathbf{y}}\right)^{\dagger} \frac{\partial \mathcal{F}(T_{\mathbf{y}}\mathbf{x})}{\partial \mathbf{y}}\right\},$$

where $\mathcal{F}$ is the Fourier transform. Starting from the following identity:

$$\frac{\partial \mathcal{F}_i(T_{\mathbf{y}}\mathbf{x})}{\partial \mathbf{y}} = I e^{I\mathbf{k}_i{}^T\mathbf{y}}\mathbf{k}_i c_i(\mathbf{x}), \quad \mathbf{c}(\mathbf{x}) = \mathcal{F}(\mathbf{x}),$$

we have:

$$\mathbf{F} = \frac{1}{\sigma^2}\mathbb{E}\left\{\sum_{i=1}^{N}\mathbf{k}_i\mathbf{k}_i{}^T|c_i(\mathbf{x})|^2\right\}$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}\mathbf{k}_i\mathbf{k}_i{}^T\mathbb{E}(|c_i(\mathbf{x})|^2)$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}\frac{\mathbf{k}_i\mathbf{k}_i{}^T}{\|\mathbf{k}_i\|_2^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N}\mathbf{g}_i\mathbf{g}_i{}^T$$
$$= \frac{1}{\sigma^2}\mathbf{G}\mathbf{G}^T \tag{15}$$

where we used the fact that the averaged power spectrum $\mathbb{E}(|c_i(\mathbf{x})|^2) \approx \|\mathbf{k}_i\|_2^{-2}$, we define the unit norm vector $\mathbf{g}_i = \mathbf{k}_i / \|\mathbf{k}_i\|$ and we defined $\mathbf{G}$ as the matrix with $\mathbf{g}_i$ as columns.

The question we address in the next paragraphs is: for which set of $\mathbf{k}_i$ is the CRB achieved? In other words, we are looking for the values of $\mathbf{k}_i$ for which the neuronal population is providing an estimate of the variable $\mathbf{y}$ with minimal variance. In particular we consider the following minimization problem:

$$\arg\min_{\{\mathbf{k}_i\}_{i=1}^{N}} \left\|\mathbf{F}^{-1}\right\|_{Frob}^2 . \tag{16}$$

Before we add a result that take into account the presence of pairwise constant correlated noise in the encoding.

#### 4.1.1 The case of constant pairwise noise correlation

In the case of constant pairwise noise correlation the covariance matrix $\mathbf{C}$ reads as:

$$\mathbf{C} = \sigma_1^2\mathbf{I} + \sigma_2^2\mathbf{1}\mathbf{1}^T. \tag{17}$$

Its inverse can be written, thanks to the Woodbury's identity as:

$$\mathbf{C}^{-1} = (\sigma_1^2\mathbf{I}+\sigma_2^2\mathbf{1}\mathbf{1}^T)^{-1} = \frac{1}{\sigma_1^2}\left(I - \frac{\sigma_2^2}{\sigma_1^2 + d\sigma_2^2}\mathbf{1}\mathbf{1}^T\right) \tag{18}$$

with $\sigma_{1,2} > 0, \ d = 2$. Repeating the same calculations done in the previous section we have that in the case of pairwise constant correlation noise the Fisher information gains an extra term of the form:

$$\frac{\sigma_2^2}{\sigma_1^2 + 2\sigma_2^2}\mathbf{G}\mathbf{1}\mathbf{1}^T\mathbf{G}^T = const\ \|\mathbf{G}\mathbf{1}\|_2^2 . \tag{19}$$

Note that $\mathbf{G1}$ is the vector whose components are the sum of the frame element coordinates. Interestingly, for balanced frames (those whose sum of the elements is zero) this contribution is null.

### 4.2 Optimal estimator and connection with frame theory

In this section we derive the proof of the main result of the paper.

**Theorem 1** *Under the hypotheses* $\mathbf{H}(1, 2, 3)$ *the minimal variance position encoded by a set of N neurons is achieved when the set of frequency vectors form a tight frame. Further, if we ask for robustness to pairwise constant neuronal noise together with minimal number of neurons the set of frequency vectors is uniquely determined and is:*

$$f = \{\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3\} = \{(cos(2\pi j/3), sin(2\pi j/3)),\ j = 1, 2, 3\}.$$

*the so called Mercedes Benz frame.*

*Proof* Using the fact that $\mathbf{F}$ is semi-positive definite we can decompose it as $\mathbf{F} = \mathbf{V}^T\Lambda\mathbf{V}$ where $\mathbf{V}$ is unitary and $\Lambda$ is diagonal. According to the Cramer-Rao bound, the variance is bounded from below by the inverse of the Fisher Information. Calculating the Frobenius norm of the Fisher matrix inverse, we have:

$$\left\|\mathbf{F}^{-1}\right\|^2 = Tr(\mathbf{V}^T(\Lambda^{-1})^2\mathbf{V}) = Tr((\Lambda^{-1})^2) = \sum_{i=1}^{N}\frac{1}{\lambda_i^2} \tag{20}$$

where $\lambda_i$ are the eigenvalues of $\mathbf{F}$.

It is easy to prove that the minimum of Eq. (20) is reached when all the eigenvalues are equal i.e.

$$\mathbf{F} = \lambda\mathbf{I} \tag{21}$$

i.e. the set $\mathbf{k_i}$ form a tight frame. Considering solutions with minimal number of neurons, i.e. $N = 2, 3$ we have (see Goyal and Kovacevic (2001), pg 210): for $N = 2$, the orthogonal frame, for $N = 3$ the so called Mercedes-Benz frame (or any rotated version of them):

$$\mathbf{k}_j = \left(\cos\left(\frac{2\pi j}{3}\right), \sin\left(\frac{2\pi j}{3}\right)\right) \quad j = 1, \cdots, 3.$$

But the orthogonal frame is not balanced. So the solution with minimal number of neurons is the Mercedes Benz frame. □

### 4.3 Algorithmic formulation

In this article we suppose a 2-phases learning process:

(1) Learning of the Fourier components by simple cells using Oja's synaptic updating rule;
(2) Learning of the selection of simple cells, performed by the complex cell that minimize, according to the Cramer-Rao bound, the norm of the inverse of the Fisher Information.

Solving phase 1 simply corresponds to the extraction of the principal components of neural input. As for phase 2 the minimization problem for the complex cell is:

$$\arg\min_{\{\mathbf{k_i}\}_{i=1}^{N}} \quad \left\| F^{-1} \right\|^2 \quad \text{with } N \quad \text{minimal.}$$

As we saw in Eq. (15), the minimization of $\left\| F^{-1} \right\|^2$ is achieved when the set $\{\mathbf{g}_i\}$ forms a tight frame. Tight frames are minima of the so called *frame potential* (see Casazza et al. (2006)), calculated as:

$$FP(\{\mathbf{g}_i\}) = \sum_{ij=1}^{N} |\mathbf{g}_i^T \mathbf{g}_j|^2.$$

Let us calculate the frame potential in our case. Here, we denote with $W$ the matrix whose columns are the vectors $\mathbf{w}_i$, simple cells receptive fields. Hence the response matrix (i.e. the simple cells output) will be $A = X^T W$ where $X$ is the dataset corresponding to the initial stimuli. The complex grid cells will then aggregate some of the responses, i.e. they will calculate $AJ$ where $J$ is a vector of zeros and ones selecting which simple cells are meant to aggregate (we will have a zero whether the simple cell is not selected in the aggregation process and one elsewhere). We can now use this notation to write the Fisher information as follows:

$$F = -\left(\frac{\partial \mathbf{r}(\mathbf{y})}{\partial \mathbf{y}}\right)^{\dagger} \frac{\partial \mathbf{r}(\mathbf{y})}{\partial \mathbf{y}} = J^T \dot{A}^T \dot{A} J = J^T R J.$$

with the dot indicating the derivative and $R = \dot{A}^T \dot{A}$. In order to minimize the number of simple cells pooled by the complex cell, we add a sparsifying term in $\|J\|_0$ or its relaxation $\|J\|_1$. Given the above reasoning, our minimization problem boils down to:

$$\underset{J}{\operatorname{argmin}} \quad \left\| J^{T7} R J \right\|^2 + \lambda \, \|J\|_1 \, .$$

To find the solution we adopt a gradient descent strategy with shrinkage; a calculation shows that the update rule for $J$ is:

$$J \rightarrow thr(J - \lambda \, RJJ^T RJ)$$

where the $thr$ threshold is enforcing the sparsity constraint.

Simulations showed weak dependence of the λ parameter. On the contrary we observed that the hexagonality of the grid cell RF critically depends on the initialization of the $J$ vector. Further investigation is needed to fully understand this behaviour.

**Acknowledgements** M. M. M. is supported by grants from the Swiss National Science Foundation (grant 320030-169206), the Fondation Asile des aveugles, and a grantor advised by Carigest SA. B. F. is supported by the Fondation Asile des aveugles. F. A. acknowledges the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 and the financial support of the AFOSR projects FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS - DLV-777826. F. A. is also supported by the Italian Institute of Technology.

**Author Contributions** F.A conceptualized the problem. F. A. and B.F. developed, implemented, and tested the model. F.A , B.F and M. M. M. wrote the manuscript.

## References

Aapo, H., Hoyer, P.O., Hurri, J. (2009). Natural image statistics: a probabilistic approach to early computational vision. *Book*.

Banino, A. et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*, 05.

Bicanski, A., & Burgess, N. (2019). A computational model of visual recognition memory via grid cells. *Current Biology*, *29*(6), 979–990, e4.

Blair, H.T., Welday, A.C., Zhang, K. (2007). Scale-invariant memory representations emerge from moire interference between grid fields that produce theta oscillations: a computational model. *Journal of Neuroscience*, *27*, 3211–3229.

Burak, Y., & Fiete, I.R. (2009). Accurate path integration in continuous attractor network models of grid cells. *PLoS Computational Biology*, 5.

Burgess, N., Barry, C., O'Keefe, J. (2007). An oscillatory interference model of grid cell firing. *Hippocampus*, *17*, 801–812.

Carandini, M., & Heeger, D.J. (2011). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62.

Casazza, G.P., Fickus, M., Kovačević, J., Leon, M.T., Tremain, C.J. (2006). A physical interpretation of tight frames. Harmonic analysis and applications. *Applied and Numerical Harmonic Analysis*.

Castro, L., & Aguiar, P. (2014). A feedforward model for the formation of a grid field where spatial information is provided solely from place cells. *Biological Cybernetics*, *108*(2), 133–143.

Cheng, D. (2018). Hexadirectional modulation of theta power in human entorhinal cortex during spatial navigation. *Current Biology*, *28*, 20, 3310–3315.

Constantinescu, A.O., O'Reilly, J.X., Behrens, T.E.J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*, 1464–1468.

Deneve, S., Peter, E., Latham, Alexandre, P. (1999). Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, *2*(8), 740–745.

Dordek, Y., Soudry, D., Meir, R., Derdikman, D. (2016). Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife*, *5*, e10094.

Eagleson, R. (1992). Measurement of the 2D affine Lie group parameters for visual motion analysis. *Spatial Vision*, *6*, 3.

Field, D.J. (1999). Wavelets, vision and the statistics of natural scenes. *Philosophical Transactions, Mathematical, Physical and Engineering Sciences*, *357*(1760), 2527.

Fiete, I.R., Burak, Y., Brookings, T. (2008). What grid cells convey about rat location. *Journal of Neuroscience*, *28*(27), 6858–6871.

Franzius, M., Sprekeler, H., Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *Plos Computational Biology*, *3*(8), 1605–1622.

Fuhs, M.C., & Touretzky, D.S. (2006). A spin glass model of path integration in rat medial entorhinal cortex. *Journal of Neuroscience*, *6*, 4266–4276.

Goyal, V.K., & Kovacevic, J. (2001). Quantized frame expansions with erasures. *Applied and Computational Harmonic Analysis*, *10*, 203–233.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., Moser, E.I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*, 801–806.

Hasselmo, M.E., Giocomo, L.M., Zilli, E.A. (2007). Grid cell firing may arise from interference of theta frequency membrane potential oscillations in single neurons. *Hippocampus*, *17*, 1252–1271.

Hebb, D.O. (1949). *The organization of behavior: a neuropsychological theory*. Wiley.

Heys, J.G., MacLeod, K.M., Moss, C.F., Hasselmo, M.E. (2013). Bat and rat neurons differ in theta frequency resonance despite similar coding of space. *Science*, *340*, 363–367.

Hubel, D.H., & Wiesel, T.N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, *28*(2), 229.

Hubel, D.H., & Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215.

Jacobs, J. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature Neuroscience*, *16*, 1188–1190.

Kay, S.M. (1993). *Fundamentals of statistical signal processing: estimation theory*. New Jersey: Englewood Cliffs.

Keinath, A., Epstein, R.A., Balasubramanian, V. (2018). Environmental deformations dynamically shift the grid cell spatial metric. *eLife*, *7*, 10.

Kim, M. (2019). Can we study 3d grid codes non-invasively in the human brain? Methodological considerations and fmri findings. *NeuroImage*, *186*, 667–678.

Kovacevic, J., & Chebira, A. (2007). Life beyond bases: the advent of frames (part i). *IEEE Signal Processing Magazine*, *24*(4), 86–104.

Kropff, E., & Treves, A. (2008). The emergence of grid cells: intelligent design or just adaptation? *Hippocampus*, *18*, 1256–1269.

Krupic, J., Bauza, M., Burton, S., Lever, C., O'Keefe, J. (2014). How environment geometry affects grid cell symmetry and what we can learn from it. *Philosophical Transactions of the Royal Society of London*, 369.

Krupic, J., Burgess, N., O'Keefe, J. (2012). Neural representations of location composed of spatially periodic bands. *Science*, *337*(6096), 853–857.

Mathis, A., Herz, A.V.M., Stemmler, M. (2012). Optimal population codes for space: grid cells outperform place cells. *Neural Computation*, *24*(9), 2280–2317.

McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser, E.I., Moser, M.B. (2006). Path integration and the neural basis of the 'cognitive map'. *Nature Reviews in the Neurosciences*, *7*, 663–678.

Mhatre, H., Gorchetchnikov, A., Grossberg, S. (2012). Grid cell hexagonal patterns formed by fast self-organized learning within entorhinal cortex. *Hippocampus*, *22*(2), 320–334.

Moser, E.I., & Moser, M.-B. (2013). Grid cells and neural coding in high-end cortices. *Neuron*, 80.

Moser, E.I., Roudi, Y., Witter, M.P., Kentros, C., Bonhoeffer, T., Moser, M.-B. (2014). Grid cells and cortical representation. *Nature Reviews Neuroscience*, *15*, 466–481.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*(3), 267–273.

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks*, *5*(6), 927–935.

Orchard, J., Yang, H., Ji, X. (2013). Does the entorhinal cortex use the fourier transform? *Frontiers in Computational Neuroscience*, *7*, 179.

Renart, A., Song, P., Wang, X.J. (2003). Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron*, *38*, 473–485.

Domínguez, U.R., & Caplan, J.B. (2018). A hexagonal fourier model of grid cells. *Hippocampus*, 09.

Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B.L., Witter, M.P., Moser, M.-B., Moser, E.I. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, *312*(5774), 758–762.

Schmidt-Hieber, C., & Häusser, M. (2013). Cellular mechanisms of spatial navigation in the medial entorhinal cortex. *Nature Neuroscience*, *16*, 325–331.

Botvinick, M.M., Stachenfeld, K.L., Gershman, S.J. (2017). The hippocampus as a predictive map. *Nature Neuroscisnce*.

Stachenfeld, K.L., Botvinick, M.M., Gershman, S.J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*, 1643–1653.

Staudigl, T., Leszczynski, M., Jacobs, J., Sheth, S.A., Schroeder, C.E., Jensen, O., Doeller, C.F. (2018). Hexadirectional modulation of high-frequency electrophysiological activity in the human anterior medial temporal lobe maps visual space. *Current Biology*, *28*(20), 3325–3329.e4.

Urdapilleta, E., Troiani, F., Stella, F., Treves, A. (2015). Can rodents conceive hyperbolic spaces? *Journal of the Royal Society Interface*, *12*, 107.

Vágó, L., & Ujfalussy, B.B. (2018). Robust and efficient coding with grid cells. *PLOS Computational Biology*, *14*(1), 1–28.

Yartsev, M.M., Witter, M.P., Ulanovsky, N. (2011). Grid cells without theta oscillations in the entorhinal cortex of bats. *Nature*, *479*, 103–107.

Yoon, H., & Sompolinsky, H. (1998). The effect of correlations on the fisher information of population codes. In *Proceedings of the 11th International Conference on Neural Information Processing Systems, NIPS 98* (pp. 167–173). Cambridge: MIT Press.

## Affiliations

**Fabio Anselmi**[1,2,3] · **Micah M. Murray**[4,5,6,7] · **Benedetta Franceschiello**[4,5]

[1] Center for Brains, Minds, and Machines, MIT, Cambridge,
MA, USA

[2] Laboratory for Computational and Statistical Learning
(LCSL) and IIT, Genova, Italy

[3] Center for Neuroscience and Artificial Intelligence
Department of Neuroscience, Baylor College of Medicine,
Houston, TX, USA

[4] The LINE (Laboratory for Investigative Neurophysiology),
Department of Radiology, University Hospital Center
and University of Lausanne, Lausanne, Switzerland

[5] Ophthalmology Department, Fondation Asile des aveugles
and University of Lausanne, Lausanne, Switzerland

[6] Sensory, Perceptual, and Cognitive Neuroscience Section,
Center for Biomedical Imaging (CIBM), Lausanne,
Switzerland

[7] Department of Hearing and Speech Sciences,
Vanderbilt University, Nashville, TN, USA