Learning place cells, grid cells and invariances with excitatory and inhibitory plasticity

Simon N. Weber[1] **and Henning Sprekeler**[1*]

***For correspondence:**
h.sprekeler@tu-berlin.de (HS)

[1]Modelling of Cognitive Processes, Institute of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Germany

Abstract Neurons in the hippocampus and adjacent brain areas show a large diversity in their tuning to location and head direction. The underlying circuit mechanisms are not resolved. In particular, it is unclear why certain cell types are selective to one spatial variable, but invariant to another. For example, place cells are typically invariant to head direction. We propose that all observed spatial tuning patterns – in both their selectivity and their invariance – arise from the same mechanism: Excitatory and inhibitory synaptic plasticity that is driven by the spatial tuning statistics of synaptic inputs. Using simulations and a mathematical analysis, we show that combined excitatory and inhibitory plasticity can lead to localized, grid-like or invariant activity. Combinations of different input statistics along different spatial dimensions reproduce all major spatial tuning patterns observed in rodents. The model is robust to changes in parameters, develops patterns on behavioral timescales and makes distinctive experimental predictions.

Introduction

Neurons in the hippocampus and the adjacent regions exhibit a broad variety of spatial activation patterns that are tuned to position, head direction or both. Common observations in these spatial dimensions are localized, bell shaped tuning curves (***O'Keefe, 1976***; ***Taube et al., 1990***), periodically repeating activity (***Fyhn et al., 2004***; ***Hafting et al., 2005***) and invariances (***Muller et al., 1994***; ***Burgess et al., 2005***), as well as combinations of these along different spatial dimensions (***Sargolini et al., 2006a***; ***Krupic et al., 2012***). For example, head direction cells are often invariant to location (***Burgess et al., 2005***), and place cells are commonly invariant to head direction (***Muller et al., 1994***). The cellular and network mechanisms that give rise to each of these firing patterns are subject to extensive experimental and theoretical research. Several computational models have been suggested to explain the emergence of grid cells (***Fuhs and Touretzky, 2006***; ***McNaughton et al., 2006***; ***Franzius et al., 2007a***; ***Burak and Fiete, 2009***; ***Couey et al., 2013***; ***Burgess et al., 2007***; ***Kropff and Treves, 2008***; ***Bush and Burgess, 2014***; ***Castro and Aguiar, 2014***; ***Dordek et al., 2016***; ***Stepanyuk, 2015***; ***Giocomo et al., 2011***; ***Zilli, 2012***; ***D'Albis and Kempter, 2017***; ***Monsalve-Mercado and Leibold, 2017***), place cells (***Tsodyks and Sejnowski, 1995***; ***Battaglia and Treves, 1998***; ***Arleo and Gerstner, 2000***; ***Solstad et al., 2006***; ***Franzius et al., 2007b***; ***Burgess and O'Keefe, 2011***; ***Franzius et al., 2007a***) and head direction cells (***McNaughton et al., 1991***; ***Redish et al., 1996***; ***Zhang, 1996***; ***Franzius et al., 2007a***).

Most of these models are designed to explain the spatial selectivity of one particular cell type and do not consider invariances along other dimensions, although the formation of invariant representations is a non-trivial problem (***DiCarlo and Cox, 2007***).

In view of the variety of spatial tuning patterns, the question arises if the differences in tuning of different cells in different areas reflect differences in microcircuit connectivity, single cell properties or plasticity rules, or if there is a unifying principle. In this paper we suggest that both the observed spatial selectivities and invariances can be explained by a common mechanism – interacting excitatory and inhibitory synaptic plasticity – and that the observed differences in the response profiles of grid, place and head direction cells result from differences in the spatial tuning of excitatory and inhibitory synaptic afferents. Here, we explore this hypothesis in a computational model of a feedforward network of rate-based neurons. Simulations as well as a mathematical analysis indicate that the model reproduces the large variety of response patterns of neurons in the hippocampal formation and adjacent areas and make predictions for the input statistics of each cell type.

Results

We study the development of spatial representations in a network of rate-based neurons with interacting excitatory and inhibitory plasticity. A single model neuron that represents a cell in the hippocampal formation or adjacent areas receives feedforward input from excitatory and inhibitory synaptic afferents. As a simulated rat moves through an environment, these synaptic afferents are weakly modulated by spatial location and in later sections also by head direction. This modulation is irregular and non-localized with multiple maxima (***Buetfering et al., 2014***); see ***Figure 1a*** and ***Methods and Materials***. Importantly, different inputs show different modulation profiles and each profile is temporally stable. We also show results for localized, i.e., place cell-like, input (***O'Keefe and Dostrovsky, 1971***; ***Marshall et al., 2002***; ***Wilent and Nitz, 2007***). The output rate is given by a weighted sum of the excitatory and inhibitory inputs.

In our model, both excitatory and inhibitory synaptic weights are subject to plasticity. The excitatory weights change according to a Hebbian plasticity rule (***Hebb, 1949***) that potentiates the weights in response to simultaneous pre- and postsynaptic activity. The inhibitory synapses evolve according to a plasticity rule that changes their weights in proportion to presynaptic activity and the difference between postsynaptic activity and a target rate (1 Hz in all simulations). This rule has previously been shown to balance excitation and inhibition such that the firing rate of the output neuron approaches the target rate (***Vogels et al., 2011***; ***D'amour and Froemke, 2015***). We assume the inhibitory plasticity to act fast enough to track changes of excitatory weights, so that excitation and inhibition are approximately balanced at all times.

The relative spatial smoothness of the excitatory and inhibitory input determines the firing pattern of the output neuron

We first simulate a rat that explores a linear track (***Figure 1***). The spatial tuning of each input neuron is stable in time and depends smoothly on the location of the animal, but is otherwise random (e.g., ***Figure 1a***). As a measure of smoothness, we use the spatial autocorrelation length. In the following, this is the central parameter of the input statistics, which is chosen separately for excitation and inhibition. In short, we assume that temporally stable spatial information is presynaptically present but we have minimal requirements on its format, aside from the spatial autocorrelation length.

At the beginning of each simulation, all synaptic weights are random. As the animal explores the track, the excitatory and inhibitory weights change in response to pre- and postsynaptic activity, and the output cell gradually develops a spatial activity pattern. We find that this pattern is primarily determined by whether the excitatory or inhibitory inputs are smoother in space. If the inhibitory tuning is smoother than the excitatory tuning (***Figure 1b***), the output neuron develops equidistant firing fields, reminiscent of grid cells on a linear track (***Hafting et al., 2008***). If instead the excitatory tuning is smoother, the output neuron fires close to the target rate of 1 Hz everywhere (***Figure 1c***); it develops a spatial invariance. For spatially untuned inhibitory afferents (***Grienberger et al., 2017***), the output neuron develops a single firing field, reminiscent of a one-dimensional place cell (***Figure 1d***); compare (***Clopath et al., 2016***).

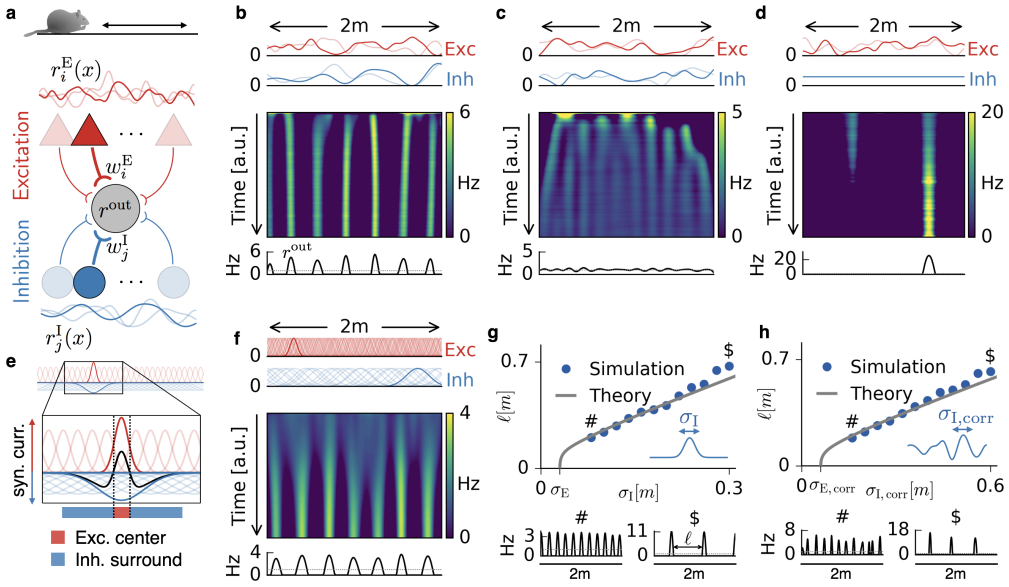

Figure 1. Emergence of periodic, invariant and single field firing patterns. **a**) Network model for a linear track. A threshold-linear output neuron (gray) receives input from excitatory (red) and inhibitory (blue) cells, which are spatially tuned (curves on top and bottom). **b**) Spatially tuned input with smoother inhibition than excitation. The fluctuating curves (top) show two exemplary spatial tunings (one is highlighted) of excitatory and inhibitory input neurons. Interacting excitatory and inhibitory synaptic plasticity gradually changes an initially random response of the output neuron (firing rate r^{out}) into a periodic, grid cell-like activity pattern. **c**) If the spatial tuning of inhibitory input neurons is less smooth than that of excitatory input neurons the interacting excitatory and inhibitory plasticity leads to a spatially invariant firing pattern. The output neuron fires close to the target rate of 1 Hz everywhere. **d**) For very smooth or spatially untuned inhibitory inputs, the output neuron develops a single firing field, reminiscent of a place cell. **e**) The mechanism, illustrated for place cell-like input. When a single excitatory weight is increased relative to the others, the balancing inhibitory plasticity rule leads to an immediate increase of inhibition at the associated location. If inhibitory inputs are smoother than excitatory inputs, the resulting approximate balance creates a center surround field: a local overshoot of excitation (firing field) surrounded by an inhibitory corona. The next firing field emerges at a distance where the inhibition has faded out. Iterated, this results in a spatially periodic arrangement of firing fields. **f**) Inputs with place field-like tuning. Gaussian curves (top) show the spatial tuning of excitatory and inhibitory input neurons (one neuron of each kind is highlighted, 20 percent of all inputs are displayed). A grid cell firing pattern emerges from an initially random weight configuration. **g**) Grid spacing ℓ scales with inhibitory tuning width σ_{I}. Simulation results (dots) agree with a mathematical bifurcation analysis (solid). Output firing rate examples at the two indicated locations are shown at the bottom. **h**) Inhibitory smoothness $\sigma_{\text{I,corr}}$ controls grid spacing; arrangement as in **d**. Note that the time axes in **b,c,d,f** are different, because the speed at which the patterns emerge is determined by both the learning rates of the plasticity and the firing rate of the input neurons. We kept the learning rate constant and adjusted the simulation times to achieve convergence. Choosing identical simulation times, but different learning rates, leads to identical results (***Figure 1–Figure supplement 2***).
Rat clip art from (***hartmut, 2015***).

Figure 1–Figure supplement 1. Statistics of the synaptic weights.

Figure 1–Figure supplement 2. Different learning rates lead to identical results.

The emergence of these firing patterns can be best explained in the simplified scenario of place field-like input tuning (***Figure 1e,f***). The spatial smoothness is then given by the size of the place fields. Let us assume that the output neuron fires at the target rate everywhere (see ***Methods and Materials***). From this homogeneous state, a small potentiation of one excitatory weight leads to an increased firing rate of the output neuron at the location of the associated place field (highlighted red curve in ***Figure 1e***). To bring the output neuron back to the target rate, the inhibitory learning rule increases the synaptic weight of inhibitory inputs that are tuned to the same location (highlighted blue curve in ***Figure 1e***). If these inhibitory inputs have smaller place fields than the excitatory inputs (***Figure 1c***), this restores the target rate everywhere (***Vogels et al., 2011***). Hence, inhibitory plasticity can stabilize spatial invariance if the inhibitory inputs are sufficiently precise (i.e., not too smooth) in space. In contrast, if the spatial tuning of the inhibitory inputs is smoother than that of the excitatory inputs, the target firing rate cannot be restored everywhere. Instead, the compensatory potentiation of inhibitory weights increases the inhibition in a spatial region that has at least the size of the inhibitory place fields. This leads to a corona of inhibition, in which the output neuron cannot fire (***Figure 1e***, blue region). Outside of this inhibitory surround the output neuron can fire again and the next firing field develops. Iterated, this results in a periodic arrangement of firing fields (***Figure 1f*** and ***Figure 7b*** for a depiction of the input currents). Spatially untuned inhibition corresponds to a large inhibitory corona that exceeds the length of the linear track, so that only a single place field remains. From a different perspective, spatially untuned input can also be understood as a limit case of vanishing spatial variation in the firing rate rather than a limit of infinite smoothness. Consistent with this view, a development of grid patterns or invariance requires a sufficiently strong spatial modulation of the inhibitory inputs (***Methods and Materials***).

The argument of the preceding paragraph can be extended to the scenario where input is irregularly modulated by space. For non-localized input tuning (***Figure 1b,c,d***), any weight change that increases synaptic input in one location will also increase it in a surround that is given by the smoothness of the input tuning (see ***Methods and Materials*** for a mathematical analysis). In the simulations, the randomness manifests itself in occasional defects in the emerging firing pattern (***Figure 1h***, bottom, and ***Figure 1–Figure supplement 1***). The above reasoning suggests that the width of individual firing fields is determined by the smoothness of the excitatory input tuning, while the distance between grid fields, i.e., the grid spacing, is set by the smoothness of the inhibitory input tuning. Indeed, both simulations and a mathematical analysis (***Methods and Materials***) confirm that the grid spacing scales linearly with the inhibitory smoothness in a large range, both for localized (***Figure 1g***) and non-localized input tuning (***Figure 1h***). The analysis also reveals a weak logarithmic dependence of the grid spacing on the ratio of the learning rates, the mean firing rates and the number of afferents of the excitatory and inhibitory population (***Equation 78*** and ***Figure 8b***).

In summary, the interaction of excitatory and inhibitory plasticity can lead to spatial invariance, spatially periodic activity patterns or single place fields depending on the spatial statistics of the excitatory and inhibitory input.

Emergence of hexagonal firing patterns

When a rat navigates in a two-dimensional arena, the spatial firing maps of grid cells in the medial entorhinal cortex (mEC) show a pronounced hexagonal symmetry (***Hafting et al., 2005***; ***Fyhn et al., 2004***) with different grid spacings and spatial phases. To study whether a hexagonal firing pattern can emerge from interacting excitatory and inhibitory plasticity, we simulate a rat in a quadratic arena. The rat explores the arena for 10 hours, following trajectories extracted from behavioral data (***Sargolini et al., 2006b***); ***Methods and Materials***. To investigate the role of the input statistics, we consider three different classes of input tuning: i) place cell-like input (***Figure 2a***), ii) sparse non-localized input, in which the tuning of each input neuron is given by the sum of 100 randomly located place fields (***Figure 2b***) and iii) dense non-localized input, in which the tuning of each input is a random function with fixed spatial smoothness (***Figure 2c***). For all input classes, the spatial

tuning of the inhibitory inputs is smoother than that of the excitatory inputs.

Initially, all synaptic weights are random and the activity of the output neuron shows no spatial symmetry. While the rat forages through the environment, the output cell develops a periodic firing pattern for all three input classes, reminiscent of grid cells in the mEC (***Fyhn et al., 2004***; ***Hafting et al., 2005***) and typically with the same hexagonal symmetry. This hexagonal arrangement is again due to the smoother inhibitory input tuning, which generates a spherical inhibitory corona around each firing field (compare ***Figure 1e***). These center-surround fields arrange in a hexagonal pattern – the closest packing of spheres in two dimensions; compare (***Turing, 1952***). We find that the spacing of this pattern is determined by the inhibitory smoothness. The similarity between cells in terms of orientation and phase of the grid depend – in decreasing order – on whether they receive the same inputs, on the trajectories on which the tuning was learned and on the initial synaptic weights (***Figure 2–Figure supplement 1***). Two grid cells can thus have a different phase and orientation, even if they share a large fraction or all of their inputs.

For the linear track, the randomness of the non-localized inputs leads to defects in the periodicity of the grid pattern. In two dimensions, we find that the randomness leads to distortions of the hexagonal grid. To quantify this effect, we simulated 500 random trials for each of the three input scenarios and plotted the grid score histogram (***Appendix 1***) before and after 10 hours of spatial exploration (***Figure 2d,e,f***). Different trials have different trajectories, different initial synaptic weights and different random locations of the input place fields (for sparse input) or different random input functions (for dense input). For place cell-like input, most of the output cells develop a positive grid score during 10 hours of spatial exploration (33% before to 86% after learning, ***Figure 2d***). Even for low grid scores, the firing rate maps look grid-like after learning but exhibit a distorted symmetry (***Figure 2d***). For sparse non-localized input, the fraction of output cells with a positive grid score increases from 35% to 87% and for dense non-localized input from 16% to 68% within 10 hours of spatial exploration (***Figure 2e,f***). The excitatory and inhibitory inputs are not required to have the same tuning statistics. Grid patterns also emerge when excitation is localized and inhibition is non-localized (***Figure 2–Figure supplement 6***).

In summary, the interaction of excitatory and inhibitory plasticity leads to grid-like firing patterns in the output neuron for all three input scenarios. The grids are typically less distorted for sparser input (***Figure 2g***).

Rapid appearance of grid cells and their reaction to modifications of the environment

In unfamiliar environments, neurons in the mEC exhibit grid-like firing patterns within minutes (***Hafting et al., 2005***). Moreover, grid cells react quickly to changes in the environment (***Fyhn et al., 2007***; ***Savelli et al., 2008a***; ***Barry et al., 2012***). These observations challenge models for grid cells that require gradual synaptic changes during spatial exploration. In principle, the time scale of plasticity-based models can be augmented arbitrarily by increasing the synaptic learning rates. For stable patterns to emerge, however, significant weight changes must occur only after the animal has visited most of the environment. To explore the edge of this trade-off between speed and stability, we increased the learning rates to a point where the grids are still stable but where further increase would reduce the stability (***Figure 3–Figure supplement 1***). For place cell-like input, periodic patterns can be discerned within 10 minutes of spatial exploration, starting with random initial weights (***Figure 3a,b***). The pattern further emphasizes over time and remains stable for many hours (***Figure 3c*** and ***Figure 3–Figure supplement 3***).

To investigate the robustness of this phenomenon we ran 500 realizations with different trajectories, initial synaptic weights and locations of input place fields. In all simulations, a periodic pattern emerged within the first 30 minutes, and a majority of patterns exhibited hexagonal symmetry after three hours (increasing from 33% to 81%, ***Figure 3c,d***). For non-localized input, the emergence of the final grids typically takes longer, but the first grid fields are also observed within minutes and are still present in the final grid, as observed in experiments (***Hafting et al., 2005***) (***Figure 3–Figure***

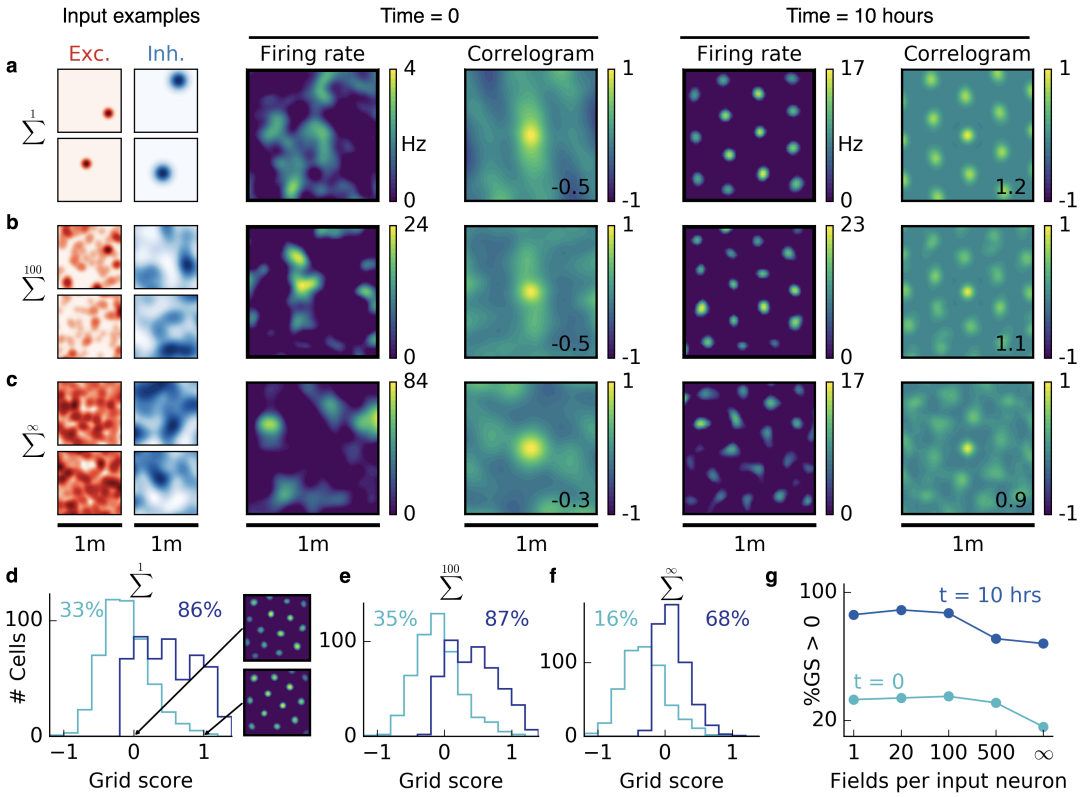

Figure 2. Emergence of two dimensional grid cells. **a,b,c**) Columns from left to right: Spatial tuning of excitatory and inhibitory input neurons (two examples each); spatial firing rate map of the output neuron and corresponding autocorrelogram before and after spatial exploration of 10 hours. The number on the correlogram shows the associated grid score. Different rows correspond to different spatial tuning characteristics of the excitatory and inhibitory input. For all figures the spatial tuning of inhibitory input neurons is smoother than that of excitatory input neurons. **a**) Each input neuron is a place cell with random location. **b**) The tuning of each input neuron is given as the sum of 100 randomly located place fields. **c**) The tuning of each input neuron is a random smooth function of the location. This corresponds to the sum of infinitely many randomly located place fields. Before learning, the spatial tuning of the output neuron shows no symmetry. After 10 hours of spatial exploration the output neuron developed a hexagonal pattern. **d**) Grid score histogram for 500 output cells with place cell-like input. Before learning (light blue) 33% of the output cells have a positive grid score. After 10 hours of spatial exploration (dark blue) this value increases to 86%. Two example rate maps are shown. The arrows point to the grid score of the associated rate map. Even for low grid scores the learned firing pattern looks grid-like. **e,f**) Grid score histograms for input tuning as in **b,c**, arranged as in **d**. **g**) Fraction of neurons with positive grid score before (light blue) and after learning (dark blue) as a function of the number of fields per input neuron. Note that in order to learn within 10 hours of exploration time, we used different learning rates for different input scenarios. Using identical learning rates for all input scenarios but adjusting the simulation times to achieve convergence leads to identical results (***Figure 2–Figure supplement 5***).

Figure 2–Figure supplement 1. Influence of random simulation parameters on the final grid pattern.

Figure 2–Figure supplement 2. Boundary effects in simulations with place field-like input.

Figure 2–Figure supplement 3. Distribution of input fields.

Figure 2–Figure supplement 4. Synaptic weight normalization does not influence the grids.

Figure 2–Figure supplement 5. Different learning rates lead to identical results.

Figure 2–Figure supplement 6. Using different input statistics for different populations also leads to hexagonal firing patterns.

supplement 2).

Above, we modeled the exploration of a previously unknown room by assuming the initial synaptic weights to be randomly distributed. If the rat had previous exposure to the room or to a similar room, a structure might already have formed in some of the synaptic weights. This structure could aid the development of the grid in similar rooms or hinder it in a novel room. To study this, we simulate a network that first learns the synaptic weights in one room. We then introduce a graded modification of the room by remapping the firing fields of a fraction of input neurons to random locations. We find that the output firing pattern is robust to such perturbations, even if more than half of the inputs are remapped (*Figure 3–Figure supplement 3*). If all inputs are changed, corresponding to a novel room, a grid pattern is learned anew. The strong initial pattern in the weights does not hinder this development (*Figure 3–Figure supplement 3*).

Recently, Wernle et al. (*Wernle et al., 2018*) discovered that in an arena separated by a wall, single grid cells form two independent grid patterns — one on each side of the wall — that coalesce once the wall is removed. They find that grid fields close to the partition wall move in order to establish a more coherent pattern. In contrast, fields far away from the partition wall do not change their locations. Rosay et al. reproduced this experimental finding by simulating grid fields as interacting particles (*Rosay et al., 2018*). They also demonstrated how it could be reproduced by a feedforward model for grid cells based on firing rate adaptation (*Rosay et al., 2018*; *Kropff and Treves, 2008*). Inspired by these experiments, we simulate a rat that first explores one half of a quadratic arena and then the other half, for 2.5 hours each (*Figure 4a*). A grid pattern emerges in each compartment (*Figure 4b,c*). We then remove the partition wall and the rat explores the entire arena for another 5 hours (*Figure 4a*). As observed experimentally, grid fields close to the former partition line rearrange to make the two grids more coherent and grid fields far away from the partition line basically stay where they were (*Figure 4d*).

In summary, periodic patterns emerge rapidly in our model and the associated time scale is limited primarily by how quickly the animal visits its surroundings, i.e., by the same time scale that limits the experimental recognition of the grids.

Place cells, band cells and stretched grids

In addition to grids, the mEC and adjacent brain areas exhibit a plethora of other spatial activity patterns including spatially invariant (*Burgess et al., 2005*), band-like (*Krupic et al., 2012*) (periodic along one direction and invariant along the other), and spatially periodic but non-hexagonal patterns (*Krupic et al., 2012*; *Hardcastle et al., 2017*; *Diehl et al., 2017*). Note that it is currently debated whether or not some of the observed spatially periodic but non-hexagonal firing patterns are artifacts of poorly isolated single cell data in multi-electrode recordings (*Navratilova et al., 2016*; *Krupic et al., 2015b*). In contrast to spatially periodic tuning, place cells in the hippocampus proper are typically only tuned to a single or few locations in a given environment (*O'Keefe and Dostrovsky, 1971*; *Moser et al., 2008*; *Leutgeb et al., 2005*). If the animal traversed the environment along a straight line, all of these cells would be classified as periodic, localized or invariant (*Figure 1*), although the classification could vary depending on the direction of the line. Based on this observation, we hypothesized that all of these patterns could be the result of an input autocorrelation structure that differs along different spatial directions.

We first verified that also in a two-dimensional arena, place cells emerge from a very smooth inhibitory input tuning (*Figure 5a,b*). The emergence of place cells is independent of the exact shape of the excitatory input. Non-localized inputs (*Figure 5a*) lead to similar results as grid cell-like inputs of different orientation and grid spacing (*Figure 5b*, *Methods and Materials*); for other models for the emergence of place cells from grid cells see (*Solstad et al., 2006*; *Franzius et al., 2007b*; *Rolls et al., 2006*; *Molter and Yamaguchi, 2008*; *Ujfalussy et al., 2009*; *Savelli and Knierim, 2010*). Next we verified that also in two dimensions, spatial invariance results when excitation is broader than inhibition (*Figure 5c*). We then varied the smoothness of the inhibitory inputs independently along two spatial directions. If the spatial tuning of inhibitory inputs is smoother than the tuning of the

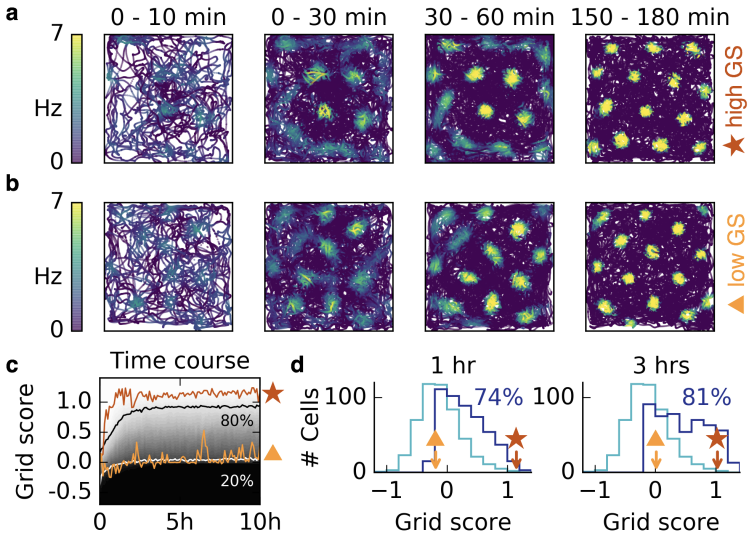

Figure 3. Grid patterns form rapidly during exploration and remain stable for many hours. **a**,**b**) Rat trajectories with color-coded firing rate a of cell that receives place cell-like input. The color depicts the firing rate at the time of the location visit, not after learning. Bright colors indicate higher firing rates. The time interval of the trajectory is shown above each plot. Initially all synaptic weights are set to random values. Part **a** and **b** show two different realizations with a good (red star) and a bad (orange triangle) grid score development. After few minutes a periodic structure becomes visible and enhances over time. **c**) Time course of the grid score in the simulations shown in **a** (red) and **b** (orange). While the periodic patterns emerge within minutes, the manifestation of the final hexagonal pattern typically takes a couple of hours. Once the pattern is established it remains stable for many hours. The gray scale shows the cumulative histogram of the grid scores of 500 realizations (black=0, white=1). The solid white and black lines indicate the 20% and 80% percentile, respectively. **d**) Histogram of grid scores of the 500 simulations shown in **c**. Initial histogram in light blue, histogram after 1 hour and after 3 hours in dark blue. Numbers show the fraction of cells with positive grid score at the given time. Rat trajectories taken from (***Sargolini et al., 2006b***).

Figure 3–Figure supplement 1. Too fast learning leads to unstable grids.

Figure 3–Figure supplement 2. Rapid development of grid patterns from non-localized input.

Figure 3–Figure supplement 3. Influence of input remapping on grid patterns.

excitatory inputs along one dimension but less smooth along the other, the output neuron develops band cell-like firing patterns (***Figure 5d***). If inhibitory input is smoother than excitatory input, but not isotropic, the output cell develops stretched grids with different spacing along two axes (***Figure 5e***). For these anisotropic cases, stretched hexagonal grids and rectangular arrangements of firing fields appear similarly favorable (compare ***Figure 5e***, second row and column). A hexagonal arrangement is favored by a dense packing of inhibitory coronas, whereas a rectangular arrangement would maximize the proximity of the excitatory centers, given the inhibitory corona (***Figure 5–Figure supplement 1***).

In summary, the relative spatial smoothness of inhibitory and excitatory input determines the symmetry of the spatial firing pattern of the output neuron. The requirements for the input tuning that support invariance, periodicity and localization apply individually to each spatial dimension, opening up a combinatorial variety of spatial tuning patterns.

Spatially tuned input combined with head direction selectivity leads to grid, conjunctive and head direction cells

Many cells in and around the hippocampus are tuned to the head direction of the animal (***Taube et al., 1990***; ***Taube, 1995***; ***Chen et al., 1993***). These head direction cells are typically tuned to a single head direction, just like place cells are typically tuned to a single location. Moreover, head direction cells are often invariant to location (***Burgess et al., 2005***), just like place cells are commonly invariant

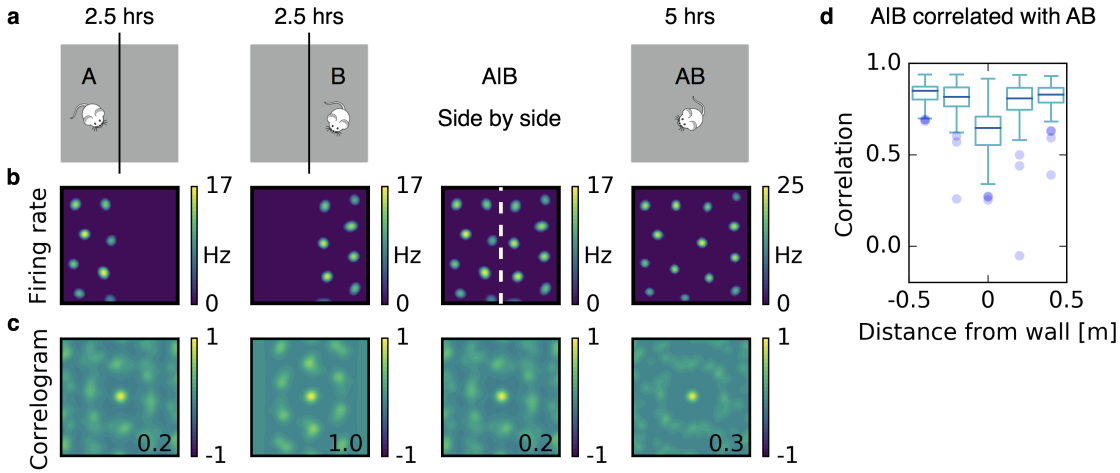

Figure 4. Grids coalesce in contiguous environments. **a**) Illustration of the experiment. A quadratic arena (gray box) is divided into two rectangular compartments by a wall (black line). The animal explores one compartment (A) and then the other compartment (B) for 2.5 hours each. Then the wall is removed and the rat explores the entire arena (AB) for 5 hours. **b**) Firing rate maps. From left to right: After learning in A; after learning in B; the maps from A and B shown side by side (A|B); after learning in AB. **c**) Autocorrelograms of the rate maps shown in **b**. The number inside the correlogram shows the grid score. **d**) Box plot of the correlations of the firing rate map A|B with the firing rate map AB as a function of distance from the partition wall. Close to the partition wall the correlation is low, far away from the partition wall it is high. This indicates that grid fields rearrange only locally. Each box extends from the first to the third quartile, with a dark blue line at the median. The whiskers extend from the first and third quartile by 1.5 the interquartile range. Dots show flier points. Data: 100 realizations of experiments as in **a,b,c**. For simulation details see ***Appendix 1***.
Mouse clip art from (***lemmling, 2006***).

to head direction (***Muller et al., 1994***). There are also cell types with conjoined spatial and head direction tuning. Conjunctive cells in the mEC fire like grid cells in space, but only in a particular head direction (***Sargolini et al., 2006a***), and many place cells in the hippocampus of crawling bats also exhibit a head direction tuning (***Rubin et al., 2014***).
To investigate whether these tuning properties could also result in our model, we simulated a rat that moves in a square box, whose head direction is constrained by the direction of motion (***Appendix 1***). Each input neuron is tuned to both space and head direction (see ***Figure 6*** for localized and ***Figure 6–Figure supplement 1*** for non-localized input).

In line with the previous observations, we find that the *spatial* tuning of the output neuron is determined by the relative *spatial* smoothness of the excitatory and inhibitory inputs and the *head direction* tuning of the output neuron is determined by the relative smoothness of the *head direction* tuning of the inputs from the two populations. If the head direction tuning of excitatory input neurons is smoother than that of inhibitory input neurons, the output neuron becomes invariant to head direction (***Figure 6a***). If instead only the excitatory input is tuned to head direction, the output neuron develops a single activity bump at a particular head direction (***Figure 6b,c***). The concurrent spatial tuning of the inhibitory input neurons determines the spatial tuning of the output neuron. For spatially smooth inhibitory input, the output neuron develops a hexagonal firing pattern (***Figure 6a,b***) and for less smooth inhibitory input the firing of the output neuron is invariant to the location of the animal (***Figure 6c***).

In summary, the relative smoothness of inhibitory and excitatory input neurons in space and in head direction determines whether the output cell fires like a pure grid cell, a conjunctive cell or a pure head direction cell (***Figure 6d***).

We find that the overall head direction tuning of conjunctive cells is broader than that of individual grid fields (***Figure 6e***). This results from variations in the preferred head direction of different grid fields. Typically, however, these variations remain small enough to preserve an overall

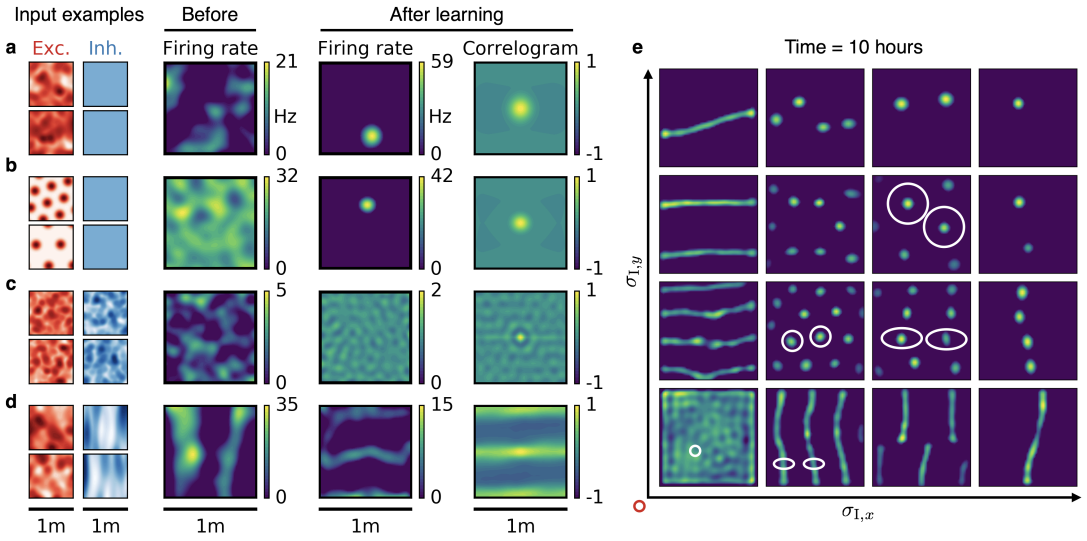

Figure 5. Emergence of spatially tuned cells of diverse symmetries. **a,b,c,d**) Arrangement as in ***Figure 2***. **a,b**) Place cells emerge if the inhibitory autocorrelation length exceeds the box length or if the inhibitory neurons are spatially untuned. The type of tuning of the excitatory input is not crucial: Place cells develop for non-localized input (**a**) as well as for grid cell input (**b**). **c**) The output neuron develops an invariance, if the spatial tuning of inhibitory input neurons is less smooth than the tuning of excitatory input neurons. **d**) Band cells emerge if the spatial tuning of inhibitory input is asymmetric, such that the autocorrelation length is larger than that of excitatory input along one direction (here the y-direction) and smaller along the other (here the x-direction). **e**) Overview of how the shape of the inhibitory input tuning determines the firing pattern of the output neuron. Each element depicts the firing rate map of the output neuron after 10 hours. White ellipses of width $2\sigma_{I,x}$ and $2\sigma_{I,y}$ in $x-$ and $y-$direction indicate the direction-dependent standard deviation of the spatial tuning of the inhibitory input neurons. For simplicity, the width of the excitatory tuning fields, σ_E, is the same in all simulations. It determines the size of the circular firing fields. The red circle at the axis origin is of diameter $2\sigma_E$.

Figure 5–Figure supplement 1. Arrangement of firing fields for asymmetric input.

head direction tuning of the cell, because individual grid fields tend to align their head direction tuning (compare to ***Figure 5–Figure supplement 1***, but in three dimensions). Whether a narrower head direction of individual grid fields or a different preferred direction for different grid fields is present also in rodents is not resolved (***Figure 6–Figure supplement 2***).

Discussion

We presented a self-organization model that reproduces the experimentally observed spatial and head direction tuning patterns in the hippocampus and adjacent brain regions. Its core mechanism is an interaction of Hebbian plasticity in excitatory synapses and homeostatic Hebbian plasticity in inhibitory synapses (***Vogels et al., 2011***; ***D'amour and Froemke, 2015***). The main prediction of the model is that the spatial autocorrelation structure of excitatory and inhibitory inputs determines – and should thus be predictable from – the output pattern of the cell. Investigations of the tuning of individual cells (***Wertz et al., 2015***) or even synapses (***Wilson et al., 2016***) that project to spatially tuned cells would thus be a litmus test for the proposed mechanism.

Origin of spatially tuned synaptic input

The origin of synaptic input to spatially tuned cells is not fully resolved (***Van Strien et al., 2009***). Given that our model is robust to the precise properties of the input, it is consistent with input from higher sensory areas (***Tanaka, 1996***; ***Quiroga et al., 2005***) that could inherit a spatial tuning from their sensory tuning in a stable environment (***Arleo and Gerstner, 2000***; ***Franzius et al., 2007a***). This is in line with the observation that grid cells lose their firing profiles in darkness (***Chen et al., 2016***; ***Pérez-Escobar et al., 2016***) and that the hexagonal pattern rotates when a visual cue card is

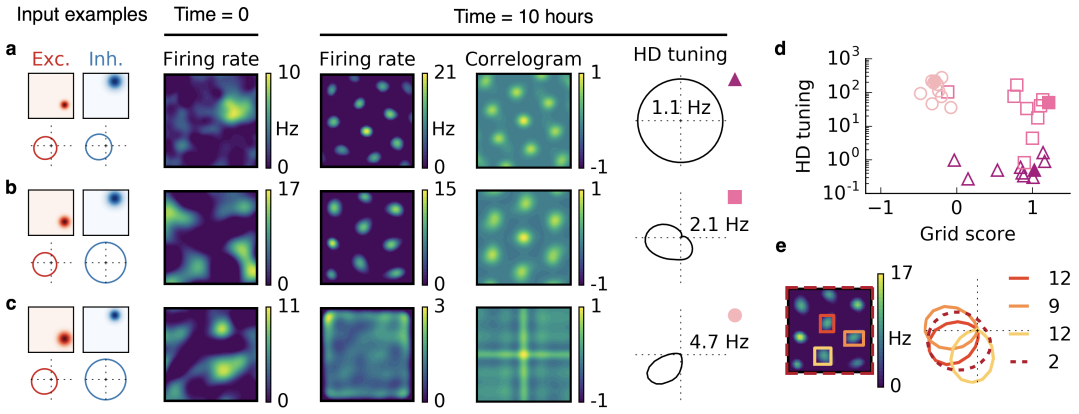

Figure 6. Combined spatial and head direction tuning. **a,b,c**) Columns from left to right: Spatial tuning and head direction tuning (polar plot) of excitatory and inhibitory input neurons (one example each); spatial firing rate map of the output neuron before learning and after spatial exploration of 10 hours with corresponding autocorrelogram; head direction tuning of the output neuron after learning. The numbers in the polar plots indicate the peak firing rate at the preferred head direction after averaging over space. **a**) Wider spatial tuning of inhibitory input neurons than of excitatory input neurons combined with narrower head direction tuning of inhibitory input neurons leads to a grid cell-like firing pattern in space with invariance to head direction, i.e., the output neuron fires like a pure grid cell. **b**) The same spatial input characteristics as in **a** combined with head direction-invariant inhibitory input neurons leads to grid cell-like activity in space and a preferred head direction, i.e. the output neuron fires like a conjunctive cell. **c**) If the spatial tuning of inhibitory input neurons is less smooth than that of excitatory neurons and the concurrent head direction tuning is wider for inhibitory than for excitatory neurons, the output neuron is not tuned to space but to a single head direction, i.e. the output neuron fires like a pure head direction cell. **d**) Head direction tuning and grid score of 10 simulations of the three cell types. Each symbol represents one realization with random input tuning. The markers correspond to the tuning properties of the input neurons as depicted in **a, b, c**: grid cell (triangles), conjunctive cell (squares), head direction cell (circles). The values that correspond to the output cells in **a, b, c** are shown as filled symbols. **e**) In our model, the head direction tuning of individual grid fields is sharper than the overall head direction tuning of the conjunctive cell. Depicted is a rate map of a conjunctive cell (left) and the corresponding head direction tuning (right, dashed). For three individual grid fields, indicated with colored squares, the head direction tuning is shown in the same polar plot. The overall tuning of the grid cell (dashed) is a superposition of the tuning of all grid fields. Numbers indicate the peak firing rate (in Hz) averaged individually within each of the four rectangles in the rate map.

Figure 6–Figure supplement 1. Combined spatial and head direction tuning with non-localized inputs.

Figure 6–Figure supplement 2. Head direction tuning of individual grid fields is difficult to assess from grid cells with few firing fields.

rotated (***Pérez-Escobar et al., 2016***).

The input could also stem from within the hippocampal formation, where spatial tuning has been observed in both excitatory (***O'Keefe, 1976***) and inhibitory (***Marshall et al., 2002***; ***Wilent and Nitz, 2007***; ***Hangya et al., 2010***) neurons. For example, the notion that mEC neurons receive input from hippocampal place cells is supported by several studies: Place cells in the hippocampus emerge earlier during development than grid cells in the mEC (***Langston et al., 2010***; ***Wills et al., 2010***), grid cells lose their tuning pattern when the hippocampus is deactivated (***Bonnevie et al., 2013***) and both the firing fields of place cells and the spacing and field size of grid cells increase along the dorso-ventral axis (***Jung et al., 1994***; ***Brun et al., 2008b***; ***Stensola et al., 2012***). Moreover, entorhinal stellate cells, which often exhibit grid-like firing patterns, receive a large fraction of their input from the hippocampal CA2 region (***Rowland et al., 2013***), where many cells are tuned to the location of the animal (***Martig and Mizumori, 2011***).

Inhibition is usually thought to arise from local interneurons – but see (***Melzer et al., 2012***)) – suggesting that spatially tuned inhibitory input to mEC neurons originates from the entorhinal cortex itself. Interneurons in mEC display a spatial tuning (***Buetfering et al., 2014***; ***Savelli et al., 2008b***; ***Frank et al., 2001***) that could be inherited from hippocampal place cells, other grid cells

(*Couey et al., 2013*; *Pastoll et al., 2013*; *Winterer et al., 2017*) or from entorhinal cells with nongrid spatial tuning (*Diehl et al., 2017*; *Hardcastle et al., 2017*). The broader spatial tuning required for the emergence of spatial selectivity could be established, e.g., by pooling over cells with similar tuning or through a non-linear input-output transformation in the inhibitory circuitry. If inhibitory input is indeed local, the increase in grid spacing along the dorso-ventral axis (*Brun et al., 2008b*) suggests that the tuning of inhibitory interneurons gets smoother along this axis. For smoother tuning functions, less neurons are needed to cover the whole environment, in accordance with the decrease in interneuron density along the dorso-ventral axis (*Beed et al., 2013*).

The excitatory input to hippocampal place cells could originate from grid cells in entorhinal cortex (*Figure 5b*), which is supported by anatomical (*Van Strien et al., 2009*) and lesion studies (*Brun et al., 2008a*). The required untuned inhibition could arrive from interneurons in the hippocampus proper that often show very weak spatial tuning (*Marshall et al., 2002*). In addition to grid cell input, place cells are also thought to receive inputs from other cells types, such as border cells (*Muessig et al., 2015*), and other brain regions such as the medial septum (*Wang et al., 2015*).

Dissociation from continuous attractor network models

The observed spatial tuning patterns have also been explained by other models. In continuous attractor networks (CAN), each cell type could emerge from a specific recurrent connectivity pattern, combined with a mechanism that translates the motion of the animal into shifts of neural activity on an attractor. How the required connectivity patterns – which lie at the core of any CAN model – could emerge is subject to debate (*Widloski and Fiete, 2014*). Our model is qualitatively different in that it does not rely on attractor dynamics in a recurrent neural network, but on experience-dependent plasticity of spatially modulated afferents to an individual output neuron (*Mehta et al., 2000*). A measurable distinction of our model from CAN models is its response to a rapid global reduction of inhibition. While a modification of inhibition typically changes the grid spacing in CAN models of grid cells (*Couey et al., 2013*; *Widloski and Fiete, 2015*), the grid field locations generally remain untouched in our model. The grid fields merely change in size, until inhibition is recovered by inhibitory plasticity (*Figure 7a*). This can be understood by the colocalization of the grid fields and the peaks in the excitatory membrane current (*Figure 7b,c*). A reduction of inhibition leads to an increased protrusion of these excitatory peaks and thus to wider firing fields. Grid patterns in mEC are temporally stable in spite of dopaminergic modulations of GABAergic transmission (*Cilz et al., 2013*) and the spacing of mEC grid cells remains constant during the silencing of inhibitory interneurons (*Miao et al., 2017*). Both observations are in line with our model.

Moreover, we found that for localized input tuning, the inhibitory membrane current typically also peaks at the locations of the grid fields. This co-tuning breaks down for non-localized input (*Figure 7b*). In contrast, CAN models predict that the inhibitory membrane current has the same periodicity as the grid (*Schmidt-Hieber and Häusser, 2013*), but possibly phase shifted.

The grid patterns of topologically nearby grid cells in the mEC typically have the same orientation and spacing but different phases (*Hafting et al., 2005*). Moreover, the coupling between anatomically nearby grid cells – e.g., the difference in spatial phase – is more stable to changes of the environment than the firing pattern of individual grid cells (*Yoon et al., 2013*). These properties are immanent to CAN models. In contrast, single cell models (*Burgess et al., 2007*; *Kropff and Treves, 2008*; *Castro and Aguiar, 2014*; *Stepanyuk, 2015*; *Dordek et al., 2016*; *D'Albis and Kempter, 2017*; *Monsalve-Mercado and Leibold, 2017*) require additional mechanisms to develop a coordination of neighboring grid cells. The challenge for any mechanism is to correlate the grid orientations, but leave the grid phases uncorrelated. The most obvious candidate, recurrent connections among different grid cells (*Si et al., 2012*), requires an intricate combination of mechanisms to perform this balancing act. We assume that an appropriate recurrent connectivity would not be simpler in our model.

CAN models predict that all grid fields in a conjunctive (grid x head direction) cell have the same head direction tuning, whereas our model predicts that there could be differences between

different grid fields (***Figure 6e***). Our preliminary analysis suggests that an in-depth evaluation would require data for central grid fields without trajectory biases (***Figure 6–Figure supplement 2***), which are at present not publicly available.

In addition, CAN models require that conjunctive (grid x head direction) cells are positively modulated by running speed. Such a modulation has been observed in experiments (***Kropff et al., 2015***). In our model, we could introduce a running speed dependence, e.g., as a global modulation of the input signals. We expect that in this case, the output neuron would inherit a speed tuning from the input but would otherwise develop similar spatial tuning patterns.

A recent analysis has shown that the periodic firing of entorhinal cells in rats that move on a linear track can be assessed as slices through a hexagonal grid (***Yoon et al., 2016***), which arises naturally in a two dimensional CAN model. In our model, we would obtain slices through a hexagonal grid if the rat learns the output pattern in two dimensions and afterwards is constrained to move on a linear track that is part of the same arena. If the rat learns the firing pattern on the linear track from scratch, the firing fields would be periodic.

Rapid appearance and rearrangement of grids

Models that learn grid cells from spatially tuned input do not have to assume a preexisting connectivity pattern or specific mechanisms for path integration (***Burgess et al., 2007***), but are challenged by the fast emergence of hexagonal firing patterns in unfamiliar environments (***Hafting et al., 2005***). Most plasticity-based models require slow learning, such that the animal explores the whole arena before significant synaptic changes occur. Therefore, grid patterns typically emerge slower than experimentally observed (***Dordek et al., 2016***). This delay is particularly pronounced in models that require an extensive exploration of both space and movement direction (***Kropff and Treves, 2008***; ***Franzius et al., 2007a***; ***D'Albis and Kempter, 2017***). In contrast to these models, which give center stage to the temporal statistics of the animal's movement, our approach relies purely on the spatial statistics of the input and is hence insensitive to running speed.

For the mechanism we suggested, the self-organization was very robust and allowed a rapid pattern formation on short time scales, similar to those observed in rodents (***Figure 3***). This speed could be further increased by accelerated reactivation of previous experiences during periods of rest (***Lee and Wilson, 2002***). By this means, the exploration time and the time it takes to activate all input patterns could be decoupled, leading to a much faster emergence of grid cells in all trajectory-independent models with associative learning. Other models that explain the emergence of grid patterns from place cell input through synaptic depression and potentiation also develop grid cells in realistic times (***Castro and Aguiar, 2014***; ***Stepanyuk, 2015***; ***Monsalve-Mercado and Leibold, 2017***). These models differ from ours in that they do not require inhibition, but instead specific forms of rate dependent synaptic depression and potentiation that change the synaptic weights such that place cell-like input leads to grid cell-like output. How these models generalize to potentially non-localized input is yet to be shown.

Learning the required connectivity in CAN models can take a long time (***Widloski and Fiete, 2014***). However, as soon as the required connectivity and translation mechanism is established, a grid pattern would be observed immediately, even in a novel room. For different rooms this pattern could have different phases and orientations, but a similar grid spacing (***Fyhn et al., 2007***). Similarly, we found that room switches in our model lead to grid patterns of the same grid spacing but different phases and orientations. The pattern emerges rapidly, but is not instantaneously present (***Figure 3–Figure supplement 3***). It would be interesting to study if a rotation of a fraction of the input would lead to a bimodal distribution of grid rotations: No rotation and co-rotation with the rotated input, as recently observed in experiments where distal cues were rotated but proximal cues stayed fixed (***Savelli et al., 2017***).

Recently, it was discovered that in an arena separated by a wall, single grid cells form two independent grid patterns – one on each side – that coalesce once the wall is removed (***Wernle et al., 2018***; ***Rosay et al., 2018***). This coalescence is local, i.e., grid fields close to the partition wall readjust,

whereas grid fields far away do not change their locations. Feedforward models like ours can explain such a local rearrangement (***Figure 4***; ***Rosay et al., 2018***).

Boundary effects

Experiments show that the pattern and the orientation of grid cells is influenced by the geometry of the environment. In a quadratic arena, the orientation of grid cells tends to align – with a small offset – to one of the box axes (***Stensola et al., 2015***). In trapezoidal arenas, the hexagonality of grids is distorted (***Krupic et al., 2015a***). We considered quadratic and circular arenas with rat trajectories from behavioral experiments and found that the boundaries distort the grid pattern also in our simulations, particularly for localized inputs (***Figure 2–Figure supplement 2***). In trapezoidal geometries, we expect this to lead to non-hexagonal grids. However, we did not observe a pronounced alignment to quadratic boundaries if the input place fields are randomly located (***Figure 2–Figure supplement 2***).

Conclusion

We found that interacting excitatory and inhibitory plasticity serves as a simple and robust mechanism for rapid self-organization of stable and symmetric patterns from spatially modulated feedforward input. The suggested mechanism ports the robust pattern formation of attractor models from the neural to the spatial domain and increases the speed of self-organization of plasticity-based mechanisms to time scales on which the spatial tuning of neurons is typically measured. It will be interesting to explore how recurrent connections between output cells can help to understand the role of local inhibitory (***Couey et al., 2013***; ***Pastoll et al., 2013***) and excitatory connections (***Winterer et al., 2017***) and the presence or absence of topographic arrangements of spatially tuned cells (***O'Keefe et al., 1998***; ***Stensola et al., 2012***; ***Giocomo et al., 2014***). We illustrated the properties and requirements of the model in the realm of spatial representations. Since invariance and selectivity are ubiquitous properties of receptive fields in the brain, the interaction of excitatory and inhibitory synaptic plasticity might be essential to form stable representations from sensory input also in other brain areas (***Constantinescu et al., 2016***; ***Clopath et al., 2016***).

Methods and Materials

Code availability

The code for reproducing the essential findings of this article is available on https://github.com/sim-web/spatial_patterns (***Weber, 2018***) under the GNU General Public License v3.0.

Network architecture and neuron model

We study a feedforward network where a single output neuron receives synaptic input from N_E excitatory and N_I inhibitory neurons (***Figure 1a***) with synaptic weight vectors $\mathbf{w}^\mathrm{E} \in \mathbb{R}^{N_\mathrm{E}}$, $\mathbf{w}^\mathrm{I} \in \mathbb{R}^{N_\mathrm{I}}$ and spatially tuned input rates $\mathbf{r}^\mathrm{E}(\mathbf{x}) \in \mathbb{R}^{N_\mathrm{E}}$, $\mathbf{r}^\mathrm{I}(\mathbf{x}) \in \mathbb{R}^{N_\mathrm{I}}$, respectively. Here $\mathbf{x} \in \mathbb{R}^{\text{dimensions}}$ denotes the location and later also the head direction of the animal. For simplicity and to allow a mathematical analysis we use a rate-based description for all neurons. The firing rate of the output neuron is given by the rectified sum of weighted excitatory and inhibitory inputs:

$$r^\text{out}(\mathbf{x}(t)) = \left[\sum_{i=1}^{N_\mathrm{E}} w_i^\mathrm{E}(t) r_i^\mathrm{E}(\mathbf{x}(t)) - \sum_{j=1}^{N_\mathrm{I}} w_j^\mathrm{I}(t) r_j^\mathrm{I}(\mathbf{x}(t)) \right]_+ , \tag{1}$$

where $[\cdot]_+$ denotes a rectification that sets negative firing rates to zero. To comply with the notion of excitation and inhibition, all weights are constrained to be positive. In most simulations we use four times as many excitatory as inhibitory input neurons. ***Tables 3*** to ***5*** list values for all simulation parameters used in each figure.

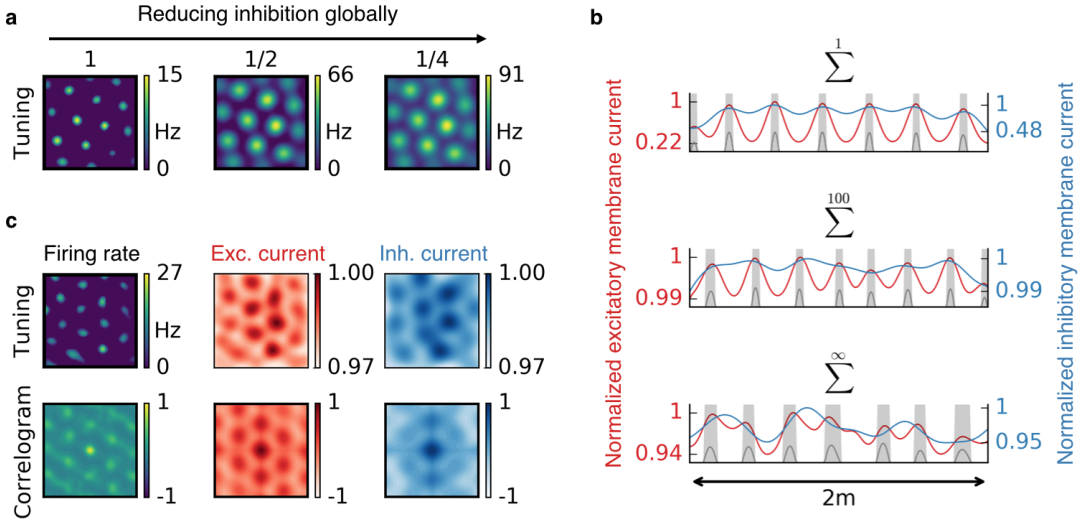

Figure 7. The effect of reduced inhibition on grid cell properties. **a**) Reducing the strength of inhibitory synapses to a fraction of its initial value (from left to right: 1, 1/2, 1/4) leads to larger grid fields but an unchanged grid spacing in our model. In continuous attractor network models, the same reduction of inhibition would affect not only the field size but also the grid spacing. **b**) Excitatory (red) and inhibitory (blue) membrane current to a cell with grid like firing pattern (gray) on a linear track. The currents are normalized to a maximum value of 1. Different rows correspond to different spatial tuning characteristics of the input neurons. From top to bottom: Place cell-like tuning, sparse non-localized tuning (sum of 100 randomly located place fields), dense non-localized tuning (Gaussian random fields). Peaks in excitatory membrane current are co-localized with grid fields (shaded area) for all input statistics. In contrast, the inhibitory membrane current is not necessarily correlated with the grid fields for non-localized input. Moreover, the dynamic range of the membrane currents is reduced for non-localized input. A reduction of inhibition as shown in ***a*** corresponds to a lowering of the blue curve. **c**) Excitatory and inhibitory membrane current to a grid cell receiving sparse non-localized input (sum of 100 randomly located place fields) in two dimensions. Top: Tuning of output firing rate, normalized excitatory and inhibitory membrane current. Bottom: Autocorrelograms thereof. The grid pattern is more apparent in the spatial tuning of the excitatory membrane current than in the inhibitory membrane current.

Excitatory and inhibitory plasticity

In each unit time step ($\Delta t = 1$), the excitatory weights are updated according to a Hebbian rule:

$$\Delta \mathbf{w}^{\mathrm{E}} = \eta_{\mathrm{E}} \mathbf{r}^{\mathrm{E}}(\mathbf{x}) r^{\mathrm{out}}(\mathbf{x}) \quad \text{(and normalization)}\,. \tag{2}$$

The excitatory learning rate η_{E} is a constant that we chose individually for each simulation. To avoid unbounded weight growth, we use a quadratic multiplicative normalization, i.e., we keep the sum of the squared weights of the excitatory population $\sum_{i=1}^{N_{\mathrm{E}}} (w_i^{\mathrm{E}})^2$ constant at its initial value, by rescaling the weights after each unit time step. However, synaptic weight normalization is not a necessary ingredient for the emergence of firing patterns (***Figure 2–Figure supplement 4***). We model inhibitory synaptic plasticity using a previously suggested learning rule (***Vogels et al., 2011***):

$$\Delta \mathbf{w}^{\mathrm{I}} = \eta_{\mathrm{I}} \mathbf{r}^{\mathrm{I}}(\mathbf{x}) (r^{\mathrm{out}}(\mathbf{x}) - \rho_0)\,, \tag{3}$$

with inhibitory learning rate η_{I} and target rate $\rho_0 = 1\,\mathrm{Hz}$. Negative inhibitory weights are set to zero.

Rat trajectory

In the linear track model (one dimension, ***Figures 1*** and ***7***), we create artificial run-and-tumble trajectories $x(t)$ constrained on a line of length L with constant velocity $v = 1\,\mathrm{cm}$ per unit time step and persistence length $L/2$ (***Appendix 1***).

In the open arena model (two dimensions, ***Figures 2***, ***3***, ***5*** and ***7***), we use trajectories $\mathbf{x}(t)$ from behavioral data (***Sargolini et al., 2006b***) of a rat that moved in a $1\,\mathrm{m} \times 1\,\mathrm{m}$ quadratic enclosure (***Appendix 1***). In the simulations with a separation wall (***Figure 4***), we created trajectories as a two

dimensional persistent random walk (***Appendix 1***).
In the model for neurons with head direction tuning (three dimensions, ***Figure 6***), we use the same behavioral trajectories as in two dimensions and model the head direction as noisily aligned to the direction of motion (***Appendix 1***).

Spatially tuned inputs

The firing rates of excitatory and inhibitory synaptic inputs $r_i^{\mathrm{E}}, r_j^{\mathrm{I}}$ are tuned to the location $\mathbf{x}$ of the animal. In the following, we use x and y for the first and second spatial dimension and z for the head direction.

For place field-like input, we use Gaussian tuning functions with standard deviation σ_{E}, σ_{I} for the excitatory and inhibitory population, respectively. In ***Figure 5*** the standard deviation is chosen independently along the x and y direction. The centers of the Gaussians are drawn randomly from a distorted lattice (***Figure 2–Figure supplement 3***). This way we ensure random but spatially dense tuning. The lattice contains locations outside the box to reduce boundary effects.

For sparse non-localized input with $N_{\mathrm{P}}^{\mathrm{f}}$ fields per neuron of population P, we first create $N_{\mathrm{P}}^{\mathrm{f}}$ distorted lattices, each with N_{P} locations. We then assign $N_{\mathrm{P}}^{\mathrm{f}}$ of the resulting $N_{\mathrm{P}}^{\mathrm{f}} N_{\mathrm{P}}$ locations at random and without replacement to each input neuron (see also ***Appendix 1***).

For dense non-localized input, we convolve Gaussians with white noise and increase the resulting signal to noise ratio by setting the minimum to zero and the mean to 0.5 (***Appendix 1***). The Gaussian convolution kernels have different standard deviations for different populations. For each input neuron we use a different realization of white noise. This results in arbitrary tuning functions of the same autocorrelation length as the – potentially asymmetric – Gaussian convolution kernel. For grid cell-like input, we place Gaussians of standard deviation σ_{E} on the nodes of perfect hexagonal grids whose spacing and orientation is variable. In ***Figure 5b*** we drew the grid spacing of each input from a normal distribution of mean $6\sigma_{\mathrm{E}}$ and standard deviation $\sigma_{\mathrm{E}}/6$. The grid orientation was drawn from a uniform distribution between -30 and 30 degrees.

For input with combined spatial and head direction tuning, we use the Gaussian tuning curves described above for the spatial tuning and von Mises distributions along the head direction dimension (***Appendix 1***).

For all input tunings, the standard deviation of the firing rate is of the same order of magnitude as the mean firing rate (***Appendix 1***).

Initial synaptic weights and global reduction of inhibition

We specify a mean for the initial excitatory and inhibitory weights, respectively, and randomly draw each synaptic weight from the corresponding mean $\pm 5\%$. The excitatory mean is chosen such that the output neuron would fire above the target rate everywhere in the absence of inhibition; we typically take this mean to be 1 (***Table 4*** and ***Appendix 1***). The mean inhibitory weight is then determined such that the output neuron would fire close to the target rate, if all the weights were at their mean value (***Table 5*** and ***Appendix 1***). Choosing the weights this way ensures that initial firing rates are random, but neither zero everywhere, nor inappropriately high.
We modeled a global reduction of inhibition by scaling all inhibitory weights by a constant factor, after the grid has been learned.

Mathematical analysis of the learning rules

In the following, we derive the spacing of periodic firing patterns as a function of the simulation parameters for the linear track.

We first show that homogeneous weights, chosen such that the output neuron fires at the target rate, are a fixed point for the time evolution of excitatory and inhibitory weights under the assumption of slow learning. We then perturb this fixed point and study the time evolution of the perturbations in Fourier space. The translational invariance of the input overlap leads to a decoupling of spatial frequencies and leaves a two dimensional dynamical system for each spatial

frequency. For smoother spatial tuning of inhibitory input than excitatory input, the eigenvalue spectrum of the dynamical system has a unique maximum, which indicates the most unstable spatial frequency. This frequency accurately predicts the grid spacing. We first consider place cell-like input (Gaussians) and then non-localized input (Gaussians convolved with white noise).

At the end of the analysis you find a glossary of the notation. Whenever we use P as a sub- or superscript instead of E or I, this implies that the equation holds for neurons of the excitatory and the inhibitory population.

The analysis is written as a detailed and comprehensible walk-through. The reader who is only interested in the result can jump to ***Equations 78*** and ***104***.

Assumption of slow learning

The firing rate of the output neuron is the weighted sum of excitatory and inhibitory input rates:

$$r^{\text{out}} = \left[\mathbf{w}^{\text{E}}\mathbf{r}^{\text{E}} - \mathbf{w}^{\text{I}}\mathbf{r}^{\text{I}}\right]_{+} . \tag{4}$$

where $[\ldots]_{+}$ indicates that negative firing rates are set to zero.
Written as a differential equation, the excitatory learning rule with quadratic multiplicative normalization is given by:

$$\frac{\mathrm{d}\mathbf{w}^{\text{E}}}{\mathrm{d}t} = \eta_{\text{E}} \left(\mathbb{1} - \frac{\mathbf{w}^{\text{E}}\mathbf{w}^{\text{E}T}}{\|\mathbf{w}^{\text{E}}\|^2} \right) \mathbf{r}^{\text{E}} r^{\text{out}} , \tag{5}$$

where $\mathbb{1}$ is the $N_{\text{E}} \times N_{\text{E}}$ identity matrix. The projection operator $\frac{\mathbf{w}^{\text{E}}\mathbf{w}^{\text{E}T}}{\|\mathbf{w}^{\text{E}}\|^2}$ ensures that the weights are constrained to remain on the hypersphere whose radius is defined by the initial value of the sum over the squares of all excitatory weights (***Miller and MacKay, 1994***). The inhibitory learning rule is given by:

$$\frac{\mathrm{d}\mathbf{w}^{\text{I}}}{\mathrm{d}t} = \eta_{\text{I}} \mathbf{r}^{\text{I}} \left(r^{\text{out}} - \rho_0 \right) . \tag{6}$$

We assume the rat to learn slowly, such that it forages through the environment before significant learning (i.e., weight change) occurs. Therefore we can coarsen the time scale and rewrite ***Equations 5*** and ***6*** as

$$\frac{\mathrm{d}\mathbf{w}^{\text{E}}}{\mathrm{d}t} = \eta_{\text{E}} \left\langle \left(\mathbb{1} - \frac{\mathbf{w}^{\text{E}}\mathbf{w}^{\text{E}T}}{\|\mathbf{w}^{\text{E}}\|^2} \right) \mathbf{r}^{\text{E}} r^{\text{out}} \right\rangle_x \tag{7}$$

and

$$\frac{\mathrm{d}\mathbf{w}^{\text{I}}}{\mathrm{d}t} = \eta_{\text{I}} \left\langle \mathbf{r}^{\text{I}} \left(r^{\text{out}} - \rho_0 \right) \right\rangle_x , \tag{8}$$

respectively, where the spatial average, $\langle \ldots \rangle_x$, is defined as

$$\langle (\ldots) \rangle_x = \frac{1}{L} \int_{-L/2}^{+L/2} (\ldots)\, \mathrm{d}x \tag{9}$$

and L is the length of the linear track.

High density assumption and continuum limit for place cell-like input

We assume a high density of input neurons and formulate the system in continuous variables. More precisely, we assume the distance between two neighboring firing fields to be much smaller than the width of the firing fields, i.e., $L/N_{\text{P}} \ll \sigma_{\text{P}}$. Furthermore, we assume that the linear track is very long compared to the width of the firing fields, i.e., $\sigma_{\text{P}} \ll L$.
We replace the neuron index with the continuous variable μ and denote the weight w^{P}_{μ} and the tuning function $r^{\text{P}}(\mu, x)$ associated with a place field that is centered at μ in the continuum limit as:

$$w^{\text{P}}_i \to w^{\text{P}}(\mu) \;\text{ and }\; r^{\text{P}}_i(x) \to r^{\text{P}}(\mu, x) . \tag{10}$$

The distance between two neighboring place fields is given by $\Delta\mu = L/N_\mathrm{P}$. For sums over all neurons we thus get the following integral in the continuum limit:

$$\sum_{i=1}^{N_\mathrm{P}} f_i = \frac{1}{\Delta\mu}\sum_{i=1}^{N_\mathrm{P}} f_i\,\Delta\mu \to \frac{N_\mathrm{P}}{L}\int_{-L/2}^{+L/2} f(\mu)\,\mathrm{d}\mu\,. \tag{11}$$

In the following we will switch between the discrete and the continuous formulation and use whatever is more convenient.

For place cell-like input we take Gaussian tuning curves:

$$r_i^\mathrm{P}(x) = \alpha_\mathrm{P}\exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_\mathrm{P}^2}\right\}, \tag{12}$$

with height α_P and standard deviation σ_P. In the continuum limit we thus get:

$$r_i^\mathrm{P}(x) \to r^\mathrm{P}(\mu,x) = r^\mathrm{P}(|x-\mu|) = \alpha_\mathrm{P}\exp\left\{-\frac{(x-\mu)^2}{2\sigma_\mathrm{P}^2}\right\}. \tag{13}$$

Because of the translational invariance of $r^\mathrm{P}(\mu,x)$, integration over space gives the same result as integration over all center locations and the mean of all inputs is the same:

$$\left\langle r_i^\mathrm{P}(x)\right\rangle_x = \left\langle r^\mathrm{P}(\mu,x)\right\rangle_x \tag{14}$$

$$= \frac{1}{L}\int_{-L/2}^{+L/2} r^\mathrm{P}(\mu,x)\,\mathrm{d}x \tag{15}$$

$$= \frac{1}{L}\int_{-L/2}^{+L/2} r^\mathrm{P}(\mu,x)\,\mathrm{d}\mu \approx \frac{\alpha_\mathrm{P}}{L}\sqrt{2\pi\sigma_\mathrm{P}^2} = M_\mathrm{P}/L \tag{16}$$

where we introduced $M_\mathrm{P} := \alpha_\mathrm{P}\sqrt{2\pi\sigma_\mathrm{P}^2}$ for the area under the tuning curves. Accordingly, we get a summarized input activity that is independent of location:

$$\sum_{i=1}^{N_\mathrm{P}} r_i^\mathrm{P}(x) = \frac{N_\mathrm{P}}{L}\int_{-L/2}^{+L/2} r^\mathrm{P}(\mu,x)\,\mathrm{d}\mu \approx \frac{N_\mathrm{P}}{L}M_\mathrm{P}\,. \tag{17}$$

Equal weights form a fixed point

In the following, we will show that equal weights $w^\mathrm{E}(\mu) = w_0^\mathrm{E}$ and $w^\mathrm{I}(\mu') = w_0^\mathrm{I}$, $\forall\mu,\mu'$ form a fixed point if w_0^I is chosen such that the output neuron fires at the target rate, ρ_0, throughout the arena. With equal weights we get a constant firing rate r_0^out,

$$r^\mathrm{out}(x) = r_0^\mathrm{out} = \left[w_0^\mathrm{E}\sum_i r_i^\mathrm{E}(x) - w_0^\mathrm{I}\sum_i r_i^\mathrm{I}(x)\right]_+, \tag{18}$$

which according to ***Equation 17*** does not depend on x. Furthermore, according to ***Equation 14***, $\left\langle r_i^\mathrm{P}(x)\right\rangle_x$ does not depend on the neuron index i. Now the stationarity of the excitatory weight evolution follows from ***Equation 7***:

$$\frac{\mathrm{d}w_i^\mathrm{E}}{\mathrm{d}t} = \eta_\mathrm{E}\left\langle r^\mathrm{out}\sum_j r_j^\mathrm{E}\left(\delta_{ij} - \frac{w_i^\mathrm{E}w_j^\mathrm{E}}{\sum_k w_k^{\mathrm{E}2}}\right)\right\rangle_x \tag{19}$$

$$= \eta_\mathrm{E} r_0^\mathrm{out}\sum_j\left[\left\langle r_j^\mathrm{E}\right\rangle_x\left(\delta_{ij} - \frac{w_0^{\mathrm{E}2}}{N_\mathrm{E}w_0^{\mathrm{E}2}}\right)\right] \tag{20}$$

$$= \frac{r_0^\mathrm{out}\eta_\mathrm{E}M_\mathrm{E}}{L}\sum_{j=1}^{N_\mathrm{E}}\left(\delta_{ij} - \frac{1}{N_\mathrm{E}}\right) = 0\,, \tag{21}$$

i.e., excitatory weights are stationary for all values of w_0^E and w_0^I (here δ_{ij} denotes the Kronecker delta which is 1 if $i = j$ and 0 otherwise). This holds for all input functions for which $\left\langle r_j^\mathrm{E}(x)\right\rangle_x$ is

independent of j. If $r^{\text{out}} = \rho_0$ it immediately follows from ***Equation 6*** that $\frac{\mathrm{d}w^{\mathrm{I}}}{\mathrm{d}t} = 0$, so the inhibitory weights are stationary if

$$\rho_0 = \mathbf{w}^{\mathrm{E}}\mathbf{r}^{\mathrm{E}} - \mathbf{w}^{\mathrm{I}}\mathbf{r}^{\mathrm{I}} = w_0^{\mathrm{E}} \sum_i r_i^{\mathrm{E}} - w_0^{\mathrm{I}} \sum_i r_i^{\mathrm{I}}\,, \tag{22}$$

which is fulfilled if

$$w_0^{\mathrm{I}} = \frac{w_0^{\mathrm{E}} \sum_i r_i^{\mathrm{E}} - \rho_0}{\sum_i r_i^{\mathrm{I}}} = \frac{w_0^{\mathrm{E}} N_{\mathrm{E}} M_{\mathrm{E}} - \rho_0}{N_{\mathrm{I}} M_{\mathrm{I}}}\,. \tag{23}$$

Linear stability analysis

In the following, we will show that the fixed point of equal weights, the *homogeneous steady state*, is unstable, when the spatial tuning of inhibitory inputs is broader than that of the excitatory inputs. In this case, perturbations of the fixed point will grow and a particular spatial frequency will grow fastest. We will show that this spatial frequency predicts the spacing of the resulting periodic pattern (***Figure 1g***).

We disturb the fixed point

$$w^{\mathrm{E}}(\mu) = w_0^{\mathrm{E}} + \delta w^{\mathrm{E}}(\mu), \quad w^{\mathrm{I}}(\mu) = w_0^{\mathrm{I}} + \delta w^{\mathrm{I}}(\mu) \tag{24}$$

and look at the time evolution of the perturbations $\frac{\mathrm{d}\delta w^{\mathrm{E}}}{\mathrm{d}t}$ and $\frac{\mathrm{d}\delta w^{\mathrm{I}}}{\mathrm{d}t}$ of the excitatory and inhibitory weights around the fixed point.

Close to the fixed point the output neuron fires around the target rate ρ_0. We thus ignore the rectification in ***Equation 4***, i.e., $r^{\text{out}} = \rho_0 + \delta r^{\text{out}}$, with $\delta r^{\text{out}} = \sum_k \delta w_k^{\mathrm{E}} r_k^{\mathrm{E}} - \sum_{k'} \delta w_{k'}^{\mathrm{I}} r_{k'}^{\mathrm{I}}$.

Time evolution of perturbations of the inhibitory weights

We start with the time evolution of the inhibitory weight perturbations:

$$\begin{aligned}
\frac{\mathrm{d}\delta w_i^{\mathrm{I}}}{\mathrm{d}t} = \frac{\mathrm{d}w_i^{\mathrm{I}}}{\mathrm{d}t} &= \eta_{\mathrm{I}} \left\langle \left(r^{\text{out}} - \rho_0\right) r_i^{\mathrm{I}} \right\rangle_x && (25)\\
&= \eta_{\mathrm{I}} \left\langle \left(\rho_0 + \delta r^{\text{out}} - \rho_0\right) r_i^{\mathrm{I}} \right\rangle_x && (26)\\
&= \eta_{\mathrm{I}} \left\langle r_i^{\mathrm{I}} \delta r^{\text{out}} \right\rangle_x && (27)\\
&= \eta_{\mathrm{I}} \left\langle r_i^{\mathrm{I}} \left(\sum_k \delta w_k^{\mathrm{E}} r_k^{\mathrm{E}} - \sum_{k'} \delta w_{k'}^{\mathrm{I}} r_{k'}^{\mathrm{I}} \right) \right\rangle_x && (28)\\
&= \eta_{\mathrm{I}} \left(\sum_{k=1}^{N_{\mathrm{E}}} \left\langle r_i^{\mathrm{I}} r_k^{\mathrm{E}} \right\rangle_x \delta w_k^{\mathrm{E}} - \sum_{k'=1}^{N_{\mathrm{I}}} \left\langle r_i^{\mathrm{I}} r_{k'}^{\mathrm{I}} \right\rangle_x \delta w_{k'}^{\mathrm{I}} \right), && (29)
\end{aligned}$$

where we used that only the rates $\mathbf{r}^{\mathrm{P}}$ depend on x. Intuitively, the first term in ***Equation 29*** means that the rate of change of the inhibitory weight perturbation of the weight associated to one location depends on the excitatory perturbations of the weights associated to every other location, weighted with the overlap (the cross correlation) of the two associated tuning functions (analogous for inhibitory weight perturbations). In the continuum limit, the sums are:

$$\begin{aligned}
\eta^{\mathrm{P}} \sum_{k=1}^{N_{\mathrm{P}'}} \left\langle r_i^{\mathrm{P}} r_k^{\mathrm{P}'} \right\rangle_x \delta w^{\mathrm{P}'}{}_k &\to \eta^{\mathrm{P}} \frac{N_{\mathrm{P}'}}{L} \int_{-L/2}^{+L/2} \left\langle r^{\mathrm{P}}(\mu) r^{\mathrm{P}'}(\mu') \right\rangle_x \delta w^{\mathrm{P}'}(\mu')\,\mathrm{d}\mu' && (30)\\
&= \int_{-L/2}^{+L/2} K^{\mathrm{PP}'}(\mu,\mu') \delta w^{\mathrm{P}'}(\mu')\,\mathrm{d}\mu'\,, && (31)
\end{aligned}$$

where we introduced overlap kernels

$$K^{\mathrm{PP}'}(\mu,\mu') := \eta^{\mathrm{P}} \frac{N_{\mathrm{P}'}}{L} \left\langle r^{\mathrm{P}}(\mu) r^{\mathrm{P}'}(\mu') \right\rangle_x \quad \mathrm{P},\mathrm{P}' \in \{\mathrm{E},\mathrm{I}\}\,. \tag{32}$$

The overlap $\left\langle r^{\mathrm{P}}(\mu) r^{\mathrm{P}'}(\mu') \right\rangle_x$ only depends on the distance of the Gaussian fields, i.e.,

$$K^{\mathrm{PP}'}(\mu,\mu') = K^{\mathrm{PP}'}(\mu - \mu')\,. \tag{33}$$

Taking $L \to \infty$, the time evolution of the perturbations of the inhibitory weights can thus be written as convolutions:

$$\frac{\mathrm{d}\delta w^{\mathrm{I}}(\mu)}{\mathrm{d}t} = (K^{\mathrm{IE}} * \delta w^{\mathrm{E}})(\mu) - (K^{\mathrm{II}} * \delta w^{\mathrm{I}})(\mu)\,, \tag{34}$$

where $*$ denotes a convolution.

Time evolution of perturbations of the excitatory weights

To derive the time evolution of the excitatory weights, we first show that the weight normalization term in ***Equation 7***, expressed through the projection operator $P_{ij} = \frac{w_i w_j}{\sum_k w_k^2}$, leads to a term that balances homogeneous weight perturbations and a term that can be neglected in the continuum limit.

Let P be the projection operator responsible for the normalization of the excitatory weights by projecting a weight update onto a vector that is orthogonal to the hypersphere of constant $\sum_{i=1}^{N_{\mathrm{E}}}(w_i^{\mathrm{E}})^2$. We now determine the projection operator around the fixed point (We drop the index 'E' in the following, to improve readability):

$$P_{ij} = \frac{(w_0 + \delta w_i)(w_0 + \delta w_j)}{\sum_k (w_0 + \delta w_k)^2} \equiv P_{ij}(\mathbf{w} + \delta\mathbf{w})\,. \tag{35}$$

Using Taylor's theorem

$$P_{ij}(\mathbf{w} + \delta\mathbf{w}) = P_{ij}(\mathbf{w}) + \sum_{l=1}^{N} \delta w_l \frac{\mathrm{d}P_{ij}(\mathbf{w})}{\mathrm{d}w_l} + \mathcal{O}(\delta\mathbf{w}^2) \tag{36}$$

and $w_l = w_0\ \forall\ l$ we get

$$P_{ij}(\mathbf{w}) = \frac{w_i w_j}{\sum_k w_k^2} = 1/N\,, \tag{37}$$

$$\frac{\mathrm{d}P_{ij}(\mathbf{w})}{\mathrm{d}w_l} = \frac{\delta_{il} w_j}{\sum_k w_k^2} + \frac{\delta_{jl} w_i}{\sum_k w_k^2} - \frac{w_i w_j 2 w_l}{(\sum_k w_k^2)^2} = \frac{\delta_{il}}{N w_0} + \frac{\delta_{jl}}{N w_0} - \frac{2}{N^2 w_0}\,. \tag{38}$$

In summary this gives:

$$P_{ij} = \underbrace{\frac{1}{N_{\mathrm{E}}}}_{\equiv P_0 \propto \mathcal{O}(1)} + \underbrace{\frac{1}{N_{\mathrm{E}} w_0^{\mathrm{E}}}\left(\delta w_i^{\mathrm{E}} + \delta w_j^{\mathrm{E}} - \frac{2\sum_{l=1}^{N_{\mathrm{E}}} \delta w_l^{\mathrm{E}}}{N_{\mathrm{E}}}\right)}_{\equiv \delta P_{ij} \propto \mathcal{O}(\delta\mathbf{w})} + \mathcal{O}(\delta\mathbf{w}^2)\,. \tag{39}$$

Using the perturbed projection operator ***Equation 39*** with ***Equation 7*** we obtain the time evolution of the excitatory weight perturbation to linear order:

$$\frac{\mathrm{d}\delta w_i^{\mathrm{E}}}{\mathrm{d}t} = \frac{\mathrm{d}w_i^{\mathrm{E}}}{\mathrm{d}t} \tag{40}$$

$$= \eta_{\mathrm{E}} \left\langle r^{\mathrm{out}} \sum_j (\delta_{ij} - P_{ij}) r_j^{\mathrm{E}} \right\rangle_x \tag{41}$$

$$= \eta_{\mathrm{E}} \left\langle (\rho_0 + \delta r^{\mathrm{out}}) \sum_j (\delta_{ij} - P_0 - \delta P_{ij}) r_j^{\mathrm{E}} \right\rangle_x \tag{42}$$

$$= \eta_{\mathrm{E}} \underbrace{\left\langle \rho_0 \sum_j (\delta_{ij} - P_0) r_j^{\mathrm{E}} \right\rangle_x}_{=0, \text{cf. } \textit{Equation 19}} + \left\langle \delta r^{\mathrm{out}} \sum_j (\delta_{ij} - P_0) r_j^{\mathrm{E}} \right\rangle_x - \left\langle \rho_0 \sum_j \delta P_{ij} r_j^{\mathrm{E}} \right\rangle_x + \mathcal{O}(\delta\mathbf{w}^2) \tag{43}$$

$$= \eta_{\mathrm{E}} \left(\underbrace{\left\langle r_i^{\mathrm{E}} \delta r^{\mathrm{out}} \right\rangle_x}_{(1)} - \underbrace{P_0 \left\langle \delta r^{\mathrm{out}} \sum_j r_j^{\mathrm{E}} \right\rangle_x}_{(2)} - \underbrace{\rho_0 \left\langle \sum_j \delta P_{ij} r_j^{\mathrm{E}} \right\rangle_x}_{(3)} \right) + \mathcal{O}(\delta\mathbf{w}^2) \tag{44}$$

Term (1) in ***Equation 44*** has a similar structure as in the inhibitory case (***Equation 27***) and will lead to analogous convolutions. In the continuum limit the second term is given by

$$(2) = \frac{1}{N_\mathrm{E}} \left\langle \left(\sum_k r_k^\mathrm{E} \delta w_k^\mathrm{E} - \sum_{k'} r_{k'}^\mathrm{I} \delta w_{k'}^\mathrm{I} \right) \sum_j r_j^\mathrm{E} \right\rangle_x \tag{45}$$

$$= \frac{M_\mathrm{E}}{L} \left\langle \sum_k r_k^\mathrm{E} \delta w_k^\mathrm{E} - \sum_{k'} r_{k'}^\mathrm{I} \delta w_{k'}^\mathrm{I} \right\rangle_x \tag{46}$$

$$= \frac{M_\mathrm{E}}{L} \left(\sum_k \left\langle r_k^\mathrm{E} \right\rangle_x \delta w_k^\mathrm{E} - \sum_{k'} \left\langle r_{k'}^\mathrm{I} \right\rangle_x \delta w_{k'}^\mathrm{I} \right) \tag{47}$$

$$= \frac{M_\mathrm{E}}{L^2} \left(M_\mathrm{E} \sum_k \delta w_k^\mathrm{E} - M_\mathrm{I} \sum_{k'} \delta w_{k'}^\mathrm{I} \right) \tag{48}$$

$$\text{cont. limit} \to \frac{M_\mathrm{E}}{L^3} \left(N_\mathrm{E} M_\mathrm{E} \int_{-L/2}^{+L/2} \delta w^\mathrm{E}(\mu') \, \mathrm{d}\mu' - N_\mathrm{I} M_\mathrm{I} \int_{-L/2}^{+L/2} \delta w^\mathrm{I}(\mu'') \, \mathrm{d}\mu'' \right) \tag{49}$$

and the third term by

$$(3) = \frac{\rho_0}{N_\mathrm{E} w_0^\mathrm{E}} \left\langle \sum_j r_j^\mathrm{E} \left(\delta w_i^\mathrm{E} + \delta w_j^\mathrm{E} - \frac{2}{N_\mathrm{E}} \sum_l \delta w_l^\mathrm{E} \right) \right\rangle_x \tag{50}$$

$$= \frac{\rho_0}{N_\mathrm{E} w_0^\mathrm{E}} \sum_j \left\langle r_j^\mathrm{E} \right\rangle_x \left(\delta w_i^\mathrm{E} + \delta w_j^\mathrm{E} - \frac{2}{N_\mathrm{E}} \sum_l \delta w_l^\mathrm{E} \right) \tag{51}$$

$$= \frac{\rho_0 M_\mathrm{E}}{N_\mathrm{E} w_0^\mathrm{E} L} \sum_j \left(\delta w_i^\mathrm{E} + \delta w_j^\mathrm{E} - \frac{2}{N_\mathrm{E}} \sum_l \delta w_l^\mathrm{E} \right) \tag{52}$$

$$= \frac{\rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \left(\delta w_i^\mathrm{E} + \frac{1}{N_\mathrm{E}} \sum_j \delta w_j^\mathrm{E} - \frac{2}{N_\mathrm{E}} \sum_l \delta w_l^\mathrm{E} \right) \tag{53}$$

$$= \frac{\rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \left(\delta w_i^\mathrm{E} - \frac{1}{N_\mathrm{E}} \sum_j \delta w_j^\mathrm{E} \right) \tag{54}$$

$$\text{cont. limit} \to \frac{\rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \left(\delta w^\mathrm{E}(\mu) - \frac{1}{L} \int_{-L/2}^{+L/2} \delta w^\mathrm{E}(\mu') \, \mathrm{d}\mu' \right) \tag{55}$$

$$= \frac{\rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \int_{-L/2}^{+L/2} \mathrm{d}\mu' \delta w^\mathrm{E}(\mu') \left[\delta(\mu - \mu') - \frac{1}{L} \right], \tag{56}$$

where $\delta(\mu - \mu')$ denotes the Dirac delta function. Together this leads to the time evolution of the excitatory weight perturbations:

$$\frac{\mathrm{d}\delta w^\mathrm{E}(\mu)}{\mathrm{d}t} = \int_{-L/2}^{+L/2} \mathrm{d}\mu' \delta w^\mathrm{E}(\mu') \left[K^\mathrm{EE}(\mu - \mu') - \frac{\eta_\mathrm{E} \rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \delta(\mu - \mu') \right. \tag{57}$$

$$\left. + \frac{\eta_\mathrm{E} M_\mathrm{E}}{L^2} \left(\frac{\rho_0}{w_0^\mathrm{E}} - \frac{N_\mathrm{E} M_\mathrm{E}}{L} \right) \right] \tag{58}$$

$$- \int_{-L/2}^{+L/2} \mathrm{d}\mu'' \delta w^\mathrm{I}(\mu'') \left[K^\mathrm{EI}(\mu - \mu'') - \frac{\eta_\mathrm{E} N_\mathrm{I} M_\mathrm{E} M_\mathrm{I}}{L^3} \right]. \tag{59}$$

We now take $L \to \infty$ and write everything as convolutions, also trivial ones:

$$\frac{\mathrm{d}\delta w^\mathrm{E}(\mu)}{\mathrm{d}t} = \left(\left[K^\mathrm{EE} - \frac{\eta_\mathrm{E} \rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \delta + \frac{\eta_\mathrm{E} M_\mathrm{E}}{L^2} \left(\frac{\rho_0}{w_0^\mathrm{E}} - \frac{N_\mathrm{E} M_\mathrm{E}}{L} \right) \right] * \delta w^\mathrm{E} \right)(\mu)$$

$$- \left(\left[K^\mathrm{EI} - \frac{\eta_\mathrm{E} N_\mathrm{I} M_\mathrm{E} M_\mathrm{I}}{L^3} \right] * \delta w^\mathrm{I} \right)(\mu). \tag{60}$$

Decoupling of spatial frequencies

The convolutions in ***Equations 34*** and ***60*** show how the excitatory and inhibitory weight perturbations at one location influence the time evolution of weights at every other location. Transforming the system to frequency space leads to a drastic simplification: The time evolution of a perturbation of a particular spatial frequency only depends on the excitatory and inhibitory perturbation of the same spatial frequency, i.e., the Fourier components decouple. We define the Fourier transform $f(k) \equiv \mathcal{F}[f(\mu)]$ with wavevector k of a function $f(\mu)$ of location μ as:

$$f(k) \equiv \int_{-\infty}^{+\infty} f(\mu) e^{-ik\mu} \, \mathrm{d}\mu \tag{61}$$

and note that

$$\int_{-\infty}^{+\infty} e^{-ik\mu} \, \mathrm{d}\mu = 2\pi\delta(k) \,. \tag{62}$$

Using the Convolution theorem and the linearity of the Fourier transform we get

$$\begin{aligned} \frac{\mathrm{d}\delta w^{\mathrm{E}}(k)}{\mathrm{d}t} = &\left[\frac{\eta_{\mathrm{E}} M_{\mathrm{E}}}{L^2} \left(\frac{\rho_0}{w_0^{\mathrm{E}}} - \frac{N_{\mathrm{E}} M_{\mathrm{E}}}{L} \right) \delta w^{\mathrm{E}}(k) + \frac{\eta_{\mathrm{E}} N_{\mathrm{I}} M_{\mathrm{E}} M_{\mathrm{I}}}{L^3} \delta w^{\mathrm{I}}(k) \right] 2\pi\delta(k) \\ &- \frac{\eta_{\mathrm{E}} \rho_0 M_{\mathrm{E}}}{w_0^{\mathrm{E}} L} \delta w^{\mathrm{E}}(k) + \left[K^{\mathrm{EE}}(k) \delta w^{\mathrm{E}}(k) - K^{\mathrm{EI}}(k) \delta w^{\mathrm{I}}(k) \right] \end{aligned} \tag{63}$$

and

$$\frac{\mathrm{d}\delta w^{\mathrm{I}}(k)}{\mathrm{d}t} = K^{\mathrm{IE}}(k) \delta w^{\mathrm{E}}(k) - K^{\mathrm{II}}(k) \delta w^{\mathrm{I}}(k) \,. \tag{64}$$

The $\delta(k)$ term in ***Equation 63*** balances homogeneous perturbations in such a way that the output neuron would still fire at the target rate, if not for permutations at other frequencies. In the following, we drop this term, because we are not interested in spatially homogeneous perturbations. Moreover, the continuum limit is only valid for high densities: $N_{\mathrm{P}}/L \to \infty$. We can thus drop terms of lower order than N_{X}/L, which eliminates the $\frac{\eta_{\mathrm{E}} \rho_0 M_{\mathrm{E}}}{w_0^{\mathrm{E}} L}$ term. Writing the remaining terms of ***Equations 63*** and ***64*** as a matrix leads to:

$$\begin{bmatrix} \delta\dot{w}^{\mathrm{E}} \\ \delta\dot{w}^{\mathrm{I}} \end{bmatrix} (k) = \begin{bmatrix} K^{\mathrm{EE}}(k) & -K^{\mathrm{EI}}(k) \\ K^{\mathrm{IE}}(k) & -K^{\mathrm{II}}(k) \end{bmatrix} \begin{bmatrix} \delta w^{\mathrm{E}} \\ \delta w^{\mathrm{I}} \end{bmatrix} (k) \,, \tag{65}$$

which contains no terms from the weight normalization anymore. The characteristic polynomial of the above matrix is:

$$\lambda^2 + \lambda \left(K^{\mathrm{II}} - K^{\mathrm{EE}} \right) + K^{\mathrm{EI}} K^{\mathrm{IE}} - K^{\mathrm{EE}} K^{\mathrm{II}} = 0 \tag{66}$$

The difference, $K^{\mathrm{EI}} K^{\mathrm{IE}} - K^{\mathrm{EE}} K^{\mathrm{II}}$, vanishes for Gaussian input, because:

$$K^{\mathrm{PP'}}(\mu, \mu' = 0) = \frac{\eta^{\mathrm{P}} N_{\mathrm{P'}}}{L} \left\langle r^{\mathrm{P}}(\mu) r^{\mathrm{P'}}(0) \right\rangle_x \tag{67}$$

$$= \frac{\alpha_{\mathrm{P}} \alpha_{\mathrm{P'}} \eta^{\mathrm{P}} N_{\mathrm{P'}}}{L^2} \int_{-L/2}^{+L/2} \mathrm{d}x \exp\left\{ -\frac{(x-\mu)^2}{2\sigma_{\mathrm{P}}^2} - \frac{x^2}{2\sigma_{\mathrm{P'}}^2} \right\} \tag{68}$$

$$\approx \frac{\alpha_{\mathrm{P}} \alpha_{\mathrm{P'}} \eta^{\mathrm{P}} N_{\mathrm{P'}}}{L^2} \sqrt{\frac{2\pi}{\frac{1}{\sigma_{\mathrm{P}}^2} + \frac{1}{\sigma_{\mathrm{P'}}^2}}} \exp\left\{ -\frac{\mu^2}{2(\sigma_{\mathrm{P}}^2 + \sigma_{\mathrm{P'}}^2)} \right\} , \tag{69}$$

where we completed the square and used $\int_{-\infty}^{+\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}$. Taking the Fourier transform and completing the square again gives

$$K^{\mathrm{PP'}}(k) = \frac{\eta^{\mathrm{P}} N_{\mathrm{P'}} M_{\mathrm{P}} M_{\mathrm{P'}}}{L^2} \exp\left\{ -\frac{k^2}{2} (\sigma_{\mathrm{P}}^2 + \sigma_{\mathrm{P'}}^2) \right\} . \tag{70}$$

and thus $K^{\mathrm{EI}}K^{\mathrm{IE}} - K^{\mathrm{EE}}K^{\mathrm{II}} = 0$.
For $\mathrm{P} = \mathrm{P}'$ ***Equation 70*** simplifies to:

$$K^{\mathrm{PP}}(k) = \frac{\eta^{\mathrm{P}} N_{\mathrm{P}} M_{\mathrm{P}}^2}{L^2} \exp\left\{-k^2\sigma_{\mathrm{P}}^2\right\} . \tag{71}$$

This leads to the eigenvalues:

$$\lambda_0(k) = 0 \tag{72}$$

$$\lambda_1(k) = K^{\mathrm{EE}}(k) - K^{\mathrm{II}}(k) \tag{73}$$

$$= \frac{1}{L^2}\left(\eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \exp\left\{-k^2\sigma_{\mathrm{E}}^2\right\} - \eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \exp\left\{-k^2\sigma_{\mathrm{I}}^2\right\}\right) , \tag{74}$$

which are shown in ***Figure 8a***. Perturbations with spatial frequencies for which $\lambda_1(k)$ is positive will grow. Setting $\frac{\mathrm{d}\lambda_1(k)}{\mathrm{d}k} = 0$ gives the wavevector $k_{\max}$ of the Fourier component that grows fastest:

$$\frac{2}{L^2}\left(\eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \sigma_{\mathrm{I}}^2 k_{\max} \exp\left\{-k_{\max}^2\sigma_{\mathrm{I}}^2\right\} - \eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \sigma_{\mathrm{E}}^2 k_{\max} \exp\left\{-k_{\max}^2\sigma_{\mathrm{E}}^2\right\}\right) = 0 \tag{75}$$

$$\Rightarrow \ln(\eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \sigma_{\mathrm{I}}^2) - k_{\max}^2\sigma_{\mathrm{I}}^2 = \ln(\eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \sigma_{\mathrm{E}}^2) - k_{\max}^2\sigma_{\mathrm{E}}^2 \tag{76}$$

$$\Rightarrow k_{\max} = \sqrt{\frac{\ln(\frac{\eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \sigma_{\mathrm{I}}^2}{\eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \sigma_{\mathrm{E}}^2})}{\sigma_{\mathrm{I}}^2 - \sigma_{\mathrm{E}}^2}} . \tag{77}$$

Assuming that the fastest-growing spatial frequency from the linearized system will prevail, the final spacing of the periodic pattern, ℓ, is determined by:

$$\ell = 2\pi/k_{\max} = 2\pi\sqrt{\frac{\sigma_{\mathrm{I}}^2 - \sigma_{\mathrm{E}}^2}{\ln(\frac{\eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \sigma_{\mathrm{I}}^2}{\eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \sigma_{\mathrm{E}}^2})}} = 2\pi\sqrt{\frac{\sigma_{\mathrm{I}}^2 - \sigma_{\mathrm{E}}^2}{\ln\left(\frac{\eta_{\mathrm{I}} N_{\mathrm{I}} \alpha_{\mathrm{I}}^2 \sigma_{\mathrm{I}}^4}{\eta_{\mathrm{E}} N_{\mathrm{E}} \alpha_{\mathrm{E}}^2 \sigma_{\mathrm{E}}^4}\right)}} . \tag{78}$$

Equation 78 is in exact agreement with the grid spacing obtained in simulations (***Figure 1g***). Moreover, it indicates the bifurcation point: When excitation is as smooth as inhibition ($\sigma_{\mathrm{E}} = \sigma_{\mathrm{I}}$), there is no unstable spatial frequency anymore and every perturbation gets balanced (***Figure 1g***, compare ***Equation 103***). The grid spacing also depends on the ratio of the inhibitory and excitatory parameters $\eta^{\mathrm{P}}, N_{\mathrm{P}}, \alpha_{\mathrm{P}}$ (logarithmic term in ***Equation 78***). We confirmed this dependence with simulations on the linear track where we increased either η_{I} or N_{I} or α_{I}^2 such that the product $\gamma = \eta_{\mathrm{I}} N_{\mathrm{I}} \alpha_{\mathrm{I}}^2$ increases with respect to the initial product γ_0. We find a good agreement with the theoretical prediction for all three variations (***Figure 8b***).
Note that the term $\eta^{\mathrm{P}} M_{\mathrm{P}}^2 N_{\mathrm{P}}$ in the logarithm in ***Equation 78*** is essentially a factor that determines the rate of weight change of population P: η^{P} is just the scaling factor; M_{P} is the mass under a tuning function (with quadratic influence: once directly through the firing rate of the input, once through the increased firing rate of the output neuron) ; N_{P} is the number of tuning functions. The remaining σ_{P}^2 originates specifically from the Gaussian shape of the tuning functions.

Analysis for non-localized input (Gaussian random fields)

Above, we derived the time evolution of perturbations of excitatory and inhibitory weights for place field-like input, i.e., Gaussian tuning curves. In the following we conduct a similar analysis, using non-localized input, i.e., random functions with a given spatial autocorrelation length. We show that the grid spacing is predicted by an equation that is equivalent to ***Equation 78***.
The non-localized input r_i^{P} for input neuron i of population P was obtained by rescaling a Gaussian random field (GRF) g_i^{P} to mean $1/2$ and minimum 0:

$$r_i^{\mathrm{P}}(x) = \frac{g_i^{\mathrm{P}}(x) - \min_x g_i^{\mathrm{P}}(x)}{2\left\langle g_i^{\mathrm{P}}(x) - \min_x g_i^{\mathrm{P}}(x)\right\rangle_x} , \tag{79}$$

where $\min_x$ denotes the minimum over all locations and the GRF g_i^{P} is obtained by convolving a Gaussian $\mathcal{G}^{\mathrm{P}}(x) = \exp(-x^2/2\sigma_{\mathrm{P}}^2)$ with white noise ξ_i from a uniform distribution between -0.5 an 0.5:

$$g_i^{\mathrm{P}}(x) = \int \mathcal{G}^{\mathrm{P}}\left(x - x'\right) \xi_i^{\mathrm{P}}(x')\,\mathrm{d}x' \,. \tag{80}$$

Since the white noise has zero mean, the spatial average of a GRF is also 0:

$$\left\langle g_i^{\mathrm{P}}(x)\right\rangle_x = \int \left\langle \mathcal{G}^{\mathrm{P}}\left(x - x'\right)\right\rangle_x \xi_i^{\mathrm{P}}(x')\,\mathrm{d}x' \tag{81}$$

$$\propto \int \xi_i^{\mathrm{P}}(x')\,\mathrm{d}x' = 0 \tag{82}$$

The individual minima $\min_x g_i^{\mathrm{P}}(x)$ in ***Equation 79*** would complicate the subsequent analysis. If we again consider infinitely large systems $L \to \infty$ with infinite density $N_{\mathrm{P}}/L \to \infty$, ***Equation 79*** simplifies. The mean of the distribution of GRF minima over different input neurons scales logarithmically with the number of samples (***Bovier, 2005***). Here the number of samples corresponds to the number of minima in a GRF, which scales inversely with the width of the convolution kernel that was used to obtain the GRF:

$$\text{Number of minima in a GRF} \propto L/\sigma_{\mathrm{P}} \,. \tag{83}$$

In the continuum limit the variance of the minima distribution over cells decreases and the relative difference between the mean minimum value of excitation and inhibition vanishes[1] (***Figure 8c***). We thus take the minimum value as a constant m, which does not depend on the population nor on the input neuron. This leads to a simplified expression of the input tuning functions:

$$r_i^{\mathrm{P}}(x) = \frac{1}{2}\left(1 - \frac{g_i^{\mathrm{P}}(x)}{m}\right) . \tag{85}$$

Since $\left\langle r_i^{\mathrm{P}}\right\rangle_x = 0.5$ is independent of i, equal excitatory weights are a fixed point for the excitatory learning rule ***Equation 7*** as described in ***Equation 19***. Moreover, the sum over all input neurons does not depend on the location:

$$\sum_{i=1}^{N_{\mathrm{P}}} r_i^{\mathrm{P}}(x) = \frac{1}{2}\left(\sum_{i=1}^{N_{\mathrm{P}}} 1 - \sum_{i=1}^{N_{\mathrm{P}}} g_i^{\mathrm{P}}(x)\right) = \frac{N_{\mathrm{P}}}{2} - \frac{1}{2}\int \mathcal{G}^{\mathrm{P}}\left(x - x'\right) \underbrace{\sum_{i=1}^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(x')}_{=0 \text{ in cont. limit}} \mathrm{d}x' = \frac{N_{\mathrm{P}}}{2} \,. \tag{86}$$

Therefore, given constant excitatory weights, all inhibitory weights can be set to a value w_0^{I} such that the output neuron fires at the target rate, i.e., homogeneous weights are a fixed point of the learning rules, as in the scenario with Gaussian input. Moreover, ***Equation 29*** holds also for GRF input. The analysis of the projection operator of the weight normalization lead to a term of homogeneous weight perturbations and a term that could be neglected in the high density limit. We now omit these terms a priori. The time evolution of excitatory and inhibitory weight perturbations can thus be summarized as (compare ***Equations 29*** and ***44***):

$$\frac{\mathrm{d}\delta w_i^{\mathrm{P}}}{\mathrm{d}t} = \eta^{\mathrm{P}}\left(\sum_{k=1}^{N_{\mathrm{E}}} \left\langle r_i^{\mathrm{P}}(x) r_k^{\mathrm{E}}(x)\right\rangle_x \delta w_k^{\mathrm{E}} - \sum_{k'=1}^{N_{\mathrm{I}}} \left\langle r_i^{\mathrm{P}}(x) r_{k'}^{\mathrm{I}}(x)\right\rangle_x \delta w_{k'}^{\mathrm{I}}\right) . \tag{87}$$

The above equation describes the time evolution of each synaptic weight. For the Gaussian input of the earlier sections, each synaptic weight is associated with one location. In the continuum limit we thus identified the synaptic weight associated to location μ with $w^{\mathrm{P}}(\mu)$. An increase of $w^{\mathrm{E}}(\mu)$ corresponded to an increase in firing at location μ (and in the surrounding, given by the width of

[1]

$$\frac{\log(L/\sigma_{\mathrm{E}}) - \log(L/\sigma_{\mathrm{I}})}{\log(L/\sigma_{\mathrm{E}})} = \frac{\log(\sigma_{\mathrm{I}}/\sigma_{\mathrm{E}})}{\log(L/\sigma_{\mathrm{E}})} \to 0 \,. \tag{84}$$

For the argument it doesn't matter if it scales purely logarithmically or with $\log^\gamma$, where γ is any exponent.

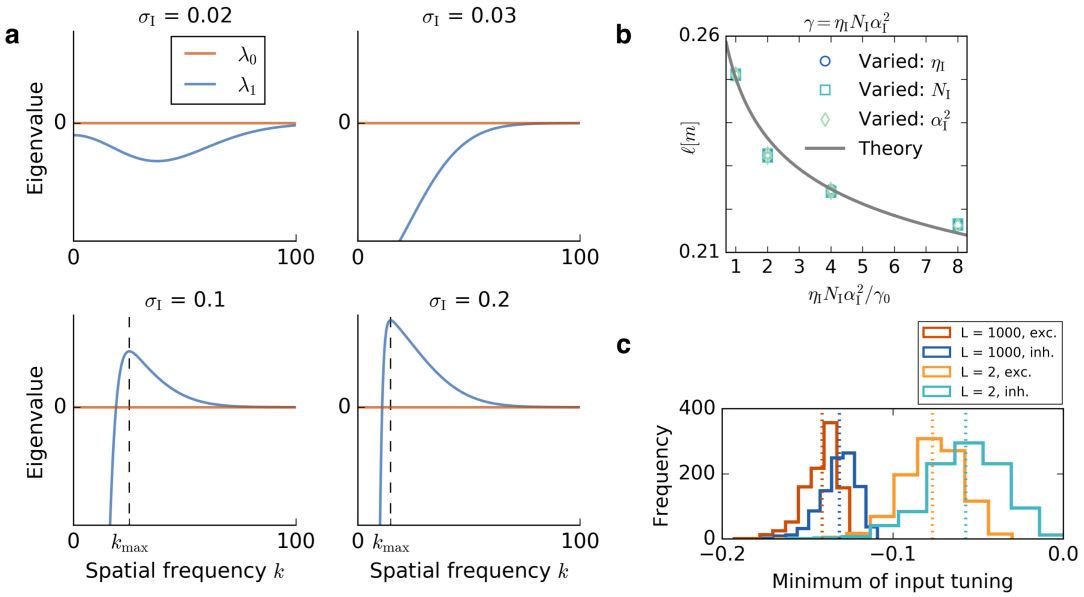

Figure 8. Results of the mathematical analysis. **a**) The eigenvalue spectrum for the eigenvalues of ***Equation 72*** for an excitatory tuning of width $\sigma_E = 0.03$. The first eigenvalue λ_0 is always 0. If the inhibitory tuning is more narrow than the excitatory tuning, i.e., $\sigma_I < \sigma_E$, the second eigenvalue λ_1 is negative for every wavevector k. For $\sigma_I > \sigma_E$ the eigenvalue spectrum has a unique positive maximum k_{max}, i.e., a most unstable spatial frequency. The wavevector k_{max} at which λ_1 is maximal is obtained from ***Equation 78*** and marked with a dashed line. **b**) The dependence of the grid spacing on learning rate η_I, number of input neurons N_I and input height α_I is accurately predicted by the theory. The gray line shows the grid spacing obtained from ***Equation 78***. We vary either the inhibitory learning rate, η_I (circles), the number of inhibitory input neurons, N_I (squares), or the square of the height of the inhibitory input place fields, α_I^2 (diamonds). The horizontal axis shows the ratio of the product $\eta_I N_I \alpha_I^2$ to the initial value of the product γ_0. We keep $\eta_E = 0.3 \times 10^{-4}$, $N_E = 800$ and $\alpha_E = 1$ in each simulation and the γ_0 parameters are: $\eta_I = 0.3 \times 10^{-3}$, $N_I = 200$, $\alpha_I = 1$. **c**) Distribution of minimal values of GRF input. Histograms show the distribution of the minimal values of 1000 Gaussian random fields for a small linear track, $L = 2$, and a large linear track $L = 1000$. Red and blue colors correspond to the tuning of excitatory and inhibitory input neurons, respectively. Each dotted line indicates the mean of the histogram of the same color. For larger systems, the distribution of the minimum values gets more narrow and the relative distance between the minima of excitatory and inhibitory neurons decreases.

the Gaussian of the excitatory tuning). Analogously, an increase of $w^I(\mu)$ caused a decrease in firing at location μ (and in the surrounding given by the width of the Gaussian of the inhibitory tuning). Because of the non-localized tuning of GRF input, each synaptic weight has an influence on the firing rate at many locations. The influence of neuron i of population P at location μ is expressed by $\xi_i^P(\mu)$. If one wanted to increase the firing rate at a specific location μ – and not just everywhere – one would thus increase all excitatory weights that have a high $\xi_i^P(\mu)$ and decrease all excitatory weights that have a low $\xi_i^P(\mu)$ (note that ξ^P can also be negative). The 'weight' that corresponds to location μ is thus expressed as:

$$w^P(\mu) := \sum_i^{N_P} w_i^P \xi_i^P(\mu), \tag{88}$$

where we weighted each synaptic weight with the value of the corresponding white noise at location μ. This corresponds to expressing the weights in a basis that is associated with the location and not with the individual input neurons. Combining ***Equation 88*** and ***Equation 87*** gives the time evolution

of the weight perturbations associated with location μ:

$$\frac{\mathrm{d}\delta w^{\mathrm{P}}(\mu)}{\mathrm{d}t} = \sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) \frac{\mathrm{d}\delta w_i^{\mathrm{P}}}{\mathrm{d}t} \tag{89}$$

$$= \eta^{\mathrm{P}} \sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) \left(\sum_{k=1}^{N_{\mathrm{E}}} \left\langle r_i^{\mathrm{P}}(x) r_k^{\mathrm{E}}(x) \right\rangle_x \delta w_k^{\mathrm{E}} - \sum_{k'=1}^{N_{\mathrm{I}}} \left\langle r_i^{\mathrm{P}}(x) r_{k'}^{\mathrm{I}}(x) \right\rangle_x \delta w_{k'}^{\mathrm{I}} \right) . \tag{90}$$

We now look at the first term of the above equation, the second term will be treated analogously:

$$\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) \sum_{k=1}^{N_{\mathrm{E}}} \left\langle r_i^{\mathrm{P}}(x) r_k^{\mathrm{E}}(x) \right\rangle_x \delta w_k^{\mathrm{E}} = \left\langle \left(\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) r_i^{\mathrm{P}}(x) \right) \left(\sum_{k=1}^{N_{\mathrm{E}}} \delta w_k^{\mathrm{E}} r_k^{\mathrm{E}}(x) \right) \right\rangle_x . \tag{91}$$

The sum containing the white noise can be simplified using the zero mean and the expression for the variance of the uniform white noise:

$$\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) r_i^{\mathrm{P}}(x) = \frac{1}{2} \left(\underbrace{\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu)}_{=0} - \frac{1}{m} \sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) g_i^{\mathrm{P}}(x) \right) \tag{92}$$

$$= -\frac{1}{2m} \sum_i^{N_{\mathrm{P}}} \int \mathcal{G}^{\mathrm{P}}\left(x - x'\right) \underbrace{\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) \xi_i^{\mathrm{P}}(x')}_{=\beta N_{\mathrm{P}} \delta(x'-\mu) \text{ in cont. limit}} \mathrm{d}x' \tag{93}$$

$$= -\frac{\beta N_{\mathrm{P}}}{2m} \mathcal{G}^{\mathrm{P}}\left(x - \mu\right) , \tag{94}$$

where β is a proportionality constant that does not depend on the population type P. The Dirac delta $\delta(x' - \mu)$ occurs, because the white noise at different locations is uncorrelated. The sum of the product of weight perturbations and input rates can be rewritten as:

$$\sum_{k=1}^{N_{\mathrm{E}}} \delta w_k^{\mathrm{E}} r_k^{\mathrm{E}}(x) = \frac{1}{2} \left(\underbrace{\sum_{k=1}^{N_{\mathrm{E}}} \delta w_k^{\mathrm{E}}}_{\text{homog. pert.}} - \frac{1}{m} \int \mathcal{G}^{\mathrm{E}}\left(x - \mu'\right) \underbrace{\sum_{k=1}^{N_{\mathrm{E}}} \delta w_k^{\mathrm{E}} \xi_k^{\mathrm{E}}(\mu')}_{=: \delta w^{\mathrm{E}}(\mu'); \textit{ Equation 88}} \mathrm{d}\mu' \right) . \tag{95}$$

The first term is independent of location x and will thus only lead to spatially homogeneous perturbations which we do not consider in the following. Inserting ***Equations 94*** and ***95*** and the analogous terms for inhibition in ***Equation 91*** leads to:

$$\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) \sum_{k=1}^{N_{\mathrm{E}}} \left\langle r_i^{\mathrm{P}}(x) r_k^{\mathrm{E}}(x) \right\rangle_x \delta w_k^{\mathrm{E}} = \frac{\beta N_{\mathrm{P}}}{4m^2} \int \left\langle \mathcal{G}^{\mathrm{P}}\left(x - \mu\right) \mathcal{G}^{\mathrm{E}}\left(x - \mu'\right) \right\rangle_x \delta w^{\mathrm{E}}(\mu') \, \mathrm{d}\mu' \tag{96}$$

$$= \frac{1}{\eta^{\mathrm{P}}} \int \hat{K}^{\mathrm{PE}}(\mu - \mu') \delta w^{\mathrm{E}}(\mu') \, \mathrm{d}\mu' \tag{97}$$

$$= \frac{1}{\eta^{\mathrm{P}}} (\hat{K}^{\mathrm{PE}} * \delta w^{\mathrm{E}})(\mu) , \tag{98}$$

where we introduced kernels for the translation invariant overlap between two Gaussians with different centers (similar to ***Equation 32***):

$$\hat{K}^{\mathrm{PP}'}(\mu - \mu') := \frac{\beta \eta^{\mathrm{P}} N_{\mathrm{P}}}{4m^2} \left\langle \mathcal{G}^{\mathrm{P}}\left(\mu\right) \mathcal{G}^{\mathrm{P}'}\left(\mu'\right) \right\rangle_x = \frac{\beta \eta^{\mathrm{P}} N_{\mathrm{P}}}{4m^2} \left\langle \mathcal{G}^{\mathrm{P}}\left(0\right) \mathcal{G}^{\mathrm{P}'}\left(|\mu - \mu'|\right) \right\rangle_x \tag{99}$$

Equation 89 can thus be written as:

$$\frac{\mathrm{d}\delta w^{\mathrm{P}}(\mu)}{\mathrm{d}t} = (\hat{K}^{\mathrm{PE}} * \delta w^{\mathrm{E}})(\mu) - (\hat{K}^{\mathrm{PI}} * \delta w^{\mathrm{I}})(\mu) , \tag{100}$$

which leads to a dynamical system for the Fourier components of the weight perturbations that is equivalent to ***Equation 65*** with eigenvalues:

$$\lambda_0(k) = 0 \tag{101}$$

$$\lambda_1(k) = \hat{K}^{\mathrm{EE}}(k) - \hat{K}^{\mathrm{II}}(k) \tag{102}$$

$$= \frac{\beta}{4m^2}\left(\eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \exp\left\{-k^2\sigma_{\mathrm{E}}^2\right\} - \eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \exp\left\{-k^2\sigma_{\mathrm{I}}^2\right\}\right) . \tag{103}$$

We thus get the same expression for the grid spacing as in the scenario of Gaussian input (with $\alpha_{\mathrm{E}} = \alpha_{\mathrm{I}} = 1$):

$$\ell = \sqrt{\frac{\sigma_{\mathrm{I}}^2 - \sigma_{\mathrm{E}}^2}{\ln\left(\frac{\eta_{\mathrm{I}}\sigma_{\mathrm{I}}^4 N_{\mathrm{I}}}{\eta_{\mathrm{E}}\sigma_{\mathrm{E}}^4 N_{\mathrm{E}}}\right)}} . \tag{104}$$

Glossary

A summary of notation:

The rat's position at time t :	$\mathbf{x}(t)$
Spatial dimensions x, y and head direction z :	$\mathbf{x} = (x, y, z)$
Population label; can be E (excitatory) or I (inhibitory) :	P
Standard deviation of Gaussian tuning of population P :	σ_{P}
Spatial autocorrelation length of input of population P :	$\sigma_{\mathrm{P,corr}}$
Number of input neurons of population P :	N_{P}
Number of place fields per input neuron of population P :	$N_{\mathrm{P}}^{\mathrm{f}}$
Firing rate of output neuron :	$r^{\mathrm{out}}(x)$
Firing rate of input neuron i of population P :	$r_i^{\mathrm{P}}(x)$
Synaptic weight of input neuron i of population P to output neuron :	$w_i^{\mathrm{P}}(t)$
Learning rates of excitation and inhibition :	$\eta_{\mathrm{E}}, \eta_{\mathrm{I}}$
Target rate of the output neuron :	ρ_0
Length of linear track :	L
Height of the Gaussian input fields :	$\alpha_{\mathrm{E}}, \alpha_{\mathrm{I}}$
Value of Gaussian with standard deviation σ_{P} at location x :	$\mathcal{G}^{\mathrm{P}}(x)$
Von Mises distribution with width σ_{P} that is periodic in $[-L/2, L/2]$:	$\mathcal{M}^{\mathrm{P}}(x)$

Simulation parameters

	$[\sigma_{\mathrm{E},x}, \sigma_{\mathrm{E},y}, \sigma_{\mathrm{E},z}]$	N_{E}	η_{E}	$w^{\mathrm{E,init}}$	$N_{\mathrm{E}}^{\mathrm{f}}$
Figure 1b	0.05	2000	2×10^{-6}	1	∞
Figure 1c	0.08	2000	2×10^{-6}	1	∞
Figure 1d	0.06	2000	2×10^{-6}	1	∞
Figure 1f	0.04	160	2×10^{-6}	1	1
Figure 1g	0.03	1600	3.6×10^{-5}	1	1
Figure 1h	0.03	10000	3.5×10^{-7}	1	∞
Figure 2a	[0.05, 0.05]	4900	6.7×10^{-5}	1	1
Figure 2b	[0.05, 0.05]	4900	2×10^{-6}	1	100
Figure 2c	[0.05, 0.05]	4900	6×10^{-6}	1	∞
Figure 3a-d	[0.05, 0.05]	4900	2×10^{-4}	1	1
Figure 4	[0.05, 0.05]	2×4900	1.3×10^{-4}	1	1
Figure 5a	[0.07, 0.07]	4900	6×10^{-6}	0.5	∞
Figure 5b	[0.07, 0.07]	400	1.3×10^{-4}	1	1
Figure 5c	[0.05, 0.05]	4900	1.1×10^{-6}	0.0455	∞
Figure 5d	[0.08, 0.08]	4900	6×10^{-6}	0.5	∞
Figure 5e	[0.05, 0.05]	4900	6.7×10^{-5}	1	1
Figure 6a	[0.07, 0.07, 0.2]	37500	1.5×10^{-5}	1	1
Figure 6b	[0.08, 0.08, 0.2]	50000	10^{-5}	1	1
Figure 6c	[0.1, 0.1, 0.2]	50000	10^{-5}	1	1
Figure 7a	[0.05, 0.05]	4900	6.7×10^{-5}	1	1
Figure 7b	0.04	2000	5×10^{-5}	1	1
	0.04	2000	5×10^{-7}	1.0	100
	0.05	2000	5×10^{-6}	0.5	∞
Figure 7c	[0.05, 0.05]	4900	2×10^{-6}	1	100
Figure 8b	0.03	800	3.3×10^{-5}	1	1

Table 1. Parameters for excitatory inputs for all figures in the manuscript. $N_{\mathrm{E}} = \infty$ indicates that the excitatory input is a Gaussian random field.

	$[\sigma_{\mathrm{I},x}, \sigma_{\mathrm{I},y}, \sigma_{\mathrm{I},z}]$	N_{I}	η_{I}	$w^{\mathrm{I,init}}$	$N_{\mathrm{I}}^{\mathrm{f}}$
Figure 1b	0.12	500	2×10^{-5}	4.4	∞
Figure 1c	0.07	2000	2×10^{-5}	1.1	∞
Figure 1d	∞	500	2×10^{-5}	4.39	∞
Figure 1f	0.13	40	2×10^{-5}	1.31	1
Figure 1g	From 0.08 to 0.3 in 0.02 steps	400	3.6×10^{-4}	***Equation 111***	1
Figure 1h	From 0.08 to 0.3 in 0.02 steps	2500	7×10^{-6}	4.03	∞
Figure 2a	[0.1, 0.1]	1225	2.7×10^{-4}	1.5	1
Figure 2b	[0.1, 0.1]	1225	8×10^{-6}	1.52	100
Figure 2c	[0.1, 0.1]	1225	6×10^{-5}	4.0	∞
Figure 3a-d	[0.1, 0.1]	1225	8×10^{-4}	1.5	1
Figure 4	[0.1, 0.1]	2×1225	5.3×10^{-4}	1.51	1
Figure 5a	$[\infty, \infty]$	1225	6×10^{-5}	2	∞
Figure 5b	$[\infty, \infty]$	1	5.3×10^{-4}	69.5	1
Figure 5c	[0.049, 0.049]	1225	4.4×10^{-5}	0.175	∞
Figure 5d	[0.3, 0.07]	1225	6×10^{-5}	2	∞
Figure 5e	[0.049, 0.049]	4900	2.7×10^{-4}	1.02	1
	[0.2, 0.1]; [0.1, 0.2]	1225	2.7×10^{-4}	1.04	1
	[2, 0.1]; [0.1, 2]	1225	2.7×10^{-4}	2.74	1
	[2, 0.2]; [0.2, 2]	1225	2.7×10^{-4}	1.38	1
	[0.1, 0.1]	1225	2.7×10^{-4}	1.5	1
	[0.2, 0.2]	1225	2.7×10^{-4}	0.709	1
	[2, 2]	1225	2.7×10^{-4}	0.259	1
	[0.1, 0.049]; [0.049, 0.1]	1225	2.7×10^{-4}	2.48	1
	[0.2, 0.049]; [0.049, 0.2]	1225	2.7×10^{-4}	1.74	1
	[2, 0.049]; [0.049, 2]	1225	2.7×10^{-4}	5.56	1
Figure 6a	[0.15, 0.15, 0.2]	9375	1.5×10^{-4}	1.55	1
Figure 6b	[0.12, 0.12, 1.5]	3125	10^{-4}	5.68	1
Figure 6c	[0.09, 0.09, 1.5]	12500	10^{-4}	2.71	1
Figure 6d	Same as	***Figure 6a,b,c***			
Figure 7a	[0.1, 0.1]	1225	2.7×10^{-4}	1.5	1
Figure 7b	0.12	500	5×10^{-4}	1.6	1
	0.12	500	5×10^{-6}	1.62	100
	0.12	500	5×10^{-5}	1.99	∞
Figure 7c	[0.1, 0.1]	1225	8×10^{-6}	1.52	100
Figure 8b	0.1	varied	varied	varied	1

Table 2. Parameters for inhibitory inputs for all figures in the manuscript. $N_{\mathrm{I}} = \infty$ indicates that the inhibitory input is a Gaussian random field. We denote spatially untuned inhibition with: $\sigma_{\mathrm{I}} = \infty$.

	t_{sim}	L
Figure 1b	2,000,000	2
Figure 1c	2,000,000	2
Figure 1d	400,000	2
Figure 1f	20,000,000	2
Figure 1g	80,000,000	14
Figure 1h	40,000,000	10
Figure 2a,b,c	1,800,000	1
Figure 3a,b,c,d	540,000	1
Figure 4	1,800,000	1
Figure 5a,c,d,e	1,800,000	1
Figure 5b	180,000	1
Figure 6a,b,c,d	1,800,000	1
Figure 7a,c	1,800,000	1
Figure 7b	400,000	2
Figure 8b	40,000,000	3

Table 3. Simulation time t_{sim} and system size L for all figures in the manuscript.

Simulation parameters of figure supplements

	$[\sigma_{\mathrm{E},x}, \sigma_{\mathrm{E},y}, \sigma_{\mathrm{E},z}]$	N_{E}	η_{E}	$w^{\mathrm{E,init}}$	$N_{\mathrm{E}}^{\mathrm{f}}$
Figure 1–Figure supplement 1	0.04	2000	5×10^{-7}	1	varied
Figure 1–Figure supplement 2	see	caption			
Figure 2–Figure supplement 1	[0.05, 0.05]	4900	6.7×10^{-5}	1	1
Figure 2–Figure supplement 2	[0.05, 0.05]	4900	6.7×10^{-5}	1	1
Figure 2–Figure supplement 4	[0.05, 0.05]	4900	0.0	1	1
Figure 2–Figure supplement 5	see	caption			
Figure 2–Figure supplement 6	[0.05, 0.05]	4900	3.3×10^{-5}	1	1
Figure 3–Figure supplement 1	see	caption			
Figure 3–Figure supplement 2	see	caption			
Figure 3–Figure supplement 3	[0.05, 0.05]	4900	0.0	1	1
Figure 6–Figure supplement 1	see	caption			

Table 4. Parameters for excitatory inputs in supplement figures. $N_{\mathrm{E}} = \infty$ indicates that the excitatory input is a Gaussian random field.

	$[\sigma_{\mathrm{I},x}, \sigma_{\mathrm{I},y}, \sigma_{\mathrm{I},z}]$	N_{I}	η_{I}	$w^{\mathrm{I,init}}$	$N_{\mathrm{I}}^{\mathrm{f}}$
Figure 1–Figure supplement 1	0.12	500	5×10^{-6}	1.61	varied
Figure 1–Figure supplement 2	see	caption			
Figure 2–Figure supplement 1	[0.1, 0.1]	1225	0.0	1.5	1
Figure 2–Figure supplement 2	[0.1, 0.1]	1225	0.0	1.5	1
Figure 2–Figure supplement 4	[0.1, 0.1]	1225	0.0	1.5	1
Figure 2–Figure supplement 5	see	caption			
Figure 2–Figure supplement 6	[0.1, 0.1]	1225	5.3×10^{-6}	0.03	50
Figure 3–Figure supplement 1	see	caption			
Figure 3–Figure supplement 2	see	caption			
Figure 3–Figure supplement 3	[0.1, 0.1]	1225	0.0	1.5	1
Figure 6–Figure supplement 1	see	caption			

Table 5. Parameters for inhibitory inputs in supplement figures. $N_{\mathrm{I}} = \infty$ indicates that the inhibitory input is a Gaussian random field. We denote spatially untuned inhibition with: $\sigma_{\mathrm{I}} = \infty$.

	t_{sim}	L
Figure 1–Figure supplement 1	48,000,000	1
Figure 1–Figure supplement 2	see	caption
Figure 2–Figure supplement 1	1,800,000	0.5
Figure 2–Figure supplement 2	1,800,000	0.5
Figure 2–Figure supplement 4	180,000	0.5
Figure 2–Figure supplement 5	see	caption
Figure 2–Figure supplement 6	1,800,000	0.5
Figure 3–Figure supplement 1	see	caption
Figure 3–Figure supplement 2	see	caption
Figure 3–Figure supplement 3	1,800,000	0.5
Figure 6–Figure supplement 1	see	caption

Table 6. Simulation time t_{sim} and system size L for supplement figures.

Acknowledgments

We are grateful to W. Gerstner for feedback on the manuscript, J. Bölts for repeating the simulations to ensure reproducibility and O. Mackwood for providing software tools. This work was funded by the German Federal Ministry for Education and Research, FKZ 01GQ1201.

References

Arleo A, Gerstner W. Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. Biological cybernetics. 2000; 83(3):287–299.

Barry C, Ginzberg LL, O'Keefe J, Burgess N. Grid cell firing patterns signal environmental novelty by expansion. Proceedings of the National Academy of Sciences. 2012; 109(43):17687–17692.

Battaglia FP, Treves A. Attractor neural networks storing multiple space representations: a model for hippocampal place fields. Physical Review E. 1998; 58(6):7738.

Beed P, Gundlfinger A, Schneiderbauer S, Song J, Böhm C, Burgalossi A, Brecht M, Vida I, Schmitz D. Inhibitory gradient along the dorsoventral axis in the medial entorhinal cortex. Neuron. 2013; 79(6):1197–1207.

Bonnevie T, Dunn B, Fyhn M, Hafting T, Derdikman D, Kubie JL, Roudi Y, Moser EI, Moser MB. Grid cells require excitatory drive from the hippocampus. Nature neuroscience. 2013; 16(3):309–317.

Bovier A, Extreme values of random processes. Citeseer; 2005.

Brun VH, Leutgeb S, Wu HQ, Schwarcz R, Witter MP, Moser EI, Moser MB. Impaired spatial representation in CA1 after lesion of direct input from entorhinal cortex. Neuron. 2008; 57(2):290–302.

Brun VH, Solstad T, Kjelstrup KB, Fyhn M, Witter MP, Moser EI, Moser MB. Progressive increase in grid scale from dorsal to ventral medial entorhinal cortex. Hippocampus. 2008; 18(12):1200–1212.

Buetfering C, Allen K, Monyer H. Parvalbumin interneurons provide grid cell-driven recurrent inhibition in the medial entorhinal cortex. Nature neuroscience. 2014; 17(5):710–718.

Burak Y, Fiete IR. Accurate path integration in continuous attractor network models of grid cells. PLoS Comput Biol. 2009; 5(2):e1000291.

Burgess N, Barry C, O'Keefe J. An oscillatory interference model of grid cell firing. Hippocampus. 2007; 17(9):801–812.

Burgess N, Cacucci F, Lever C, O'Keefe J. Characterizing multiple independent behavioral correlates of cell firing in freely moving animals. Hippocampus. 2005; 15(2):149–153.

Burgess N, O'Keefe J. Models of place and grid cell firing and theta rhythmicity. Current opinion in neurobiology. 2011; 21(5):734–744.

Bush D, Burgess N. A hybrid oscillatory interference/continuous attractor network model of grid cell firing. The Journal of Neuroscience. 2014; 34(14):5065–5079.

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

Chen G, Manson D, Cacucci F, Wills TJ. Absence of visual input results in the disruption of grid cell firing in the mouse. Current Biology. 2016; 26(17):2335–2342.

Chen LL, Lin LH, Barnes CA, McNaughton BL. Head-direction cells in the rat posterior cortex. II. Contributions of visual and ideothetic information to the directional firing. Experimental brain research. 1993; 101(1):24–34.

Cilz NI, Kurada L, Hu B, Lei S. Dopaminergic modulation of GABAergic transmission in the entorhinal cortex: concerted roles of α1 adrenoreceptors, inward rectifier K+, and T-type Ca2+ channels. Cerebral Cortex. 2013; p. bht177.

Clopath C, Vogels TP, Froemke RC, Sprekeler H. Receptive field formation by interacting excitatory and inhibitory synaptic plasticity. bioRxiv. 2016; p. 066589.

Constantinescu AO, O'Reilly JX, Behrens TE. Organizing conceptual knowledge in humans with a gridlike code. Science. 2016; 352(6292):1464–1468.

Couey JJ, Witoelar A, Zhang SJ, Zheng K, Ye J, Dunn B, Czajkowski R, Moser MB, Moser EI, Roudi Y, et al. Recurrent inhibitory circuitry as a mechanism for grid formation. Nature neuroscience. 2013; 16(3):318–324.

DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends in cognitive sciences. 2007; 11(8):333–341.

Diehl GW, Hon OJ, Leutgeb S, Leutgeb JK. Grid and nongrid cells in medial entorhinal cortex represent spatial location and environmental features with complementary coding schemes. Neuron. 2017; 94(1):83–92.

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

D'Albis T, Kempter R. A single-cell spiking model for the origin of grid-cell patterns. PLoS computational biology. 2017; 13(10):e1005782.

D'amour JA, Froemke RC. Inhibitory and excitatory spike-timing-dependent plasticity in the auditory cortex. Neuron. 2015; 86(2):514–528.

Frank LM, Brown EN, Wilson MA. A comparison of the firing properties of putative excitatory and inhibitory neurons from CA1 and the entorhinal cortex. Journal of neurophysiology. 2001; 86(4):2029–2040.

Franzius M, Sprekeler H, Wiskott L. Slowness and sparseness lead to place, head-direction, and spatial-view cells. PLoS Comput Biol. 2007; 3(8):e166.

Franzius M, Vollgraf R, Wiskott L. From grids to places. Journal of computational neuroscience. 2007; 22(3):297–299.

Fuhs MC, Touretzky DS. A spin glass model of path integration in rat medial entorhinal cortex. The Journal of Neuroscience. 2006; 26(16):4266–4276.

Fyhn M, Hafting T, Treves A, Moser MB, Moser EI. Hippocampal remapping and grid realignment in entorhinal cortex. Nature. 2007; 446(7132):190–194.

Fyhn M, Molden S, Witter MP, Moser EI, Moser MB. Spatial representation in the entorhinal cortex. Science. 2004; 305(5688):1258–1264.

Giocomo LM, Moser MB, Moser EI. Computational models of grid cells. Neuron. 2011; 71(4):589–603.

Giocomo LM, Stensola T, Bonnevie T, Van Cauter T, Moser MB, Moser EI. Topography of head direction cells in medial entorhinal cortex. Current Biology. 2014; 24(3):252–262.

Grienberger C, Milstein AD, Bittner KC, Romani S, Magee JC. Inhibitory suppression of heterogeneously tuned excitation enhances spatial coding in CA1 place cells. Nature neuroscience. 2017; 20(3):417–426.

Hafting T, Fyhn M, Bonnevie T, Moser MB, Moser EI. Hippocampus-independent phase precession in entorhinal grid cells. Nature. 2008; 453(7199):1248–1252.

Hafting T, Fyhn M, Molden S, Moser MB, Moser EI. Microstructure of a spatial map in the entorhinal cortex. Nature. 2005; 436(7052):801–806.

Hangya B, Li Y, Muller RU, Czurkó A. Complementary spatial firing in place cell–interneuron pairs. The Journal of Physiology. 2010; 588(21):4165–4175.

Hardcastle K, Maheswaranathan N, Ganguli S, Giocomo LM. A multiplexed, heterogeneous, and adaptive code for navigation in medial entorhinal cortex. Neuron. 2017; 94(2):375–387.

hartmut, https://openclipart.org/detail/216359/klara; 2015.

Hebb DO. The organization of behavior: A neuropsychological approach. John Wiley & Sons; 1949.

Jung MW, Wiener SI, McNaughton BL. Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat. The Journal of neuroscience. 1994; 14(12):7347–7356.

Kropff E, Carmichael JE, Moser MB, Moser EI. Speed cells in the medial entorhinal cortex. Nature. 2015; 523(7561):419–424.

Kropff E, Treves A. The emergence of grid cells: Intelligent design or just adaptation? Hippocampus. 2008; 18(12):1256–1269.

Krupic J, Bauza M, Burton S, Barry C, O'Keefe J. Grid cell symmetry is shaped by environmental geometry. Nature. 2015; 518(7538):232–235.

Krupic J, Burgess N, O'Keefe J, Spatially periodic cells are neither formed from grids nor poor isolation; 2015.

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

Langston RF, Ainge JA, Couey JJ, Canto CB, Bjerknes TL, Witter MP, Moser EI, Moser MB. Development of the spatial representation system in the rat. Science. 2010; 328(5985):1576–1580.

Lee AK, Wilson MA. Memory of sequential experience in the hippocampus during slow wave sleep. Neuron. 2002; 36(6):1183–1194.

lemmling, https://openclipart.org/detail/17622/simple-cartoon-mouse-1; 2006.

Leutgeb S, Leutgeb JK, Barnes CA, Moser EI, McNaughton BL, Moser MB. Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. Science. 2005; 309(5734):619–623.

Marshall L, Henze DA, Hirase H, Leinekugel X, Dragoi G, Buzsáki G. Hippocampal pyramidal cell-interneuron spike transmission is frequency dependent and responsible for place modulation of interneuron discharge. J Neurosci. 2002; 22(2).

Martig AK, Mizumori SJ. Ventral tegmental area disruption selectively affects CA1/CA2 but not CA3 place fields during a differential reward working memory task. Hippocampus. 2011; 21(2):172–184.

McNaughton B, Chen L, Markus E. "Dead reckoning," landmark learning, and the sense of direction: a neurophysiological and computational hypothesis. Journal of Cognitive Neuroscience. 1991; 3(2):190–202.

McNaughton BL, Battaglia FP, Jensen O, Moser EI, Moser MB. Path integration and the neural basis of the 'cognitive map'. Nature Reviews Neuroscience. 2006; 7(8):663–678.

Mehta MR, Quirk MC, Wilson MA. Experience-dependent asymmetric shape of hippocampal receptive fields. Neuron. 2000; 25(3):707–715.

Melzer S, Michael M, Caputi A, Eliava M, Fuchs EC, Whittington MA, Monyer H. Long-range–projecting GABAergic neurons modulate inhibition in hippocampus and entorhinal cortex. Science. 2012; 335(6075):1506–1510.

Miao C, Cao Q, Moser MB, Moser EI. Parvalbumin and somatostatin interneurons control different space-coding networks in the medial entorhinal cortex. Cell. 2017; 171(3):507–521.

Miller KD, MacKay DJ. The role of constraints in Hebbian learning. Neural Computation. 1994; 6(1):100–126.

Molter C, Yamaguchi Y. Entorhinal theta phase precession sculpts dentate gyrus place fields. Hippocampus. 2008; 18(9):919–930.

Monsalve-Mercado MM, Leibold C. Hippocampal spike-timing correlations lead to hexagonal grid fields. Physical review letters. 2017; 119(3):038101.

Moser EI, Kropff E, Moser MB. Place cells, grid cells, and the brain's spatial representation system. Neuroscience. 2008; 31(1):69.

Muessig L, Hauser J, Wills TJ, Cacucci F. A developmental switch in place cell accuracy coincides with grid cell maturation. Neuron. 2015; 86(5):1167–1173.

Muller RU, Bostock E, Taube JS, Kubie JL. On the directional firing properties of hippocampal place cells. The Journal of Neuroscience. 1994; 14(12):7235–7251.

Mégevand P, http://de.mathworks.com/matlabcentral/fileexchange/43543-watson-s-u2-statistic-based-permutation-test-for-circular-data; 2013.

Navratilova Z, Godfrey KB, McNaughton BL. Grids from bands, or bands from grids? An examination of the effects of single unit contamination on grid cell firing fields. Journal of neurophysiology. 2016; 115(2):992–1002.

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

O'Keefe J, Burgess N, Donnett JG, Jeffery KJ, Maguire EA. Place cells, navigational accuracy, and the human hippocampus. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 1998; 353(1373):1333–1340.

O'Keefe J, Dostrovsky J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain research. 1971; 34(1):171–175.

Pastoll H, Solanka L, van Rossum MC, Nolan MF. Feedback inhibition enables theta-nested gamma oscillations and grid firing fields. Neuron. 2013; 77(1):141–154.

Pérez-Escobar JA, Kornienko O, Latuske P, Kohler L, Allen K. Visual landmarks sharpen grid cell metric and confer context specificity to neurons of the medial entorhinal cortex. Elife. 2016; 5:e16937.

Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I. Invariant visual representation by single neurons in the human brain. Nature. 2005; 435(7045):1102–1107.

Redish AD, Elga AN, Touretzky DS. A coupled attractor model of the rodent head direction system. Network: Computation in Neural Systems. 1996; 7(4):671–685.

Rolls ET, Stringer SM, Elliot T. Entorhinal cortex grid cells can map to hippocampal place cells by competitive learning. Network: Computation in Neural Systems. 2006; 17(4):447–465.

Rosay S, Wernle T, Mulas M, Treves A. Modeling grid fields instead of modeling grid cells; 2018, in preparation.

Rowland DC, Weible AP, Wickersham IR, Wu H, Mayford M, Witter MP, Kentros CG. Transgenically targeted rabies virus demonstrates a major monosynaptic projection from hippocampal area CA2 to medial entorhinal layer II neurons. Journal of Neuroscience. 2013; 33(37):14889–14898.

Rubin A, Yartsev MM, Ulanovsky N. Encoding of head direction by hippocampal place cells in bats. The Journal of Neuroscience. 2014; 34(3):1067–1080.

Sargolini F, Fyhn M, Hafting T, McNaughton BL, Witter MP, Moser MB, Moser EI. Conjunctive representation of position, direction, and velocity in entorhinal cortex. Science. 2006; 312(5774):758–762.

Sargolini F, Fyhn M, Hafting T, McNaughton BL, Witter MP, Moser MB, Moser EI, http://www.ntnu.edu/kavli/research/grid-cell-data; 2006.

Savelli F, Knierim JJ. Hebbian Analysis of the Transformation of Medial Entorhinal Grid-Cell Inputs to Hippocampal Place Fields. Journal of Neurophysiology. 2010; 103(6):3167.

Savelli F, Luck J, Knierim JJ. Framing of grid cells within and beyond navigation boundaries. eLife. 2017; 6.

Savelli F, Yoganarasimha D, Knierim JJ. Influence of boundary removal on the spatial representations of the medial entorhinal cortex. Hippocampus. 2008; 18(12):1270–1282.

Savelli F, Yoganarasimha D, Knierim JJ. Influence of boundary removal on the spatial representations of the medial entorhinal cortex. Hippocampus. 2008; 18(12):1270–1282.

Schmidt-Hieber C, Häusser M. Cellular mechanisms of spatial navigation in the medial entorhinal cortex. Nature neuroscience. 2013; 16(3):325–331.

Si B, Kropff E, Treves A. Grid alignment in entorhinal cortex. Biological cybernetics. 2012; 106(8-9):483–506.

Solstad T, Moser EI, Einevoll GT. From grid cells to place cells: a mathematical model. Hippocampus. 2006; 16(12):1026–1031.

Stensola H, Stensola T, Solstad T, Frøland K, Moser MB, Moser EI. The entorhinal grid map is discretized. Nature. 2012; 492(7427):72–78.

Stensola T, Stensola H, Moser MB, Moser EI. Shearing-induced asymmetry in entorhinal grid cells. Nature. 2015; 518(7538):207–212.

Stepanyuk A. Self-organization of grid fields under supervision of place cells in a neuron model with associative plasticity. Biologically Inspired Cognitive Architectures. 2015; 13:48–62.

Tanaka K. Inferotemporal cortex and object vision. Annual review of neuroscience. 1996; 19(1):109–139.

Taube JS. Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. The Journal of neuroscience. 1995; 15(1):70–86.

Taube JS, Muller RU, Ranck JB. Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. The Journal of Neuroscience. 1990; 10(2):436–447.

Tsodyks M, Sejnowski T. Associative memory and hippocampal place cells. International journal of neural systems. 1995; 6:81–86.

Turing AM. The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 1952; 237(641):37–72.

Ujfalussy B, Kiss T, Érdi P. Parallel computational subunits in dentate granule cells generate multiple place fields. PLoS computational biology. 2009; 5(9):e1000500.

Van Strien NM, Cappaert N, Witter MP. The anatomy of memory: an interactive overview of the parahippocampal–hippocampal network. Nature Reviews Neuroscience. 2009; 10(4):272–282.

Vogels T, Sprekeler H, Zenke F, Clopath C, Gerstner W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. Science. 2011; 334(6062):1569–1573.

Wang Y, Romani S, Lustig B, Leonardo A, Pastalkova E. Theta sequences are essential for internally generated hippocampal firing fields. Nature neuroscience. 2015; 18(2):282–288.

Weber SN. spatial_patterns. GitHub. 2018; https://github.com/sim-web/spatial_patterns:2583a22.

Wernle T, Waaga T, Mørreaunet M, Treves A, Moser MB, Moser EI. Integration of grid maps in merged environments. Nature neuroscience. 2018; 21(1):92.

Wertz A, Trenholm S, Yonehara K, Hillier D, Raics Z, Leinweber M, Szalay G, Ghanem A, Keller G, Rózsa B, et al. Single-cell–initiated monosynaptic tracing reveals layer-specific cortical network modules. Science. 2015; 349(6243):70–74.

Widloski J, Fiete IR. A model of grid cell development through spatial exploration and spike time-dependent plasticity. Neuron. 2014; 83(2):481–495.

Widloski J, Fiete IR. Cortical microcircuit determination through global perturbation and sparse sampling in grid cells. bioRxiv. 2015; p. 019224.

Wilent WB, Nitz DA. Discrete place fields of hippocampal formation interneurons. Journal of neurophysiology. 2007; 97(6):4152–4161.

Wills TJ, Cacucci F, Burgess N, O'Keefe J. Development of the hippocampal cognitive map in preweanling rats. Science. 2010; 328(5985):1573–1576.

Wilson DE, Whitney DE, Scholl B, Fitzpatrick D. Orientation selectivity and the functional clustering of synaptic inputs in primary visual cortex. Nature neuroscience. 2016; 19(8):1003.

Winterer J, Maier N, Wozny C, Beed P, Breustedt J, Evangelista R, Peng Y, D'Albis T, Kempter R, Schmitz D. Excitatory microcircuits within superficial layers of the medial entorhinal cortex. Cell Reports. 2017; 19(6):1110–1116.

Yoon K, Buice MA, Barry C, Hayman R, Burgess N, Fiete IR. Specific evidence of low-dimensional continuous attractor dynamics in grid cells. Nature neuroscience. 2013; 16(8):1077–1084.

Yoon K, Lewallen S, Kinkhabwala AA, Tank DW, Fiete IR. Grid cell responses in 1D environments assessed as slices through a 2D lattice. Neuron. 2016; 89(5):1086–1099.

Zhang K. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. The journal of neuroscience. 1996; 16(6):2112–2126.

Zilli EA. Models of grid cell spatial firing published 2005–2011. Frontiers in neural circuits. 2012; 6:16.

Appendix 1

Rat trajectory

In the linear track model (one dimension, ***Figure 1***) we create artificial trajectories $x(t)$. The rat moves along a line of length L with constant velocity $v = 1\,\mathrm{cm}$ per unit time step $\Delta t = 1$. The rat always inverts its direction of motion when it hits either end of the enclosure at $-L/2$ or $L/2$. Additionally, in each unit time step it inverts its direction with a probability of $2v\Delta t/L$, resulting in a typical persistence length of $L/2$.

In the open arena model (two dimensions, ***Figures 2***, ***3*** and ***5***), we take trajectories $\mathbf{x}(t)$ from behavioral data (***Sargolini et al., 2006b***) of a rat that moved in a $1\,\mathrm{m} \times 1\,\mathrm{m}$ quadratic enclosure. The data provides coherent trajectories in intervals of 10 minutes. To get a 10 hours trajectory we concatenate 60 individual trajectories. Different trajectories in our simulations correspond to different random orders of concatenation. A 10 minute trajectory contains 30,000 locations. We update the location in every unit time step. A time step thus corresponds to $20\,\mathrm{ms}$. For simulations with a separation wall (***Figure 4***), we used a persistent random walk to constrain the motion of the rat to either side of the arena (see below).

In the model for neurons with head direction tuning (three dimensions, ***Figure 6***) we use the same behavioral trajectories as in two dimensions. To account for the experimental observation that the head direction of the animal is only roughly aligned with the direction of motion, we model the head direction as the direction of motion plus a random angle that is drawn in each unit time step from a normal distribution with standard deviation $\pi/6$.

In all dimensions and for the learning rates under consideration, we find that the precise trajectory of the rat only has a small influence on the results (see also ***Figure 2–Figure supplement 1***).

Spatial tuning of input neurons

The firing rates of excitatory and inhibitory synaptic inputs $r_i^{\mathrm{E}}, r_j^{\mathrm{I}}$ are tuned to the location $\mathbf{x}$ of the animal. In the following, we use x and y for the first and second spatial dimensions and z for the head direction. The value of x, y, z are in the range $[-L/2, L/2]$. Note that we take the interval of length L even for the dimension of head direction in order to have spatial and head direction input at the same scale. In the interpretation as head direction input, the periodic interval is to be understood as the full circle of 360 degrees.

We analyzed three different kinds of input tuning functions. Place cells (single Gaussians), several place fields (sum of multiple Gaussians) and non-localized input (Gaussians convolved with white noise). We summarize the tuning functions of neurons from the excitatory and the inhibitory population by referring to them as population P where $\mathrm{P} \in \{\mathrm{E}, \mathrm{I}\}$.

For readability we define a Gaussian of height one with standard deviation σ_{P}:

$$\mathcal{G}^{\mathrm{P}}(x) := \exp\left\{-\frac{x^2}{2\sigma_{\mathrm{P}}^2}\right\}. \tag{105}$$

The input function of the i-th neuron of population P with $N_{\mathrm{P}}^{\mathrm{f}}$ place fields per input neuron in one dimension is then given by:

$$r_i^{\mathrm{P}}(x) = \sum_{\beta=1}^{N_{\mathrm{P}}^{\mathrm{f}}} \mathcal{G}^{\mathrm{P}}\left(x - \mu_{i,\beta}^{\mathrm{P}}\right), \tag{106}$$

where $\mu_{i,\beta}^{\mathrm{P}}$ denotes the center location of field number β of input neuron i of population P. The scenario of place cell-like inputs is obtained by setting $N_{\mathrm{P}} = 1$.

For higher dimensions we define the center components as $\boldsymbol{\mu}^{\mathrm{P}}_{i,\beta} = (\mu^{\mathrm{P}}_{i,\beta,x}, \mu^{\mathrm{P}}_{i,\beta,y}, \mu^{\mathrm{P}}_{i,\beta,z})$. In two dimensions, the tuning of the i-th neuron of population P with $N^{\mathrm{f}}_{\mathrm{P}}$ place fields per input neuron is thus given by:

$$r^{\mathrm{P}}_i(\mathbf{x}) = \sum_{\beta=1}^{N^{\mathrm{f}}_{\mathrm{P}}} \mathcal{G}^{\mathrm{P}}\left(x - \mu^{\mathrm{P}}_{i,\beta,x}\right) \mathcal{G}^{\mathrm{P}}\left(y - \mu^{\mathrm{P}}_{i,\beta,y}\right) . \tag{107}$$

Here the two one dimensional Gaussians can have different standard deviations along different axes, $\sigma_{\mathrm{P},x}$ and $\sigma_{\mathrm{P},y}$, respectively. For simplicity, we constrain the resulting elliptic bell shaped curve to be aligned to the x or y axis.

In three dimensions we also consider bell-shaped tuning functions along the z-direction. However, since the head direction component is periodic we take von Mises functions that are periodic in the interval $[-L/2, L/2]$:

$$\mathcal{M}^{\mathrm{P}}(z) := \exp\left\{\left(\frac{L}{2\pi\sigma_{\mathrm{P},z}}\right)^2 \left[\cos\left(\frac{2\pi z}{L}\right) - 1\right]\right\} . \tag{108}$$

In the interpretation as head direction input, the periodic interval is to be understood as the full circle of 360 degrees. In three dimensions, the tuning of the i-th neuron of population P with $N^{\mathrm{f}}_{\mathrm{P}}$ place fields per input neurons is thus given by:

$$r^{\mathrm{P}}_i(\mathbf{x}) = \sum_{\beta=1}^{N^{\mathrm{f}}_{\mathrm{P}}} \mathcal{G}^{\mathrm{P}}\left(x - \mu^{\mathrm{P}}_{i,\beta,x}\right) \mathcal{G}^{\mathrm{P}}\left(y - \mu^{\mathrm{P}}_{i,\beta,y}\right) \mathcal{M}^{\mathrm{P}}\left(z - \mu^{\mathrm{P}}_{i,\beta,z}\right) . \tag{109}$$

The center locations $\boldsymbol{\mu}^{\mathrm{P}}$ for neurons of type P in an enclosure of side length L are drawn from a randomly distorted lattice (***Figure 2—Figure supplement 3***). First the total number of input neurons is factorized in its dimensional components $N_{\mathrm{P}} = N_{\mathrm{P},x} N_{\mathrm{P},y} N_{\mathrm{P},z}$. Then, for example along the x dimension, center locations of neurons of population P are placed equidistantly in $[-\frac{L}{2} - 3\sigma_{\mathrm{P},x}, \frac{L}{2} + 3\sigma_{\mathrm{P},x}]$. Allowing the field centers to lie a multiple of their standard deviation outside the box reduces boundary effects. Each point on the equidistant lattice is subsequently distorted with noise drawn from a uniform distribution whose range is given by the distance between two points on the undistorted lattice, i.e., $[-\frac{L}{2(N_{\mathrm{P},x}-1)}, \frac{L}{2(N_{\mathrm{P},x}-1)}]$; see ***Figure 2–Figure supplement 3***. Other dimensions are treated analogously. This procedure ensures a random but still dense coverage of the arena with few place fields. A truly random distribution of centers leads to similar results (not shown) but requires more input neurons in order to cover the arena densely. We create $N^{\mathrm{f}}_{\mathrm{P}}$ of such distorted lattices. To each input neuron we assign one center location from each of the $N^{\mathrm{f}}_{\mathrm{P}}$ lattices at random and without replacement. This guarantees that each input neuron has $N^{\mathrm{f}}_{\mathrm{P}}$ randomly located fields that together cover the arena densely.

We obtain dense non-localized input by convolving Gaussians as in ***Equations 105*** and ***107*** (with $N^{\mathrm{f}}_{\mathrm{P}} = 1$) with uniform white noise between -0.5 and 0.5. For the discretization we choose $\sigma_{\mathrm{P}}/20$ and center the Gaussian convolution kernel on an array of 8 times its standard deviation. We convolve this array with a sufficiently large array of white noise such that we only keep the values where the array of the convolution kernel is inside the array of the white noise. This way we avoid boundary effects at the edges. From the resulting function we subtract its minimum and then divide by twice the mean of the difference between the function and its minimum. This increases the signal to noise ratio and ensures that all of the inputs have a mean value of 0.5 across the arena and a minimum at 0. For each input neuron we take a different realization of white noise. This results in arbitrary tuning functions of the same autocorrelation length as the Gaussian convolution kernel. We define

the autocorrelation length as the distance at which the autocorrelation has decayed to $1/e$ of its maximum, where e is Euler's number. The above mentioned also holds for circular enclosures, only that we drop all field centers outside of a circle of radius $L/2 + 3\sigma_{\mathrm{P}}$ because they never get activated. This is not necessary but it reduces simulation time.

Learning two sides of a room independently

In ***Figure 4*** we simulated a rat that explores each half of an arena that is divided by a wall. Then the wall is removed and the animal explores the entire arena. This setup was inspired by recent experiments (***Wernle et al., 2018***) and simulations (***Rosay et al., 2018***). To simulate the two separated compartments, we use two independent sets of inputs, i.e., place cells that are randomly distributed around the entire arena (AB). One set is active when the rat explores the first compartment, the other set is active when the rat explores the second compartment. Both sets are active when the wall is removed. If we would use a single set of inputs, the grids would be merged, even before the wall is removed. The excitatory synaptic weights of the two sets are normalized independently. This is important, because otherwise the synaptic weights of inputs that are only active when the rat is in compartment A would die out while the rat explores compartment B.

To constrain the motion of the rat to one side of the arena we create artificial rat trajectories as a persistent random walk with velocity v along a direction vector $(\cos\phi, \sin\phi)$, with polar angle ϕ. In each time step, Δt, a random number drawn from a normal distribution with mean 0 and standard deviation $\sqrt{4v\Delta t/L}$ is added to ϕ, resulting in a two dimensional random walk of persistence length $L/2$. Whenever the rat hits one of the boundaries, the direction vector is modified such that the angle of incidence equals the angle of reflection. We related the trajectory to behavioral times by assuming an average rat velocity of $20\,\mathrm{cm/s}$.

Boundary effects and stability of grids

The motion of the rat is not periodic. We constrained it to either a square or a circular box. The input tuning is not periodic either. Consequently, input neurons with tuning fields that lie partially outside the boundary receive less activation. This leads to boundary effects: Excitatory weights associated to fields at the boundaries grow less, because the Hebbian learning scales with the presynaptic activation. This leads to a smaller firing rate at the boundary. According to the inhibitory learning rule, the inhibitory weights of neurons that are tuned to boundary locations then also grow less. At a distance given by the width of the excitatory firing fields, the excitatory weights grow as fast as those that are far away from the boundary. If inhibition is more broadly tuned than excitation, the inhibitory input is still reduced at these locations, though. Firing fields are thus favored at a distance from the boundary that is determined by the width of the excitatory tuning, because at this location the excitation will exceed the inhibition. This preference of firing at a certain distance from the boundary competes with the preference for hexagonal firing that is induced by the interaction of excitatory and inhibitory plasticity. For place field-like input arranged on a symmetric lattice, the alignment to the boundary can be seen in the alignment of one grid axis to the boundary in a square box (***Figure 2—Figure supplement 2a***). This alignment is not an artifact of the symmetric distribution of input fields, because it is not present in a circular arena (***Figure 2—Figure supplement 2b***). The tendency to align with the boundary can be overcome by using a random distribution of input fields ((***Figure 2—Figure supplement 2c***)) and in particular by using input with more than one place field per neuron, i.e., non-localized input. Nonetheless, we observed boundary effects in all simulations when simulating for very long times.

Distribution of initial synaptic weights

In order to start with reasonable firing rates, we take the initial weights close to the values that would correspond to the fixed point weights (see also the mathematical analysis). More precisely, initially all synaptic weights are chosen from a uniform distribution. For the spreading of the distribution we take $\pm 5\%$ of the mean value. For the mean value of the excitatory weights, w_0^{E}, we typically take $w_0^{\mathrm{E}} = 1$, see ***Table 4***. We then determine the mean of the initial inhibitory weights, w_0^{I}, such that the output neuron fires on average around the target rate:

$$\mathbf{w}^{\mathrm{E}}\mathbf{r}^{\mathrm{E}} - \mathbf{w}^{\mathrm{I}}\mathbf{r}^{\mathrm{I}} = w_0^{\mathrm{E}} \sum_{i=1}^{N_{\mathrm{E}}} r_i^{\mathrm{E}} - w_0^{\mathrm{I}} \sum_{j=1}^{N_{\mathrm{I}}} r_j^{\mathrm{I}} \stackrel{!}{=} \rho_0 \,, \tag{110}$$

so

$$w_0^{\mathrm{I}} = \frac{w_0^{\mathrm{E}} \sum_{i=1}^{N_{\mathrm{E}}} r_i^{\mathrm{E}} - \rho_0}{\sum_{j=1}^{N_{\mathrm{I}}} r_j^{\mathrm{I}}} \,. \tag{111}$$

The sums are given by:

$$\sum_{i=1}^{N_{\mathrm{P}}} r_i^{\mathrm{P}} = \frac{N_{\mathrm{P}}}{A_{\mathrm{P}}} M_{\mathrm{P}} \,, \tag{112}$$

where N_{P} is the number of input neurons, M_{P} is the area under a tuning function and A_{P} is the area in which the centers of the input tuning function can lie. For the fixed point weight relation ***Equation 111*** this leads to

$$w_0^{\mathrm{I}} = \frac{w_0^{\mathrm{E}} N_{\mathrm{E}} M_{\mathrm{E}} / A_{\mathrm{E}} - \rho_0}{N_{\mathrm{I}} M_{\mathrm{I}} / A_{\mathrm{I}}} \,. \tag{113}$$

The values for A_{P} and M_{P} depend on the dimensionality of the system.

One Dimension

For Gaussian input we have:

$$M_{\mathrm{P}} = \sqrt{2\pi} N_{\mathrm{P}}^{\mathrm{f}} \alpha_{\mathrm{P}} \sigma_{\mathrm{P}}, \quad A_{\mathrm{P}} = L + 6\sigma_{\mathrm{P}} \,. \tag{114}$$

For Gaussian random field input we have:

$$M_{\mathrm{P}} = \frac{A_{\mathrm{P}}}{2}, \quad A_{\mathrm{P}} = L \,. \tag{115}$$

Two Dimensions

For Gaussian input we have:

$$M_{\mathrm{P}} = \int\int r^{\mathrm{P}}(\mu_x, \mu_y)\, \mathrm{d}\mu_x\, \mathrm{d}\mu_y = 2\pi N_{\mathrm{P}}^{\mathrm{f}} \sigma_{\mathrm{P},x} \sigma_{\mathrm{P},y}, \quad A_{\mathrm{P}} = (L + 6\sigma_{\mathrm{P},x})(L + 6\sigma_{\mathrm{P},y}) \,. \tag{116}$$

For Gaussian random field input we have:

$$M_{\mathrm{P}} = \frac{A_{\mathrm{P}}}{2}, \quad A_{\mathrm{P}} = L^2 \,. \tag{117}$$

Three dimensions

In three dimensions we use a von Mises distribution along the third dimension to account for the periodicity of the head direction angle. We thus get

$$M_{\mathrm{P}} = \int\int\int r^{\mathrm{P}}(\mu_P, \mu_y, \mu_z)\,\mathrm{d}\mu_x\,\mathrm{d}\mu_y\,\mathrm{d}\mu_z \tag{118}$$

$$= N_{\mathrm{P}}^{\mathrm{f}} 2\pi\sigma_{\mathrm{P},x}\sigma_{\mathrm{P},y}L\frac{I_0\left[\left(\frac{L}{2\pi\sigma_{\mathrm{P},z}}\right)^2\right]}{\exp\left\{(\frac{L}{2\pi\sigma_{\mathrm{P},z}})^2\right\}} \tag{119}$$

where I_0 is the modified Bessel function. The area in which the function centers can lie is given by:

$$A_{\mathrm{P}} = (L + 6\sigma_{\mathrm{P},x})(L + 6\sigma_{\mathrm{P},y})L\,. \tag{120}$$

Grid score measure

We use the grid score suggested in (***Langston et al., 2010***). More precisely, we determine the grid score of a spatial autocorrelogram – the Pearson correlation coefficients for all spatial shifts of the firing rate map against itself – in the following way: We crop a centered doughnut shape from the correlogram. To get the inner radius of the doughnut, we clip all values in the correlogram with values smaller than 0.1 to 0. We obtain the resulting clusters that are larger than 0.1 using `scipy.ndimage.measurements.label` from the *SciPy* package for *Python* with a quadratic filter structure, $((1,1,1),(1,1,1),(1,1,1))$, for a correlogram with 51×51 pixels. We use the distance from the center to the outermost pixel of the innermost cluster as the inner radius of the doughnut. For the outer radius we try 50 values, linearly increasing from the inner radius to the corner of the quadratic arena. For each of the resulting 50 doughnuts, we rotate the doughnut around the center and correlate it with the unrotated doughnut. We determine the correlation for 30, 60, 90, 120 and 150 degrees. We define the grid score as the minimum of the correlation values at 60 and 120 degrees minus the maximum of the correlation values at 30, 90 and 150 degrees. After trying all 50 doughnuts, we take the highest resulting grid score as the grid score of the cell. A hexagonal symmetry thus leads to positive values whereas a quadratic symmetry leads to negative values.

Measure for head direction tuning

To quantify the head direction tuning of a cell, we compare the head direction tuning to a uniform circular tuning, using Watson's U^2 measure. We adopted the code from (***Mégevand, 2013***). We drew $10,000$ samples, `s_HD`, from a probability distribution created from the head direction tuning array, and $10,000$ samples, `s_uniform`, from a uniform distribution and use `watson_u2(s_uniform, s_HD)` from (***Mégevand, 2013***) to quantify the degree of non-circularity. The sharper the head direction tuning, the higher the resulting value.

Measure for grid spacing on the linear track

We define the grid spacing of one dimensional grids as the location of the first non-centered peak in the autocorrelogram of the firing pattern (***Figure 1g***). For place cell-like input we obtain the grid spacing from a single simulation.

For non-localized input the grids show defects, which results in misleading peaks in the correlogram. In this case, we used the first peak of the average of 50 correlograms to get the grid spacing (***Figure 1h***). The 50 correlograms were obtained from 50 realizations that differ only in the randomness of the input function. To avoid taking a fluctuation in the correlogram as the first peak – and thus get a misleading grid spacing – we take the

maximum between $3\sigma_E$ (to cut out the center of the correlogram) and 1 (a value larger than the largest grid spacing in ***Figure 1h***).

For high values of the spatial smoothness of inhibition, σ_I, the simulation results deviate from the analytical solution. This is because for high σ_I but small σ_E the output neuron fires very sparsely, which impedes the learning. This can be readily overcome by increasing the tuning width, σ_E, of the excitatory input.

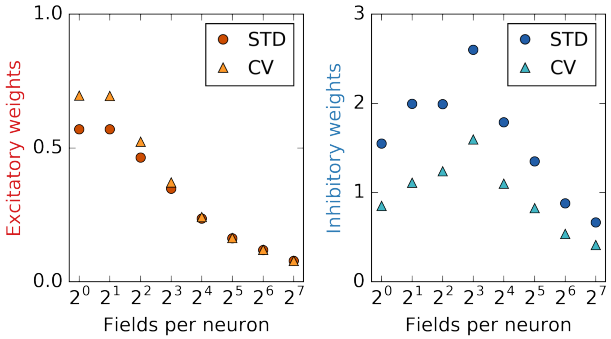

Figure 1–Figure supplement 1. Statistics of the synaptic weights. Depicted are the standard deviation (STD) and the coefficient of variation (CV) of excitatory (left) and inhibitory (right) weights as a function of the number of place fields per input neuron. The values are computed after the output neuron has established a stable grid pattern on a linear track. For excitatory weights, the CV decreases significantly with non-localized input. This indicates that different firing patterns in the output neuron are closer in 'weight space', the more non-localized the input is. In other words, to obtain a different firing pattern, the weights need to be modified by a lesser amount, i.e., the configuration and thus the output pattern is less robust: An explanation for the defects in grids with non-localized input.

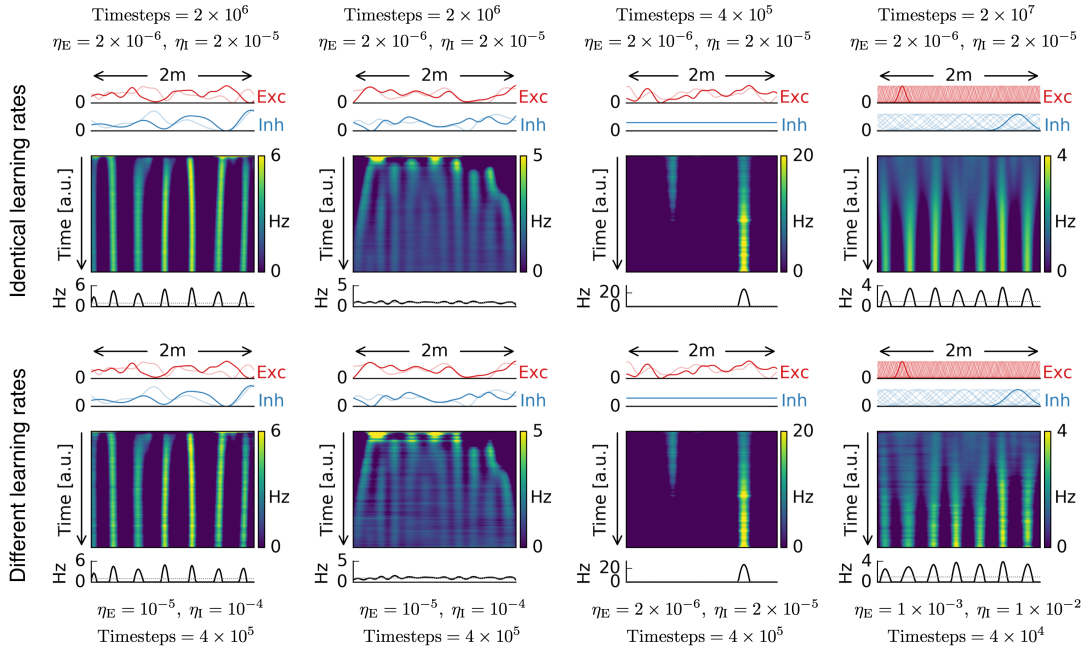

Figure 1–Figure supplement 2. Statistics of the synaptic weights. Depicted are the standard deviation (STD) and the coefficient of variation (CV) of excitatory (left) and inhibitory (right) weights as a function of the number of place fields per input neuron. The values are computed after the output neuron has established a stable grid pattern on a linear track. For excitatory weights, the CV decreases significantly with non-localized input. This indicates that different firing patterns in the output neuron are closer in 'weight space', the more non-localized the input is. In other words, to obtain a different firing pattern, the weights need to be modified by a lesser amount, i.e., the configuration and thus the output pattern is less robust: An explanation for the defects in grids with non-localized input.

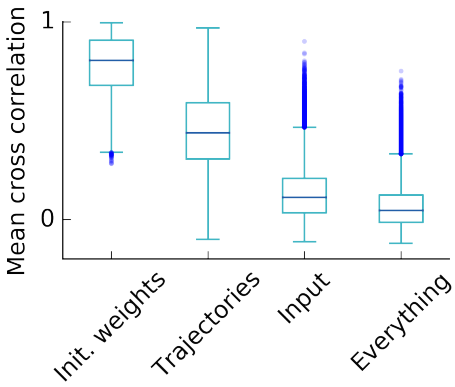

Figure 2–Figure supplement 1. Influence of random simulation parameters on the final grid pattern. Box plot of the cross correlations of the rate maps after learning of 500 simulations (i.e., $(500^2 - 500)/2 = 124750$ cross correlations). For each set of 500 simulations, only the parameter that is indicated on the $x-$axis was varied. A high cross correlation indicates that different simulations lead to similar grids and thus points towards a low influence of the varied parameter on the final grid pattern. We conclude that the influence on the final grid pattern in decreasing order is given by the parameters: Initial synaptic weights, trajectory of the rat, input tuning (i.e., locations of the randomly located input tuning curves). As expected, the correlation is lowest, if all parameters are different in each simulation (rightmost box). Each box extends from the first to the third quartile, with a dark blue line at the median. The whiskers extend from the first and third quartile by 1.5 the interquartile range. Dots show flier points. See ***Appendix 1*** for details on how trajectories, synaptic weights and inputs are varied.

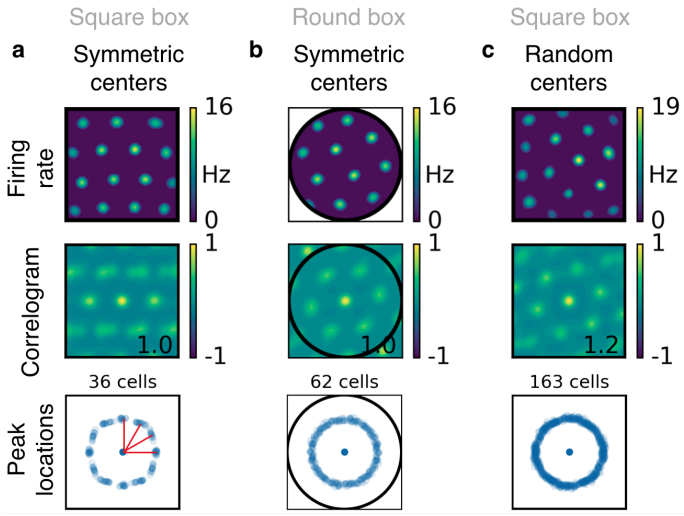

Figure 2–Figure supplement 2. Boundary effects in simulations with place field-like input. **a**) Simulations in a square box with input place fields that are arranged on a symmetric grid. From top to bottom: Firing rate map and corresponding autocorrelogram for an example grid cell; peak locations of 36 grid cells. The clusters at orientation of 0, 30, 60 and 90 degrees (red lines) indicate that the grids tend to be aligned to the boundaries. **b**) Simulations in a circular box with input place fields that are arranged on a symmetric grid. Arrangement as in **a**. The grids show no orientation preference, indicating that the orientation preference in **a** is induced by the square shape of the box. **c**) Simulations in a square box with input place fields that are arranged on a distorted grid (see ***Figure 2—Figure supplement 3***). Arrangement as in **a**. The grids show no orientation preference, indicating that the influence of the boundary on the grid orientation is small compared to the effect of randomness in the location of the input centers.

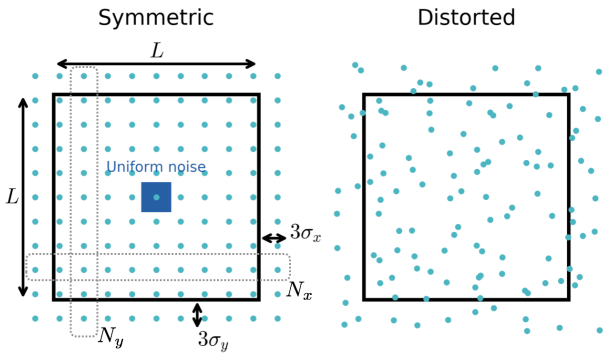

Figure 2–Figure supplement 3. Distribution of input fields. Black square box: Arena in which the simulated rat can move (side length L). Blue circles: Locations of input firing fields. To create random place field locations that cover the space densely, we use locations from a distorted lattice. To this end we first create a symmetric lattice with N_x locations along the x-direction and N_y locations along the y-direction. To reduce boundary effects, these centers can lie a certain distance outside the boundary (typically taken as threefold the width of the Gaussian input fields). We then add noise from a uniform distribution (blue square) to each location and obtain a distorted lattice (right). See ***Appendix 1*** for more details on the choice of input functions.

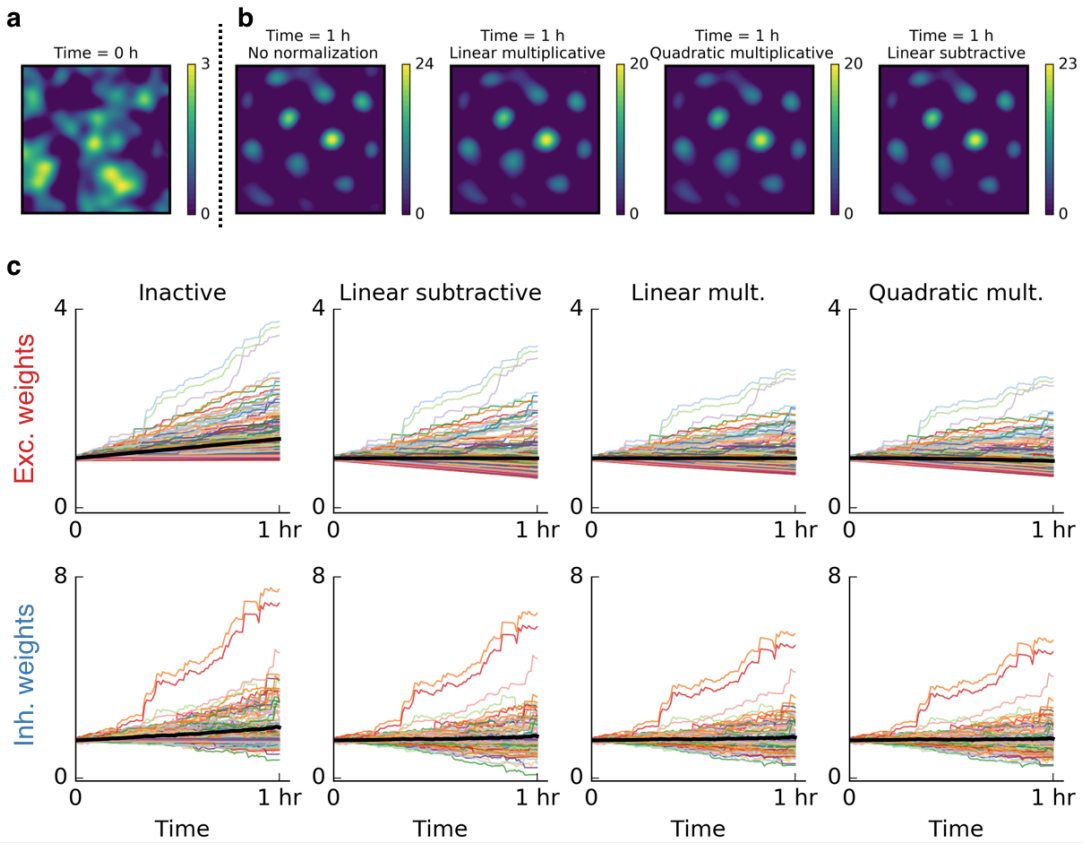

Figure 2–Figure supplement 4. Weight normalization is not crucial for the emergence of grid cells. In all simulations in the main text we used quadratic multiplicative normalization for the excitatory synaptic weights – a conventional normalization scheme. This choice was not crucial for the emergence of patterns. **a**) Firing rate map of a cell before it started exploring its surroundings. **b**) From left to right: Firing rate of the output cell after one hour of spatial exploration for inactive, linear multiplicative, quadratic multiplicative and linear subtractive normalization. **c**) Time evolution of excitatory and inhibitory weights for the simulations shown in **b**. The colored lines show 200 individual weights. The black line shows the mean over all synaptic weights. From left to right: Inactive, linear multiplicative, quadratic multiplicative and linear subtractive normalization. Without normalization, the mean of the synaptic weights grows strongest and would grow indefinitely. On the normalization schemes: Linear multiplicative normalization keeps the sum of all weights constant by multiplying each weight with a factor in each time step. Linear subtractive normalization keeps the sum of all weights roughly constant by adding or subtracting a factor from all weights and ensuring that negative weights are set to zero. Quadratic multiplicative normalization is explained in ***Methods and Materials***.

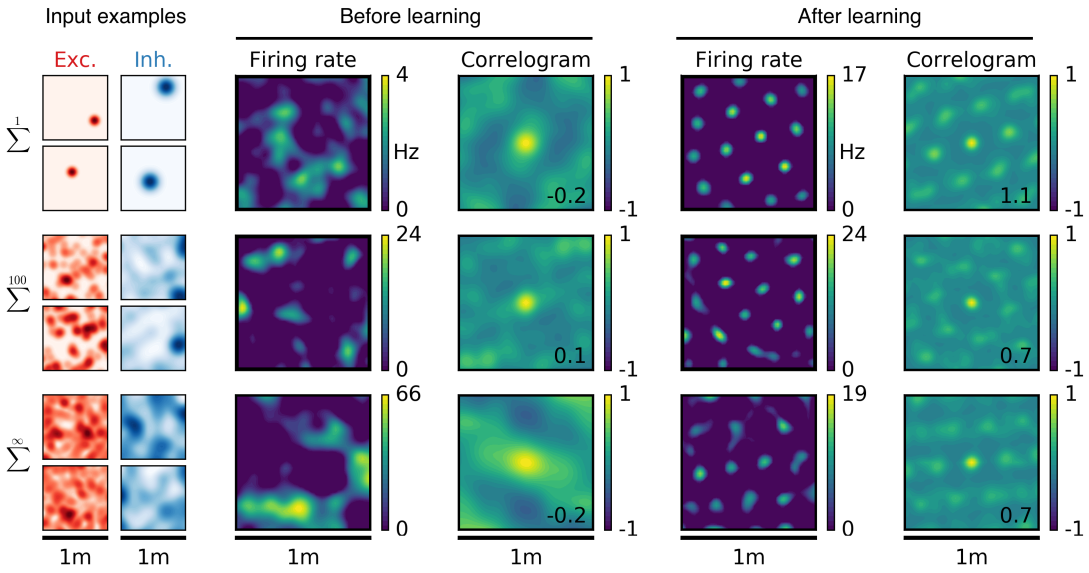

Figure 2–Figure supplement 5. Different learning rates lead to identical results. Same figure as ***Figure 2a,b,c*** but with identical learning rates for all three input scenarios: $\eta_E = 2 \times 10^{-6}$, $\eta_I = 8 \times 10^{-6}$. In order to obtain similar results as in ***Figure 2***, we need to increase the simulation times. The exploration times for the three scenarios, from top to bottom, are: 335, 10 and 50 hours. Longer simulation times are needed for inputs with lower mean firing rate, because this corresponds to an effectively lower learning rate; compare to mathematical analysis. Note that 100 fields per neurons have a larger mean firing rate than the Gaussian random fields, because the Gaussian random fields are scaled to have a mean firing rate of 0.5 Hz (see ***Methods and Materials***). As in in one dimension, scaling the excitatory and inhibitory learning rates does not have an influence on the firing pattern of the output neuron, as long as learning is sufficiently slow. Note that the patterns here are not identical to the patterns in ***Figure 2*** because of a different random initialization.

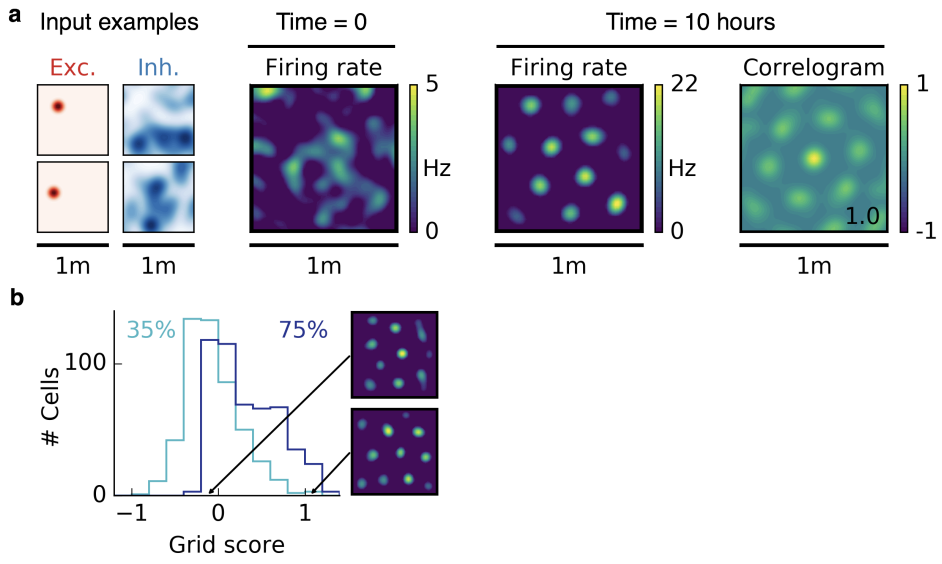

Figure 2–Figure supplement 6. Using different input statistics for different populations also leads to hexagonal firing patterns. **a**) Arrangement as in ***Figure 2a*** but with place cell-like excitatory input and sparse non-localized inhibitory input (sum of 50 randomly located place fields). A hexagonal pattern emerges, comparable to ***Figure 2a,b,c***. **b**) Grid score histogram of 500 realizations with mixed input statistics as in **a**. Arrangement as in ***Figure 2d***.

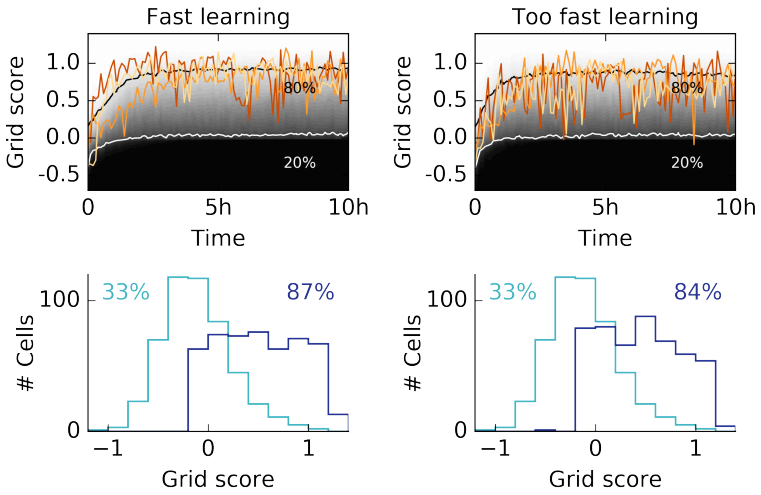

Figure 3–Figure supplement 1. Too fast learning leads to unstable grids. We showed that stable grid patterns emerge within minutes of behavioral rat trajectories (***Figure 3***) for high learning rates. Our model requires thorough spatial exploration of the rat, before significant weight changes occur. Accordingly, no stable patterns should emerge if the learning rate of the rat is too high. Left column: Same data as shown in ***Figure 3***C with three different individual traces (top). Grid score histogram of 500 realizations before (light blue) and after 10 hours of spatial exploration (dark blue). Right column: The same simulations as shown on the left, but with twice the learning rates for excitatory and inhibitory synapses. The high learning rate leads to flickering unstable grids, which is expressed in the large fluctuations in the grid score. The histogram after 10 hours of spatial exploration shows that less cells develop a hexagonal pattern, if the learning rates are very high.

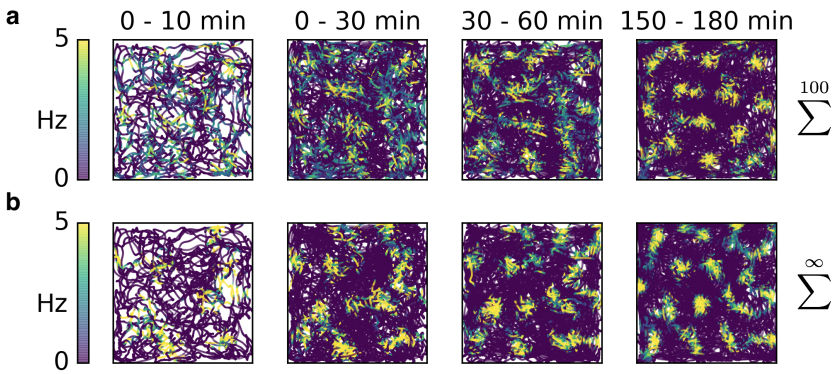

Figure 3–Figure supplement 2. Rapid development of grid patterns from non-localized input. **a**) Sparse non-localized input (sum of 100 randomly located place fields) as in ***Figure 2b***. **b**) Dense non-localized input (random function with fixed spatial smoothness) as in ***Figure 2c***. While the emergence of the final patterns takes roughly an hour – and thus longer than for place cell-like inputs (***Figure 3***) – the early firing fields are still present in the final grid, as observed in experiments (***Hafting et al., 2005***).

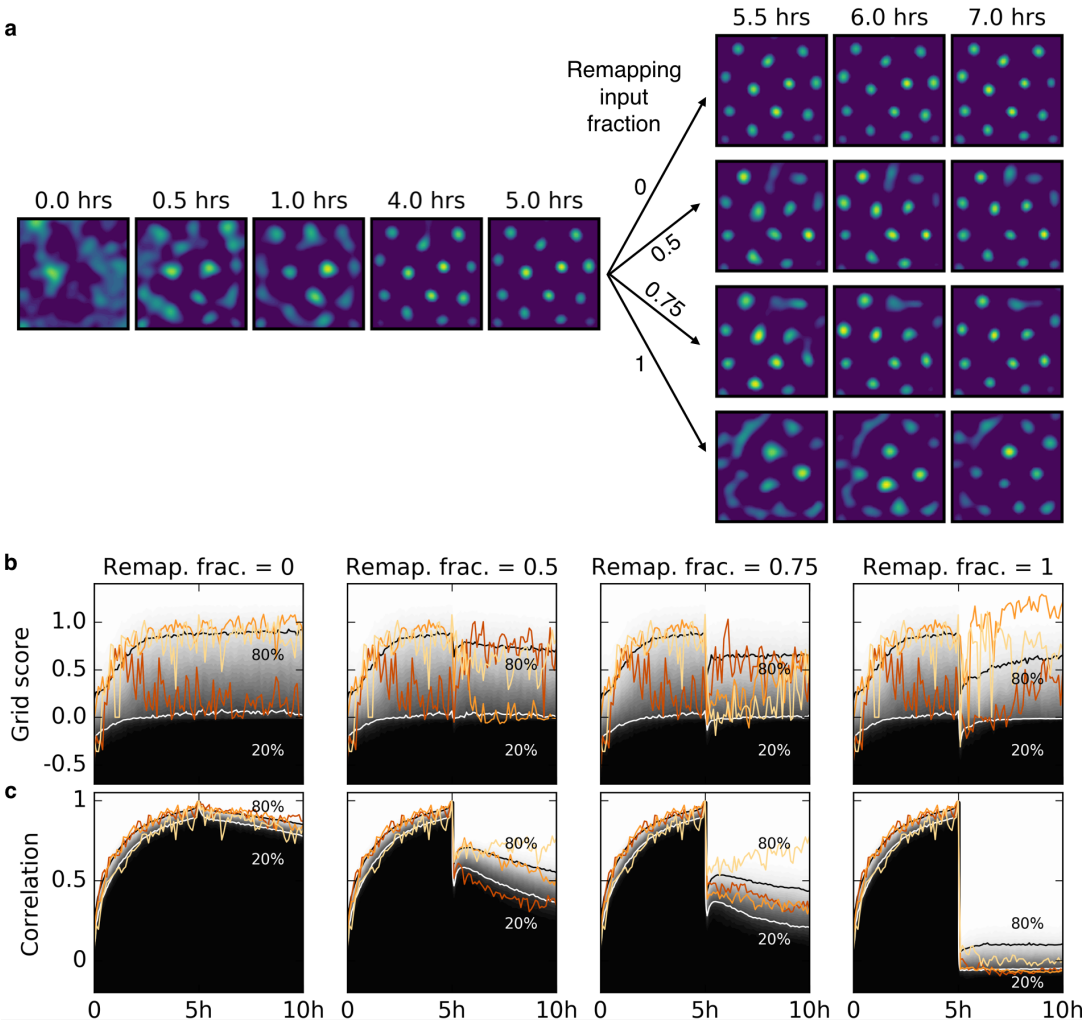

Figure 3–Figure supplement 3. Influence of input remapping on grid patterns. **a**) A grid is learned from random initial weights and place cell-like input. After 1 hour the grid pattern is apparent. After 5 hours we remap a fraction (number above arrows) of the input, i.e., the field locations of a fraction of place fields are changed to new random locations. A remapping fraction of 0 indicates that the input is unchanged and 1 indicates that all input neurons have a new place field location. The synaptic weights are not changed in this 'input remapping'. After the remapping, the rat explores and continues learning. For all four remapping scenarios, a periodic pattern is visible shortly after the remapping. For remapping fractions less than 1, it occurs faster than during the initial learning from random weights. **b**) Time courses of the grid scores for four different input remapping fractions as in **a** (Remap. frac.; shown above). The gray scale shows the cumulative histogram of the grid scores of 500 realizations (black=0, white=1). The solid white and black lines indicate the 20% and 80% percentile, respectively. Colored lines show three individual traces. The red traces correspond to the simulations shown in **a**. Varying a substantial fraction of the input often does not destroy the hexagonality of the grid patterns: Note the small dip in the 80% percentile for a remapping fraction of 0.5. **c**) Time course of the Pearson correlation of a developing rate map with the rate map of the same simulation at 5 hours (right before the input was modified) for the same simulations as in **b**, using the same color scheme and labeling. The stronger the input remapping, the lower the correlation of the grid after remapping with the grid before remapping. Note that the grid spacing is comparable for all grids, because the spatial autocorrelation length of the input is not modified during the remapping. Thus, for complete input remapping (fraction = 1) the new grid could be realigned with the old grid by a rotation and a phase shift.

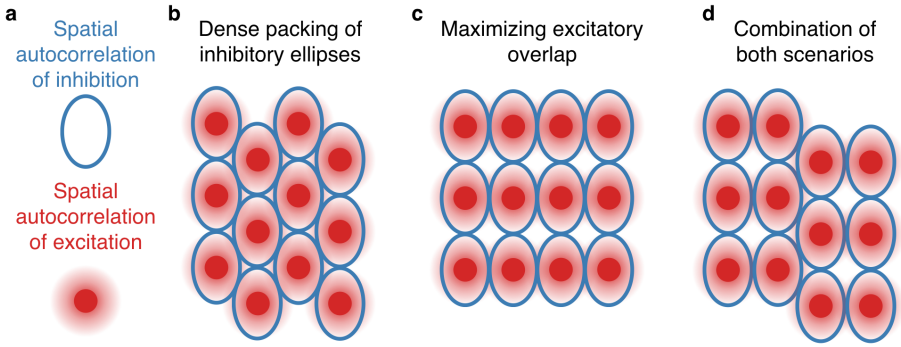

Figure 5–Figure supplement 1. Arrangement of firing fields for asymmetric input. **a**) For ellipsoidal spatial autocorrelation structures of inhibitory input (blue line), we observed band cell-like firing patterns or stretched grids (***Figure 5e***). Interestingly, the resulting patterns alternate between two different symmetries. This can be understood by two competing arrangements of ellipsoids. **b**) A dense packing of ellipsoids maximizes the area with non-zero firing and is favored by the inhibitory learning rule. This leads to stretched grids. **c**) Maximizing the overlap between excitatory input fields is favored by the excitatory learning rule and leads to quadratic grids with different periodicities along different directions. **d**) Some simulations show a combination of both patterns; compare ***Figure 5e***. The observed alignment of excitatory firing fields in **c** is particularly favored, if inhibition is very smooth along one direction. This could lead to the alignment of the head direction of individual grid fields in the simulations shown in ***Figure 6***.

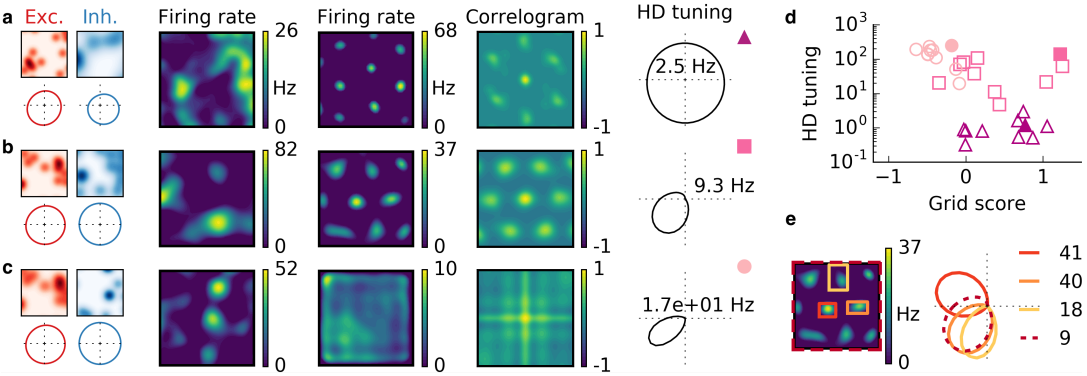

Figure 6–Figure supplement 1. Cells with combined spatial and head direction tuning with input tuning that is given by the sum of 20 randomly located Gaussian spheres. Arrangement as in ***Figure 6***.

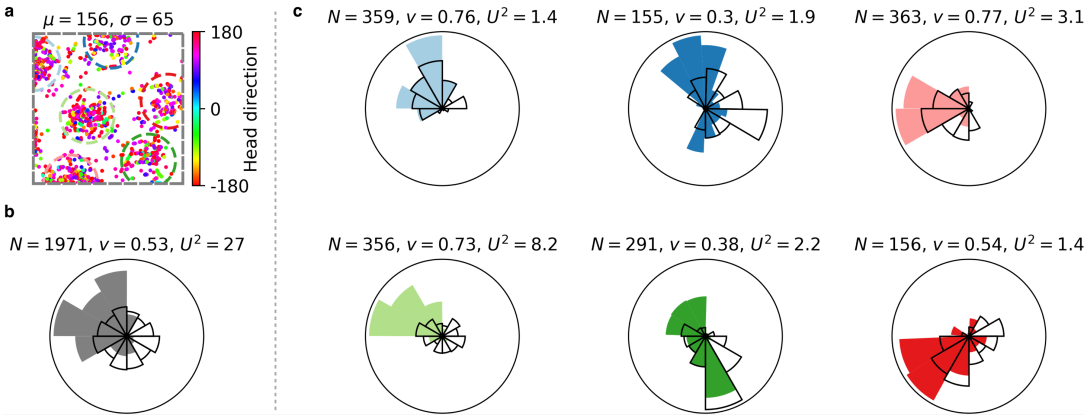

Figure 6–Figure supplement 2. Head direction tuning of individual grid fields is difficult to assess from grid cells with few firing fields. **a**) Spike locations of a grid cell color-coded with the head direction of the rat at the moment the respective spike was fired. Circular mean and circular standard deviation of head directions at spike firing ('spike head directions' in the following) shown above. **b**) Polar histogram of spike head directions (filled fan) and trajectory head directions (black fan) for all spikes of the grid cell. N: Number of spikes. v: Sum of N unit vectors whose orientation is given by the spike head directions, normalized by N. The larger this number, the more precise the head direction tuning. U^2: Watson's U^2 value (***Appendix 1***). The larger this number, the more the distribution of spike head directions deviates from the distribution of trajectory head directions. **c**) Same arrangement as in **b** but for individual grid fields. Each plot corresponds to a different grid field. The colors of the filled fans correspond to the colors of the circles around grid fields in **a**. Only spikes within these circles are considered. Individual grid fields often exhibit a sharper head direction tuning than the entire cell; compare the larger v values. However, the trajectory often exhibits a strong head direction bias; compare the directionality of black fans. The head direction tuning of a grid field tends to align with this bias; compare the smaller U^2 values. Quantitative statements about the head direction tuning of individual grid fields would be easier for cells with more firing fields, because the head direction bias of rodents is less pronounced in central parts of the arena (***Rubin et al., 2014***); compare the grid field in light green circle. From the publicly available data, we did not come to a conclusive answer on the tuning of individual grid fields. Data obtained from (***Sargolini et al., 2006a***,b).