# Cortical microcircuit determination through global perturbation and sparse sampling in grid cells

**John Widloski** [*] [†] **and Ila R. Fiete** [†]

[*]Department of Physics, and [†]Center for Learning and Memory and Department of Neuroscience, The University of Texas, Austin 70709

Submitted to Proceedings of the National Academy of Sciences of the United States of America

**Under modern interrogation, famously well-studied neural circuits such as that for orientation tuning in V1 are steadily giving up their secrets, but quite basic questions about connectivity and dynamics, including whether most computation is done by lateral processing or by selective feedforward summation, remain unresolved. We show here that grid cells offer a particularly rich opportunity for dissecting the mechanistic underpinnings of a cortical circuit, through a strategy based on global circuit perturbation combined with sparse neural recordings. The strategy is based on the theoretical insight that small perturbations of circuit activity will result in characteristic quantal shifts in the spatial tuning relationships between grid cells, which should be observable from multi- single unit recordings of a small subsample of the population. The predicted shifts differ qualitatively across candidate recurrent network mechanisms, and also distinguish between recurrent versus feedforward mechanisms. More generally, the proposed strategy demonstrates how sparse neural recordings coupled with global perturbation in the grid cell system can reveal much more about circuit mechanism as it relates to function than can full knowledge of network activity or of the synaptic connectivity matrix.**

*Significance*: **Grid cells in the mammalian brain maintain an updated record of location as animals move through space. Systems neuroscience aims to find the mechanisms of such memory and integration functions. The grid cell system offers a unique opportunity amongst cortical circuits to understand mechanism, in part because of its highly constrained response properties. We propose an experimental strategy based on global circuit perturbation combined with sparse neural recordings, that can yield surprisingly detailed information about mechanism and discriminate between distinct models currently undifferentiated by experiment. The proposed strategy demonstrates how sparse neural recordings coupled with global perturbation can reveal more about circuit mechanism as it relates to function than can full knowledge of network activity or of the synaptic connectivity matrix.**

grid cells | cortical microcircuit | neural integrator | attractor dynamics

Abbreviations: DRPS, distribution of relative phase shifts

Questions about the origin of the beautiful tuning curves often seen in sensory and cortical circuits have long consumed systems neuroscientists, both theorists who propose possible mechanisms, and experimentalists who search for them (1). Indeed, the mechanisms underlying direction tuning in the retina and cortex and orientation tuning in V1 remain unresolved and closely studied (2–6). Basic questions, like whether orientation tuning is largely attributable to selective feedforward summation or lateral interactions, are not yet settled.

Here we propose that the grid cell system provides a unique opportunity for understanding the underpinnings of computation in cortical circuits. The unusual responses of grid cells present a challenge and simultaneously, an opportunity. The challenge is to understand how such a complex cognitive response is generated; the opportunity is the availability of versatile experimental tools and a rich set of relatively detailed models (7–16) that are well-constrained by the very complexity of the grid cell response, to help meet the challenge.

The recent application of quantitative analyses to electrophysiological data reveals that the population activity of grid cells (within individual modules) is localized around a continuous low-dimensional (2D) manifold (17, 18), a finding that lends support to early models predicated on the idea of low-dimensional pattern formation through strong lateral interactions (7–9, 19, 20), as well as other models in which grid cells receive location-coded inputs and through structured feedforward connections (with the possible addition of some lateral connectivity) generate grid-patterned responses (12–14, 21).

These models are architecturally and mechanistically distinct in important ways, both large and subtle: they differ in whether grid cells perform velocity-to-location integration, in whether pattern formation originates wholly or partly within grid cells, and in the structure of their recurrent circuitry. Some of the structural differences within recurrent models which seem subtle have qualitative ramifications for how the circuit could have developed. Despite their differences, the models are difficult to distinguish on the basis of existing multiple single-unit activity records, because all of them produce grid-patterned outputs and exhibit approximate 2D continuous attractor dynamics. Worse, as we discuss at the end, neither complete single neuron-resolution activity records nor complete single synapse-resolution weight matrices will be sufficient to distinguish between proposed mechanisms.

We show how it is nevertheless possible to gain surprisingly detailed information about the grid cell circuit from a feasible experimental strategy that depends on global circuit perturbation and sparse neural recording. In this context, global means circuit-wide not brain-wide. The proposed strategy can allow the experimenter to discriminate between various distinct candidate mechanisms that are currently undifferentiated by experiment.

## Results

**Experimentally undifferentiated grid cell models.** Let us begin by considering 2D recurrent pattern forming models, in which grid cells are assumed to integrate velocity inputs and output location-coded responses. Such recurrent pattern forming models are of three main types. The first are *aperiodic* networks (9, 22), Figure 1A. In these models, activity in the cortical sheet (when neurons are appropriately rearranged – note that topography is not required in these models or in the pro-

## Reserved for Publication Footnotes

posed experiments) is grid-like and therefore periodic, but the connectivity between cells is highly localized and not periodic. In other words, connectivity does not reflect the periodicity in the activity. Taking a developmental or plasticity perspective, this network model is somewhat unusual in that strongly correlated neurons (those with the same activity phase, within or across activity bumps) are not connected as might be expected from associative learning. So if this network architecture holds in the brain, it would suggest that associative learning is curtailed once pattern formation occurs. From a functional viewpoint, aperiodic networks can require careful tuning of input at the network edges to accurately integrate their velocity inputs ( (9), but not so in (22)).

The second type are *fully periodic* networks, Figure 1B (7, 8, 19, 22–24). In these, network connectivity is itself periodic (the network has periodic boundary conditions on a rhombus), and the connectivity period equals the activity period: they each have a single period over the network (left, Figure 1B). An alternative version of the fully periodic network is to consider an aperiodic network with multiple activity bumps, but in which neurons at the centers of all the activity bumps are synaptically coupled. These two views of a fully periodic network are mathematically equivalent. Developmentally, the latter may be constructed from an aperiodic network (with multiple activity bumps) by application of associative learning post pattern-formation, so that neurons with similar phase but in different bumps end up recurrently coupled (right, Figure 1B).

The third type of the recurrent pattern forming networks is *partially periodic*, Figure 1C (9). In these, as in the aperiodic networks, the bulk connectivity is local, so that connectivity does not reflect the periodicity of the population activity patterns. However, opposite edges of the cortical sheet are identified so that the network is effectively a torus. From a developmental perspective, these networks are the strangest: bulk connectivity does not reflect the periodic activity but the boundary condition requires knowledge of it (Figure S1).

Next come a variety of models in which grid cells are the result of feedforward summation of inputs that are already spatially tuned (12–14,21). Functionally, these models suggest that path integration occurs upstream of grid cells, in different low-dimensional attractor networks (12–14). (In (21), the origin of spatial tuning in the inputs is not directly modeled; if the assumed place-cell like inputs are based on path integration then the model will display low-dimensional dynamics, so we will consider the model under this assumption). Some of these models additionally include recurrent weights in the grid cell layer (13, 21). We will call all these *feedforward* models.

**A perturbation-based probe of circuit architecture.** The conceptual idea for differentiating between recurrent models of grid cells depends on multi-single unit grid cell recording before and after a global perturbation of the network. The idea is as follows. If population activity patterning in the neural sheet is due to aperiodic recurrent connections, then globally increasing the gain of recurrent inhibition or the time-constant of neurons in network models is predicted to increase the period of stable patterns in the cortical sheet, Figure S2. These effects, not predicted by linear stability analysis, exist in simulation of dynamical models (9,22) and can be analytically derived by considering nonlinear effects (Widloski and Marder, unpublished observations).

Following the global perturbation, two cells originally in adjacent peaks of the population activity pattern and thus at the same phase of the population pattern (Figure 2A, blue), no longer will be (Figure 2A-B, red). Call the shift in pattern phase between cells in neighboring peaks one quantum (Figure 2A, circle and square, red versus blue). Then the shift in pattern phase between cells previously of the same phase and separated by exactly $K$ peaks will be $K$ quanta (Figure 2A, circle and triangle, red versus blue; explicit phase plot in Figure 2B). Across all cell pairs in the population, the shifts in phase will be quantized and will reach a maximum value of $M$ quanta (or a full phase cycle, whichever is smaller), where $M$ is the number of bumps in the population pattern.

Suppose the perturbation induces at most small phase shifts between all bumps of the population pattern (that is, $\alpha M < \frac{1}{2}$, where $\alpha = |\frac{\lambda_{post}}{\lambda_{pre}} - 1|$ is the perturbation stretch factor, and $\lambda_{pre}, \lambda_{post}$ are the population pattern periods pre- and post-perturbation, respectively; see Figure S3). Then the number of peaks in the distribution of pattern phase shifts, Figure 2C, will equal twice the number of bumps in the underlying population pattern, Figure 2A. In other words, the number of peaks in the distribution of pattern phase shifts can specify the number of bumps in the population.

However, the construction of this distribution relies on experimentally difficult-to-access quantities, namely the population pattern phase for each cell. If the network were topographically organized, this would be relatively simple to extract from a snapshot of network activity. If the network is not topographically organized, it is possible to obtain estimates of phase similarity or phase distance magnitudes between cells from patterns of coactivation or correlation across snapshots of the population activity, but such a scalar activity similarity measure cannot yield 2D phase in a 2D network.

The utility of our proposed strategy arises because the distribution of shifts in the *population pattern phase* across cells is mirrored in the distribution of shifts in the *relative phase of spatial tuning* across cells (Figure 2D). We illustrate this point in 1D, but the same idea carries directly over to 2D (Figure S4). The relative spatial tuning phase is derived from the spatial tuning of simultaneously recorded cell pairs. Cell pairs with zero relative phase in their spatial tuning pre-perturbation (because they fell on the same phase of the population pattern, albeit in different bumps) will exhibit post-perturbation shifts in relative phase that, like the shifts in the population phases, will be quantized, and for small changes in population period will be proportional to the number of bumps separating that cell pair, Figure 2C-E. This predicted *distribution of relative phase shifts* (DRPS, Figure 2E) between neurons from an aperiodic network is a property of patterning in an abstract space, independent of how neurons are actually arranged in the cortical sheet.

In 2D, relative phase is a vector, measured along the two principal axes of the spatial tuning grid. The total number of bumps in the population pattern can then be read out as equal to a quarter of the product of the number of peaks in the two relative phase shift distributions (Figure S4).

**Relating network parameters to experimental parameters.** Changes in the strength of recurrent inhibition in our model can be mapped, in the biological system, into changes in the gain of inhibitory synaptic conductances. Experimentally, this perturbation may be induced by locally infusing allosteric modulators that increase inhibitory channel conductances (e.g. benzodiazipines; (25) and personal communication with C. Barry). Changes in the time-constant of our model neurons can be mapped to changes in the EPSP time-constant in the biological system. Experimentally, the EPSP time-constant is sensitive to temperature through the Arrhenius effect and can be lengthened by cooling (26–28). However, cooling affects several other single-neuron properties. To assess what to expect experimentally from a temperature perturbation and how to correctly include temperature effects

in simpler neural models, we performed network simulations with cortical Hodgkin-Huxley neurons (29) while implementing documented temperature-dependent changes in all ionic and synaptic conductances (Experimental Procedures, SI, and Figure S5). The effect of cooling on conductance amplitudes is to shrink the population period in an aperiodic network, but its effect on conductance time-constants is to expand the period. The net effect of cooling is an expansion because temperature changes have larger effects on conductance time-constants (larger Q10 factors) than amplitudes (smaller Q10 factors) (27, 30). We therefore conclude that changes in temperature are reasonable to associate with changes in the time-constant of simple neuron models. To summarize, two global experimental perturbations capable of inducing population activity period changes are neuromodulatory infusions that alter the gain of recurrent inhibition and cortical cooling (28, 31).

**Discriminating amongst recurrent architectures.** Dynamical simulations of grid cell models reveal that the effects of the global perturbation will differ across recurrent network architectures, with consequently different predictions for the DRPS. In an aperiodic network, incremental global perturbation results in incremental expansion of the population activity pattern (Figure 3A, red, and Figure S2). Thus, the DRPS envelope will gradually and linearly widen with perturbation strength, and the separation between peaks will gradually grow (Figure 3B-C, red and Figure S2).

In a partially periodic network (with aperiodic local connectivity but with opposite boundaries connected), the number of bumps in the population activity pattern is constrained to be an integer. Thus, incrementally increasing the perturbation strength should result first in no change to the population activity period, and then a sudden change when the network can accommodate an additional bump (or an additional row of bumps in 2D, assuming the pattern does not rotate as a result of the perturbation; see Discussion) (Figure 3A, purple). Thus, incremental changes in perturbation strength should result in a stepwise change in population period and in the width of the DRPS envelope (Figure 3B-C, purple). Because the number of bumps has increased by a discrete amount, as soon as the DRPS changes, it will become maximally wide.

The fine structure of the DRPS will still be multimodal. However, counting peaks to estimate the number of bumps in the underlying population pattern will result in serious underestimation: when the pattern change is not incremental, there can be large changes in phase that are then lost in the DRPS, which is cut off at the maximal phase norm of 0.5 (Figure S3 and e.g. Figure 3B, compare peaks in the solid and dashed lines for small and large perturbations, respectively).

In the fully periodic network (Figure 1C), the same global perturbations that alter the population pattern period in the other recurrent networks (Figure 1A-B) are ineffective in inducing a corresponding change (Figure 3A, blue). This is because the periodic connectivity completely fixes the period of the pattern. Thus, the global perturbation will not affect the relative phase relationships between cells, and the DRPS is predicted to remain narrow, unimodal, and peaked at zero (Figure 3B-C, blue).

**Discriminating feedforward from recurrent architectures.** If the spatial tuning pattern or pattern components are generated upstream of grid cells and inherited or combined by them through feedforward summation (12–14), then perturbing the recurrent weights or the biophysical time-constant within only the grid cell layer is predicted to leave unchanged the population activity period, preserving the spatial tuning shapes and cell-cell relationships. As a result, the DRPS should be narrow and centered at zero, as in the case of a recurrent network with fully periodic connectivity (Figure 3C, green line).

In all recurrent model networks (Figure 1A-C), the *spatial tuning period* of cells is predicted to expand with the global perturbation, which induces a change in the efficacy with which feedforward velocity inputs shift the pattern phase over time (Figure 3D and Figure S6). This expansion in spatial tuning period with global perturbation strength is predicted to hold for all three recurrent network classes, and can be used as an assay of the effectiveness of the experimental manipulation, especially when there is no shift in the DRPS.

By contrast, in feedforward models integration occurs upstream of the grid cells and thus the spatial tuning period should remain unchanged with global perturbation (Figure 3D, green line). Response amplitudes should nevertheless change in the feedforward models, thus revealing whether the attempted global perturbations are in effect.

**Experimental feasibility of proposed method.** We consider two key data limitations. First, it is not yet experimentally feasible to record from all cells in a grid module. Even a 100 cell sample would constitute a 1-10 % subsampling of the estimated module size. With present estimates that <20 % of cells in a local patch in MEC are grid cells (32), the yield would be a meager 20 grid cells. Is this sufficient to observe the predicted quantal structure in a phase shift distribution, if it were present? Fortunately, the proposed method is tolerant to severe sub-sampling of the population: a tiny random fraction of the population (10/1600 cells) can capture the essential structure of the full DRPS, Figure 4A.

Second, spatial tuning and relative phase parameters are estimated from neural responses during a random, finite exploration trajectory in which cells respond variably. Hence, spatial tuning parameters, including phase and relative phase, are only known with a degree of uncertainty. In tests that depend only on the width of the DRPS (e.g. Figure 3), this phase uncertainty is not a serious limitation.

However, more detailed questions about the number of bumps in the population pattern in an aperiodic network depend on estimating the number of DRPS peaks, and here phase estimation uncertainty can be problematic: phase uncertainty will merge together peaks in the DRPS, Figure S7. At very small perturbation strengths, the DRPS peak spacing (in the aperiodic network) increases with the stretch factor. Thus, the larger the perturbation, the more distinguishable the peaks at a fixed phase error, Figure 4B and Figure S7. Yet increasing the stretch factor is not without a tradeoff: The two-for-one relationship between number of peaks in the DRPS and the number of bumps in the population pattern per linear dimension holds when the total induced shift in phase is small for all bumps (as before, when $\alpha < \frac{1}{2M}$, with $M$ now equal to the larger of the number of bumps along the two principal axes of the population pattern), Figure 4B. At larger stretch factors, the number of peaks in the DRPS is smaller than twice the number of bumps along the corresponding dimension of the pattern, and the discrepancy can be substantial.

Fortunately, the DRPS is computed from the relative phases between cells, which remain stable in a fixed network (17) (here fixed refers to the network while a given perturbation strength is stably maintained). This stability makes it possible to gain progressively better estimates of relative phase over time even if there is substantial drift in the spatial responses of cells, by computing the relative spatial phase over short snapshots of the trajectory then averaging together the relative phase estimates from different snapshots across a

progressively longer trajectory (similar to the methods used in (17) and (33)).

To distinguish $M = 5$ bumps per linear dimension based on structure within the DRPS would require a stretch factor of no greater than $\alpha = 1/(2M) = 0.1$, and phase noise must be reduced to at least 0.02 (Figure S7). Distinguishing 7 bumps would require $\alpha \leq 0.07$ and a phase noise of smaller than about 0.01. Based on grid cell and trajectory data (accessed through http://www.ntnu.edu/kavli/research/grid-cell-data), this would require an approximately 10 (50) -minute recording (Figure 4C).

The proposed method therefore has high tolerance to subsampling and a more limited tolerance to phase uncertainty. It will require longer-than-usual but still realistic amounts of spatial trajectory data with neural recordings to obtain adequately small error in relative phase estimation to test predictions that differentiate between models.

**A decision tree for experimental design.** We lay out a decision tree with an experimental workflow for discriminating between disparate networks, all of which exhibit 2D continuous attractor dynamics (Figure 5).

The demands from experiment are to be able to stably induce a global perturbation in one grid module, and to do so at 2-3 strengths. In all the cases, the term perturbation refers to a small change that leaves the network dynamics qualitatively unchanged while affecting its quantitative properties. The data to be collected are simultaneous recordings from several grid cells as the animal explores a familiar enclosure with no proximal spatial cues over about 20 minutes or more.

First, before applying perturbations, characterize the spatial tuning (periods) of the neurons, as well as cell-cell relationships (the relative spatial tuning phase). Next, apply a series of 2-3 global perturbations of increasing strength. At each perturbation strength, characterize the spatial tuning of cells and cell-cell relationships. A change in the amplitude of the cells' response across the different perturbations signals that the perturbation is having an effect.

If further there is no change in the spatial tuning period, it follows that the perturbations produced no change in the population pattern and velocity responsiveness, thus the network must be feedforward, Figure 5 (green). Verify that cell-cell relationships remain unchanged across perturbations, as predicted for feedforward networks.

If there is a change in the spatial tuning period, characterize the cell-cell relationships in each perturbation condition. Plot the DRPS from each perturbed condition relative to the pre-perturbation condition, and obtain its width. If the DRPS width increases steadily and linearly with perturbation strength, that implies an aperiodic recurrent architecture, Figure 5 (red). If the DRPS width exhibits a step change, it is consistent with a partially periodic recurrent network, Figure 5 (purple). A DRPS that remains narrowly peaked around zero, with no change in width with perturbation strength, is consistent with a fully periodic network, Figure 5 (blue).

Finally, if the network is either aperiodic or partially periodic, the underlying population pattern has multiple bumps. The number of peaks in the DRPS for each dimension of relative phase bounds from below the quantity $2M$, where $M$ is the number of bumps in the population pattern along that dimension. When the stretch factor $\alpha$ times the number of bumps is smaller than 1/2, and if the DRPS is quantal, the number of DRPS peaks equals twice the number of population activity bumps along the corresponding dimension.

## Discussion

**Assumptions.** The predictions made here assume that the network activity pattern is stable against rotations. Rotations of the population pattern would induce large changes in the DRPS, obscuring the predicted effects of pattern period expansion in any recurrent network. The fully periodic network is not subject to rotations, but partially periodic and aperiodic networks may be. In experimental data, the cell-cell phase relationships between grid cells are indeed very stable across time and environments (17), suggesting that the population activity undergoes no rotation. It is unclear what features of the circuit stabilize the population pattern against rotation; it is possible that slight directional anisotropies in the outgoing connectivity of neurons pin its orientation.

The simplifying observation, that spatial responses may be used to estimate the DRPS, depends on other inputs not being able to overrule the new post-perturbation cell-cell relationships. For instance, external sensory inputs or hippocampal place cells that become associated with particular configurations of grid cells may keep resetting the grid networks to express old relative phase relationships. To avoid this possibility, it may be important to assess post-perturbation cell-cell relationships only in novel environments, for which there are no previously learned associations between external cues, place cell responses, and grid cell activity.

Finally, it is important to note that if in feedforward models one were to include feedback from the grid cell layer back to the spatially tuned inputs (as in (14)), the network would effectively become a type of recurrent circuit, and perturbing the grid cell layer may result in changes in grid period and cell-cell relationships.

**Prior probabilities of different grid cell models being correct.** From theoretical arguments, we believe the candidate grid cell mechanisms are not equally probable. In particular, the partially periodic model is difficult to justify from the viewpoint of grid cell development. In (22), we see that activity-dependent rules acting on spatially informative feedforward inputs can lead to the formation of a network capable of path integration and with grid cell-like tuning. The network, post-development, has aperiodic structure. Under certain conditions, if network weights continue to undergo plasticity after the network has matured enough to expresses recurrent patterning, the network can become fully periodic as neurons with the same spatial phase become wired together (Figure S8). In fact, the addition of relatively weak coupling between neurons in nearest-neighbor activity bumps is sufficient to convert an aperiodic network into what is, functionally if not topologically, a fully periodic network (Figure S8).

Thus, it is possible to imagine mechanisms for the development of the fully periodic and fully aperiodic networks. By contrast, a partially periodic network involves local connectivity which does not depend on a neuron's spatial phase, but at the same time requires some mechanism for neurons at one end of the network to link with those at the opposite end in way that depends on spatial phase, Figure S1. It is more difficult to imagine a plausible mechanism that can satisfy both constraints. By the same argument, in feedforward models, one would expect the 1D patterned inputs to grid cells to involve fully periodic or fully aperiodic 1D networks.

**Circuit inference through perturbation and sparse activity records: outlook and alternatives.** It is interesting to compare the potential of our suggested approach with that of single synapse-level circuit reconstruction (a connectomics approach). A high-quality full-circuit connectome (with signed connections) can specify the topology of the network struc-

ture. In other words, it should be possible to reveal whether the circuit is intrinsically "local" (as in the aperiodic network of Figure 1A) (22), partially periodic (with local center-surround-like connectivity and periodic boundary conditions as in Figure 1B), or fully periodic (with center-surround-like connectivity of a width that spans the entire network together with periodic boundary conditions). It may even be possible to infer the locality of structure in the aperiodic network from an unsigned connectome.

Network topology is, however, one ingredient in circuit mechanism: Determining whether the signed connections lead to activity patterning still requires a large amount of inference (for instance, converting the connections into weights and inserting the matrix into an appropriate dynamical model). Even with further inference steps, whether the network actually performs certain functions like velocity-to-position integration or only inherits them is not answerable based on connectomics data. For instance, a network with lateral interactions may generate position-dependent responses *de novo* through integration (Figure 1A-C), or may act only to further pattern inputs that are already spatially tuned (Figure 1D-E) (13, 14, 21). Despite these functional differences, both types of networks have similar connectivity and topologies.

Single neuron-resolution records of activity within a grid cell module can be fruitfully used to understand the dimensionality and relationships of neural responses, but without perturbation, inferring actual connectivity and thus mechanisms from activity is problematic (34, 35). Hence, activity records do not distinguish between different recurrent models. In short, while connectomics and large-scale recording can provide troves of useful information, they are not sufficient for discriminating between models; as we have shown here, they may also not be immediately necessary.

As we have seen, with a perturbation approach it is possible to localize where integration occurs: if the perturbed area is performing integration, the spatial tuning period is predicted to change. Generally speaking, perturbation modulates the effect of connectivity on dynamics, and the proposed readout is neural activity. This closed-loop approach allows for detailed tests of mechanistic neural models, whose very goal is to relate architecture and dynamics, in a way not easily rivaled by non-perturbative probes of connectivity or activity.

Cooling and other perturbation experiments have been performed in V1 (5,36), but they were not as revealing about underlying mechanism as might be possible in grid cells. The reason is twofold: Recurrent models of orientation tuning in V1 are ring models, which are periodic single-bump networks, thus the predicted DRPS after cooling is essentially the same as the prediction for a feedforward network. Moreover, because the orientation response does not arise from integration of a velocity input, the spatial tuning width after cooling is also not predicted to change in a substantial way for recurrent networks. These factors make it harder to discriminate recurrent from feedforward mechanisms from perturbation. The multi-bump tuning of grid cells offers a unique opportunity to use the types of perturbative approaches used in V1 (5,36), to obtain unprecedented detail on the local circuit mechanisms that support the complex tuning of cortical cells.

## Materials and Methods

**Definitions: Population phase and relative spatial phase.** Roman subscripts (e.g., $i$ and $j$) refer to individual cells. If cells are arranged topographically based on connectivity, then $i$ refers to the location (in neuron body-length units) within the population pattern of the $i$th cell. If the period of the population pattern is $\lambda_{pop}$ (again in neuron body-length units), then the population pattern phase of cell $i$th is $\phi^i_{pop} = ((i-1) \bmod \lambda_{pop})/\lambda_{pop}$ (with the arbitrary choice, made without loss of generality, that neuron 1 has phase 0).

Next, consider the spatial tuning curves of cells $i, j$. Without respect to arrangement in the cortical sheet, let $d^{ij}$ represent the offset, in meters, of the peak closest to the origin in the cross-correlation of the two spatial tuning curves, and let $\lambda$ be the spatial tuning period (in meters) of the two cells. The relative spatial phase is defined as $\delta^{ij} = (d^{ij} \bmod \lambda)/\lambda$. Phase magnitudes are based on the usual Lee metric, $|\delta| = \min(|\delta|, 1-|\delta|)$. In 2D, the transformation of $\vec{d^{ij}}$ into $\vec{\delta^{ij}}$ is identical to that described in (17) and replicated here in SI. Analogously, using the same procedure, the 2D coordinate of the $i$th cell in the cortical sheet can be transformed into $\vec{\phi^i_{pop}}$, the population phase vector. As noted in Results, $\delta^{ij}$ is easily experimentally accessible; $\phi^i_{pop}$, much less so.

**Generation of Figures.** Figure 1 is schematic. Figure 2 is generated from ideal (imposed) periodic patterns but without dynamical neural network simulations. In Figures 2, 4A,B, S3, S4, and S7 relative spatial phase is computed for convenience (to save the computational cost of generating spatial tuning curves, then deriving relative phases) from the population phases (thus, by setting $\delta^{ij} = \phi^i_{pop} - \phi^j_{pop}$). Figures 3, S2, S6, which distinguish between different recurrent architectures, are based on dynamical neural network simulations using the mature grid cell network described in SI. Briefly, the model is a network of excitatory and inhibitory neurons (except in S8 – see figure caption for details), with linear-nonlinear Poisson (LNP) spiking dynamics (9, 22). For Figure S5, we use Hodgkin-Huxley dynamics. Structured lateral interactions between neurons lead to pattern formation in the neural population. Relative spatial phases are explicitly computed from spatial tuning curves of cells, which are obtained from spike responses to 2-minute long simulated quasi-random trajectories. Velocity inputs drive shifts of the population pattern, resulting in spatially periodic tuning. Only cells from the simulation with good spatial tuning are included in the analysis of relative phase shifts: for fully and partially periodic networks, this means all cells in the network, while for aperiodic networks this means cells in the central $3/4$ of the network. Since the inhibitory and excitatory populations share similar population patterning and spatial tuning in these simulations, we made the arbitrary choice to display the inhibitory population.


1. Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148:574–591.
2. Rivlin-Etzion M, Wei W, Feller MB (2012) Visual stimulation reverses the directional preference of direction-selective retinal ganglion cell reverses the directional preference of direction-selective retinal ganglion cells. Neuron 76:518–525.
3. Kim JS, et al. (2014) Space-time wiring specificity supports direction selectivity in the retina. Nature 509:331–6.
4. Takemura Sy, et al. (2013) A visual motion detection circuit suggested by drosophila connectomics. Nature 500:175–81.
5. Ferster D, Miller KD (2000) Neural mechanisms of orientation selectivity in the visual cortex. Annu Rev Neurosci 23:441–471.
6. Sompolinsky H, Shapley R (1997) New perspectives on the mechanisms for orientation selectivity. Curr Opin Neurobiol 7:514–22.
7. Fuhs MC, Touretzky DS (2006) A spin glass model of path integration in rat medial entorhinal cortex. J Neurosci 26:4266–4276.
8. Guanella A, Kiper D, Verschure P (2007) A model of grid cells based on a twisted torus topology. International Journal of Neural Systems 17:231–240.
9. Burak Y, Fiete IR (2009) Accurate path integration in continuous attractor network models of grid cells. PLoS Comput Biol 5:e1000291.
10. Burgess N, Barry C, O'Keefe J (2007) An oscillatory interference model of grid cell firing. Hippocampus 17:801–812.
11. Hasselmo ME, Giocomo LM, Zilli EA (2007) Grid cell firing may arise from interference of theta frequency membrane potential oscillations in single neurons. Hippocampus 17:1252–71.
12. Welday AC, Shlifer IG, Bloom ML, Zhang K, Blair HT (2011) Cosine directional tuning of theta cell burst frequencies: evidence for spatial coding by oscillatory interference. J Neurosci 31:16157–76.
13. Mhatre H, Gorchetchnikov A, Grossberg S (2012) Grid cell hexagonal patterns formed by fast self-organized learning within entorhinal cortex. Hippocampus 22:320–334.
14. Bush D, Burgess N (2014) A hybrid oscillatory interference/continuous attractor network model of grid cell firing. J Neurosci 34:5065–5079.
15. Hasselmo ME, Brandon MP (2012) A model combining oscillations and attractor dynamics for generation of grid cell firing. Front Neural Circuits 6:30.
16. Navratilova Z, Giocomo LM, Fellous JM, Hasselmo ME, McNaughton BL (2012) Phase precession and variable spatial scaling in a periodic attractor map model of medial entorhinal grid cells with realistic after-spike dynamics. Hippocampus 22:772–89.
17. Yoon K, et al. (2013) Specific evidence of low-dimensional continuous attractor dynamics in grid cells. Nat Neurosci 16:1077–84.

18. Fyhn M, Hafting T, Treves A, Moser MB, Moser EI (2007) Hippocampal remapping and grid realignment in entorhinal cortex. Nature 446:190–4.
19. Burak Y, Fiete I (2006) Do we understand the emergent dynamics of grid cell activity? The Journal of Neuroscience 26:9352–9354.
20. McNaughton BL, Battaglia FP, Jensen O, Moser EI, Moser MB (2006) Path integration and the neural basis of the 'cognitive map'. Nat Rev Neurosci 7:663–678.
21. Kropff E, Treves A (2008) The emergence of grid cells: Intelligent design or just adaptation? Hippocampus 18:1256–69.
22. Widloski JE, Fiete IR (2014) A model of grid cell development through spatial exploration and spike time-dependent plasticity. Neuron 83:481–495.
23. Pastoll H, Solanka L, van Rossum MCW, Nolan MF (2013) Feedback inhibition enables theta-nested gamma oscillations and grid firing fields. Cell 77:141–154.
24. Brecht M, et al. (2014) An isomorphic mapping hypothesis of the grid representation. Philos Trans R Soc Lond B Biol Sci 369.
25. Rudolph U, Möhler H (2004) Analysis of gabaa receptor function and dissection of the pharmacology of benzodiazepines and general anesthetics through mouse genetics. Annu Rev Pharmacol Toxicol 44:475–98.
26. Katz B, Miledi R (1965) The effect of temperature on the synaptic delay at the neuromuscular junction. J Physiol 181:656–670.
27. Thompson SM, Masukawa LM, Prince DA (1985) Temperature dependence of intrinsic membrane properties and synaptic potentials in hippocampal ca1 neurons in vitro. J Neurosci 5:817–824.
28. Moser EI, Anderson P (1994) Conserved spatial learning in cooled rats in spite of slowing of dentate field potentials. J Neurosci 14:4458–4466.
29. Pospischil M, et al. (2008) Minimal hodgkin-huxley type models for different classes of cortical and thalamic neurons. Biol Cybern 99:427–41.
30. Hodgkin AL, Huxley AF, Katz B (1952) Measurement of current-voltage relatioships in the membrane of the giant axon of loligo. J Physiol 116:424–448.
31. Long MA, Fee MS (2008) Using temperature to analyse temporal dynamics in the songbird motor pathway. Nature 456:189–194.
32. Tang Q, et al. (2014) Pyramidal and stellate cell specifity of grid and border representations in layer 2 of medial entorhinal cortex. Neuron 84:1191–1197.
33. Bonnevie T, et al. (2013) Grid cells require excitatory drive from the hippocampus. Nat Neurosci 16:309–317.
34. Roudi Y, Tyrcha J, Hertz J (2009) Ising model for neural data: Model quality and approximate methods for extracting functional connectivity. Phys Rev E 79.
35. Honey CJ, et al. (2009) Predicting human resting-state functional connectivity from structural connectivity. PNAS 106:2035–2040.
36. Michalski A, Wimborne BM, Henry GH (1993) The effect of reversible cooling of cat's primary visual cortex on the responses of area 21a neurons. J Physiol 466:133–156.
37. Hafting T, Fyhn M, Molden S, Moser MB, Moser EI (2005) Microstructure of a spatial map in the entorhinal cortex. Nature 436:801–806.

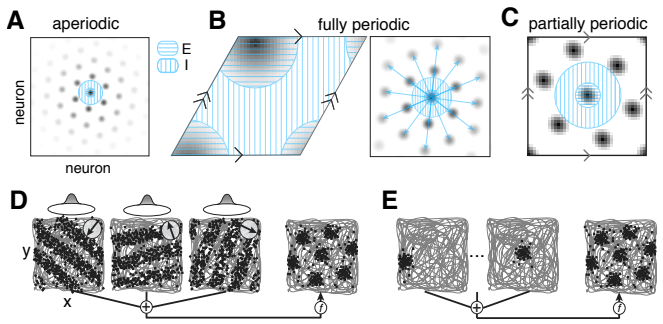


**Fig. 1. Mechanistically distinct models not distinguished by existing data.** (A-C) Recurrent 2D pattern-forming models: Activity in the cortical sheet (gray; darker indicates more activity) and the outgoing recurrent weights from a single representative cell (blue green regions centered on the cell of origin). (A) Aperiodic network: Aperiodic boundary conditions and "local" connectivity that is not determined by activity phase so that connectivity does not reflect the periodicity in activity. (B) Fully periodic network: Connectivity period equals activity period, with periodic boundary conditions on a rhombus. The two networks shown (left: single-bump network; right: multi-bump network with all bumps identified by allowing for strong recurrent connections between cells of the same activity phase) are mathematically identical. We refer to both as a single-bump network. (C) Partially periodic network: "Local" connectivity (in same sense as in (A)), with opposite edges of the cortical sheet identified so that the network boundary conditions are periodic. (D-E) Feedforward and feedforward-recurrent models: Spatially tuned (post-path integration) inputs drive grid cells (gray: spatial trajectory; spikes: red). (D) The inputs, generated from ring attractor networks (ellipses above squares) that integrate different components of animal velocity indicated by the inset compasses, have stripe-like spatial tuning (as in (13)). Feedforward summation followed by a nonlinearity produces grid-like responses (right). (E) Place-tuned inputs with selective feedforward summation, and in some models, lateral interactions, drive grid-like responses.

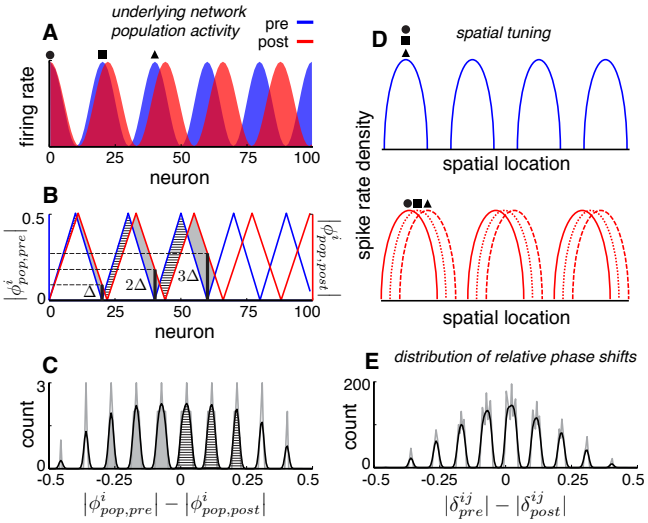


**Fig. 2. Global perturbation and phase shift analysis can reveal detailed features of population patterning.** (A) Schematic (not a dynamical neural network simulation) of population activity in a 1D aperiodic grid cell network (blue) before perturbation and (red) after a 10% pattern expansion ($\alpha = 0.1; \lambda_{pop,pre} = 20$ neurons). For illustration, cells are ordered topographically based on local connectivity and pattern expansion is centered at the left network edge. Circle, square, triangle: cells that shared the same phase no longer do post-expansion. (B) The pattern phase ($\phi^i_{pop}$; see Experimental Procedures) of cells (i) in the network, pre- (blue) and post- (red) perturbation. Cells exactly K peaks apart in the population pattern exhibit shifts in population phase equal to $K\Delta$, where $\Delta$ is the quantal phase shift. (C) The histogram of shifts, pre- to post-perturbation, in the pattern phases of all cells (n=100). Gray line: raw histogram (200 bins). Black line: smoothed histogram (convolution with 2-bin Gaussian). Negative (positive) phase shifts are from gray-shaded (vertically-striped) areas in (B). (D-E) Shift distributions for pattern phase (experimentally inaccessible) carry over to shift distributions for relative spatial tuning phase (experimentally observable). (D) The circle, square, and triangle cells originally have identical spatial tuning (schematic in blue), but post-perturbation are no longer co-active thanks to shifts in the population pattern (as in A) and thus also exhibit shifted spatial tuning curves (red). The shift for a pair is proportional to the number of activity bumps between them in the original population pattern. (E) Histogram of relative phase shifts (DRPS; gray). A relative phase shift ($\delta^{ij}$; see Experimental Procedures) between a pair of cells $i$, $j$ equals the change in their relative spatial tuning phase post-perturbation. Black: smoothed version. There are $n =$ (100 choose 2) samples because relative phase is computed pairwise.

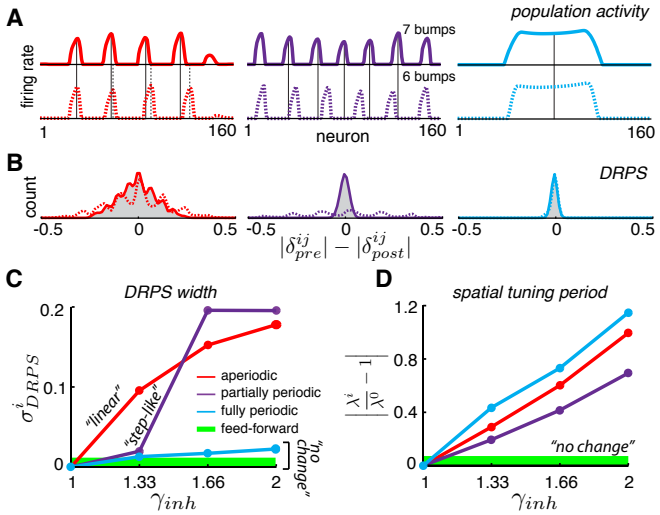


**Fig. 3. Effects of perturbation on recurrent and feedforward neural networks and predictions for experiment.** (A-B) The effect of perturbing inhibitory weights in dynamical neural network simulations of aperiodic (column 1), partially periodic (column 2), and fully periodic (column 3) recurrent architectures (see Experimental Procedures for simulation details). (A) Population pattern pre- and post-perturbation (first and second rows, with $\gamma_{inh} = 1$ and $1.33$, respectively). Vertical lines: bump centers in the unperturbed (solid) and perturbed (dotted) patterns. (B) DRPS relative to the unperturbed network. Solid line: perturbed network with $\gamma_{inh} = 1.33$; dashed line: larger perturbation of $\gamma_{inh} = 1.66$. (C) How the width of the DRPS, $\sigma^i_{DRPS}$, defined as the standard deviation of the DRPS, varies with perturbation strength. Thick green line: DRPS widths for feedforward networks (predicted, not from simulation). Note that, while the step-like shape of the DRPS width as a function of perturbation strength for the partially periodic network is general, the point at which the partially periodic network steps up will vary from trial to trial. (D) How the spatial tuning periods ($\lambda^i_x$) vary with perturbation strength in the different simulated recurrent networks and in a feedforward network (thick green line; predicted, not from simulation) (see SI for definition of spatial tuning period).

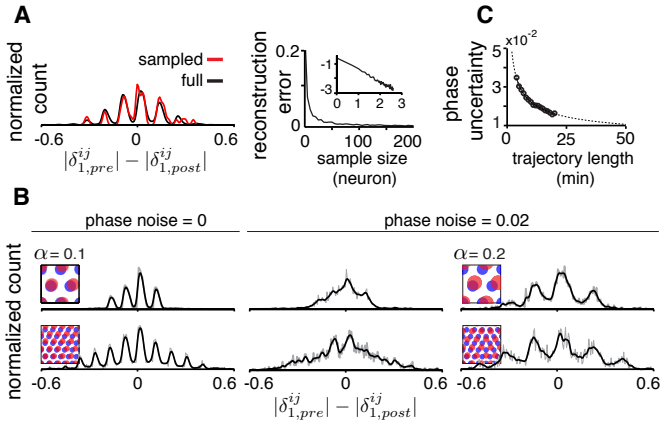


**Fig. 4. Measurement limitations and the resolvability of predictions under such constraints.** (A) Left: The quantal structure of the DRPS is apparent even in small samples of the population (black: full population DRPS; red: DRPS computed from n=10 cells; both curves smoothed with 2-bin Gaussian; bins=200). Plotted is DRPS along first principal axis of the 2D phase (see Figure S4). Right: The L2-norm difference between the full and sampled DRPS as a function of number of sampled cells. Inset: log-log scale. (B) First and second columns: DRPS (200 bins; gray line: raw; black line: smoothed with 2-bin Gaussian) for different numbers of population pattern bumps along the first principal axis of the pattern and for different amounts of phase noise (noise is sampled i.i.d. from a gaussian distribution, $\mathcal{N}(0, \sigma^2_{phase})$, and added to each component of the relative phase vector, $\vec{\delta}^{ij}$; "phase noise" is the same as $\sigma_{phase}$). Third column: Same as the second column, except for a larger stretch factor, $\alpha = 0.2$. Note that the peak-to-peak separation has increased so that the individual peaks are discernible. However, for the 5 bump network in the second row, inferring the number of bumps in the underlying population pattern would lead to an underestimate, since $M \times \alpha = 5 \times 0.2 > 1/2$. (C) The uncertainty (standard deviation) in estimating relative phase, for different amounts of data (data from (37)), from bootstrap samples of the full dataset (see SI for details). As expected, the decrease in uncertainty follows $T^{-\frac{1}{2}}$ (gray). *Parameters*: $\lambda_{pop,pre}$ = 40/3 neurons (A), = 20 neurons (B, top row), = 8 neurons (B, bottom row); $\alpha = 0.1$; $\hat{e}_1 = [1, 0]$; $\hat{e}_2 = \hat{e}_1 + 60°$; network size: $40 \times 40$ neurons.

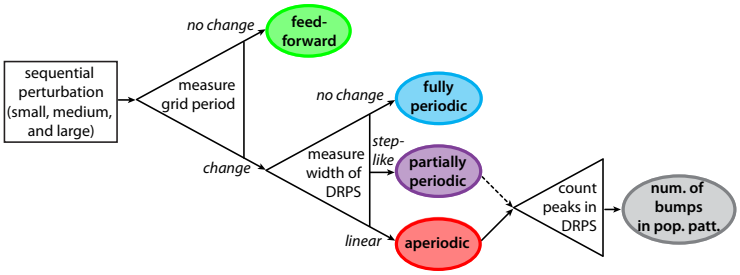


**Fig. 5. Decision tree for experimentally discriminating circuit mechanisms.** For each of three circuit perturbations of increasing strength, both spatial tuning period and relative phase shifts are measured. Recurrent networks are discriminated from feedforward and feedforward-recurrent networks by the effects of the perturbation on spatial tuning period (first open triangle). Different recurrent networks can be discriminated based on how the DRPS width varies with perturbation strength (second open triangle). The number of bumps in the multi-bump population patterns can be inferred by counting the peaks in the DRPS (third open triangle), though, for the partially periodic, only a lower bound on the number of bumps can be established (dotted line).