Reviewed Preprint
v1 • July 3, 2026
Not revised

For correspondence:
dorrellwec@gmail.com

Reviewing editor: Francesco Savelli

If Grid Cells are the Answer, What is the Question? A Review of Normative Grid Cell Theory

William Dorrell[1], James Whittington[2]

[1]Gatsby Unit, University College London, London, United Kingdom • [2]Oxford University, Oxford, United Kingdom

eLife Assessment

This Review Article puts forth a normative theory for the grid cell representations found in the entorhinal cortex. It discusses a range of theoretical models and experimental findings, organizing them around a proposed framework in which grid cells are interpreted as biologically constrained, high-fidelity codes for path integration. This framing can be potentially interesting both for readers seeking a conceptual entry point into the grid cell literature and for those more generally interested in the promises and limitations of normative theories in neuroscience. Some logical gaps and points requiring conceptual or technical clarification were nonetheless identified. Moreover, the empirical support for the path-integration account is not yet as definitive as the manuscript's framing sometimes suggests. The review would thus be strengthened by clearer justification of key arguments and fuller discussion of biological complexities, model limitations, and competing interpretations. Some stylistic choices in how arguments and literature are sometimes rhetorically framed may lessen the review's appeal for key segments of its intended audience.

https://doi.org/10.7554/eLife.111058.1.sa4

Abstract

For 20 years the beautiful structure in the grid cell code has presented an attractive puzzle: what computation do these representations subserve, and why does it manifest so curiously in neurons. The first question quickly attracted an answer: grid cells subserve path-integration, the ability to keep track of one's position as you move about the world. Subsequent work has only solidified this link: bottom-up mechanistic models that perform path-integration match the measured neural responses, while experimental perturbations that selectively disrupt grid cell activity impair performance on path-integration dependent tasks. A more controversial area of work has been top-down normative modelling: why has the brain chosen to compute like this? Floods of ink have been spilt attempting to build a precise link between the population's objective and the measured implementation. The holy grail is a normative link with broad predictive power which generalises to other neural systems. We review this literature and argue that, despite some controversies, the literature largely agrees that grid cells can be explained as a (1) biologically plausible (2) high fidelity, non-linearly decodable code for position that (3) subserves path-integration. As a rare area of neuroscience with mature theoretical and experimental work, this story holds lessons for normative theories of neural computations, and on the risks and rewards of integrating task-optimised neural networks into such theorising.

1 Introduction

It has been 20 years since the discovery of the most surprising single neuron response yet described: grid cell activity correlates with an animal's self-position, activating when the animal is in a hexagonal lattice of positions (Hafting et al.), fig. 1A. Perhaps even more surprising than their original discovery is the finding that the grid cells lattices come in discrete modules of which a rodent will have a handful (Stensola et al.), fig. 1C. Grid cells in the same module have receptive fields that are translated (but not rotated) versions of one another which uniformly tile the space of possible phases, fig. 1B. Finally, alongside the grid cells in layer II of medial entorhinal cortex, layer III hosts cells that fire at conjunctions of a hexagonal lattice of positions and a particular heading direction (Sargolini et al.), fig. 1D. There exists extensive additional phenomenology; but these four phenomena form a cohesive explanatory target:

P1. Hexagonal-lattice tuning curves
P2. For each grid cell there is a family of grid cells, called a module, which share the same tuning curve but translated, tiling the whole space.
P3. The grid cell code contains multiple modules with different lattices.
P4. The existence of paired conjunctive grid-heading direction cells.

Giving these striking findings our questions are clear: what do grid cells do? And why in this way?

A large body of work has convincingly answered the first question: the grid cell representation subserves path-integration. It has long been posited that the mammalian brain is capable of integrating its velocity to track self-position (Tolman), and as soon as grid cells were discovered they became the likely neural implementation (McNaughton et al.). In the intervening time the evidence has only built.

The second question is normative: why has biology chosen to perform path-integration using grid cells? Answering this question does not just satisfy curiosity; it promises principles to predict grid cell behaviour in novel situations, and the possibility that the same principles will generalise to other neural circuits. With the wealth of careful evidence that has accumulated the normative question seems well-posed and tractable. Despite this, there has been significant controversy in the field, producing a menagerie of different models whose commonalities and relative advantages are unclear.

This review seeks to clarify the normative grid cell theory literature. We proceed as follows:

1. We begin with path-integration. We recall perturbative and mechanistic evidence that links grid cells to path-integration. Then we intuitively link the existence of translated tuning curves, P2, to path-integration.
2. We then describe non-path-integrating ‘efficient coding’ theories that model grid cells as only a high-quality positional encoding, not as position codes that connect to one-another via path-integration. We contrast with some natural instantiations of efficient coding for which place cells rather than grid cells are optimal. Then we show that those efficient coding approaches that do generate hexagonal tuning curves, P1, are unable to match the modular structure: sets of grid cells with translated axis-aligned tuning curves, P2. We justify this by explaining how this feature is detrimental for an efficient code, but crucial for path-integration.
3. Next, we describe models that combine efficient coding with path-integration, and show that many classes of such models *are* able to capture the translated, axis-aligned, structure of grid cells, though most are limited to a single module. Further, we discuss the precise velocity update mechanism and the discrepancies between normative models and biology, in particular, P4.
4. Finally, we discuss how nonlinear encoding objectives differ qualitatively from linear. Only with a nonlinear objective, along with a path-integration constraint, do multiple modules of grid cells appear, matching data, P3.
5. We conclude with a unified normative view: theories that combine path-integration, nonlinear position encoding, and efficiency in the form of biological constraints (usually synaptic or neural activity energy efficiency and nonnegative firing rates) can cohesively capture P1, P2 and P3: multiple axis-aligned grid modules. Further, we sketch remaining puzzles, including regarding P4, and lessons for the future.

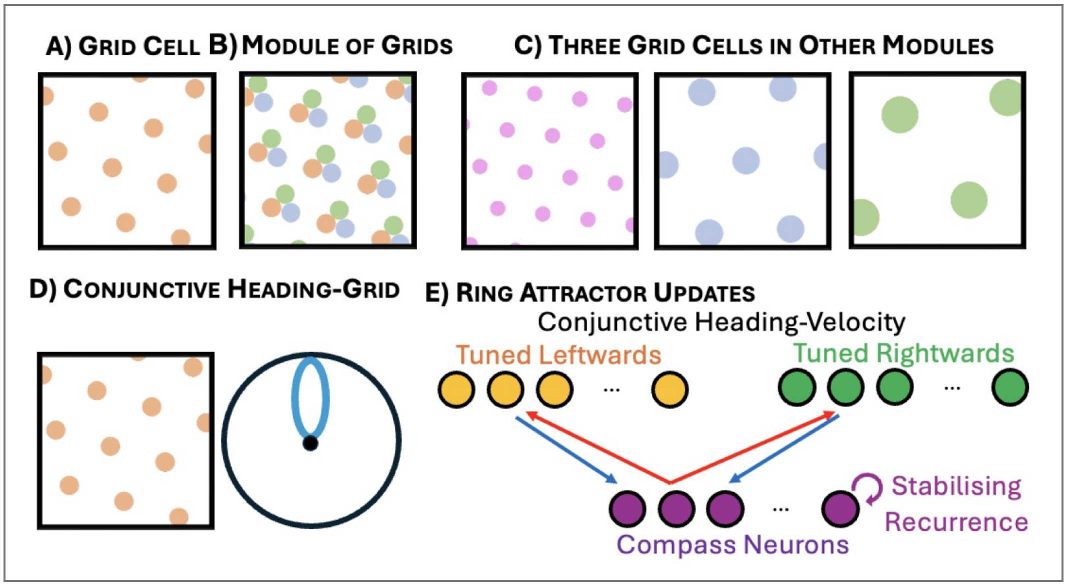

Figure 1. Structure of the grid cell code.

A: Neurons are tuned to a hexagonal lattice of positions in 2D space. **B**: They are grouped into modules: neurons in the same module have translated (but not rotated) receptive fields, and across a module they uniformly sample the phases (translations). **C**: There are only a handful of modules in one animal, each with its own lattice, and ~ 1000s neurons covering the possible phases. **D**: For each grid module there is a population of grid cells that are conjunctively tuned to both the underlying grid of the module, and a particular heading direction. **E**: These conjunctive neurons can implement path-integration by pushing the bump of neural activity around the module (Burak and Fiete), like the ring attractor in the fly central complex (Hulse and Jayaraman), using a shifted connectivity pattern: pure spatial neurons project to conjunctive neurons with the same spatial tuning profile (red connections), which project back to the spatial neurons shifted by their velocity tuning (blue connections). When the rightward neurons are more active than the leftward, this will cause the activity bump to move rightwards on the ring, implementing path-integration.

2 Grid Cells Perform Path-Integration

In this section we link the existence of a translated set of tuning curves, P2, to path-integration. We begin by reviewing evidence that grid cells are involved in path-integration. We then sketch intuitively how a translated set of tuning curves can naturally underlie path-integration.

2.1 Non-Normative Evidence Linking Grid Cells to Path-Integration

In this section we briefly review two of the key strands of evidence that suggest grid cells subserve path-integration: mechanistic models and perturbation effects.

Mechanistic Models

Mechanistic models that perform path-integration match neural observations. The most successful of these are continuous attractor neural networks (CANNs). CANNs were originally developed to model path-integration of heading direction (Skaggs et al.; Redish, Elga, and Touretzky). Their simplest implementations comprise one population of neurons that encode the animal's heading direction, and two further populations that code for conjunctions of heading direction and angular velocity, either to the left or right, fig. 1E. These conjunctive heading-velocity neurons can then be used to update the heading direction representation. First theoretically posited in the 90s, these circuits have since been verified experimentally, most beautifully in the fruit fly (Kim et al.).

Subsequent work extended CANNs to two-dimensional space, initially to model hippocampal place cells (Touretzky and Redish; Samsonovich and McNaughton; Conklin and Eliasmith). One difficulty in moving from a compact space of heading directions to an infinite space of (2D) positions is encoding the space in a finite set of neurons. Work that predated the discovery of grid cells proposed encoding space periodically, predicting lattice tuning curves but with square rather than hexagonal lattices (Samsonovich and McNaughton). Subsequent work has shown how attractor dynamics in these 2D continuous attractor circuits can naturally leads to hexagonal grid and conjunctive cells (Fuhs and Touretzky; Guanella, Kiper, and Verschure; Pastoll et al.; Burak and Fiete), and multiple modules (Kang and Balasubramanian; Khona, Chandra, and Fiete).

P4, the layer III conjunctive neurons, provide crucial evidence for these models. In a CANN each pure grid cell (i.e. tuned only to space) excites a set of conjunctive grid cells which have the same spatial tuning curve, fig. 1E, but are additional tuned to movement in particular direction, fig. 1D. In a CANN these cells implement path-integration by projecting back to the pure grid cell whose receptive field is translated along the direction of motion tuning, fig. 1E. Not only do these modelled cells match those observed in layer III, but, remarkably, measured connections between layer II and III neurons estimated from spike-time connectivity match the shifted projection pattern (Vollan et al.), presenting a ringing endorsement for the model.

There are other mechanistic models, notably the oscillatory-inteference model (Burgess, Barry, and O'keefe; Burgess; Bush and Burgess; Giocomo and Hasselmo; Hasselmo). These models were motivated by the strong theta-frequency effects in entorhinal, including grid-cell phase precession (Hafting et al.; Reifenstein et al.). However, they are unable to explain the presence of conjunctive grid cells, and more recent versions of CANN models that include theta-modulations can explain frequency effects like phase precession and theta sweeps (Vollan et al.). As such, there is strong mechanistic evidence that circuits supporting path-integration can match the measured biological effects.

Perturbation Effects

Concurrently, behavioural evidence has shown that perturbing the grid cell system impairs animals' ability to perform path-integration dependent tasks. First, lesions to the medial entorhinal cortex impair path-integration (Van Cauter et al.; Steffenach et al.). Second, disrupted spatial navigation is a known symptom of Alzheimer's disease, and this effect is thought to arise due to disruptions in grid coding in the medial entorhinal cortex. Evidence comes from genetic knock-in models of Alzheimer's which have disrupted grid cells (Jun et al.; Ying et al.), alongside impaired path-integration abilities (Ying et al.). Further, people at genetic risk of Alzheimer's show disrupted grid coding long before displaying other symptoms of Alzheimer's (Kunz et al.). Finally, and most precisely, removal of NMDA glutamate receptors from retro-hippocampal regions led to a selective disruption of grid cells while leaving other spatially selective cells intact. This perturbation caused behavioural disruptions to path integration (Gil et al.). In sum, the behavioural evidence is specific and strong.

2.2 An Intuitive Guide to the Grid Cell Solution to Path-Integration

We now outline how translational symmetry amongst tuning curves, P2, forms a natural substrate for path-integration. For simplicity, we work here with binary neurons that are either on or off, but the arguments generalise.

Path-integration involves updating your representation in response to movement. Upon taking a step, $\Delta\boldsymbol{x}$, you have to update your internal encoding of position, $\boldsymbol{g}(\boldsymbol{x})$, appropriately:

$$g\left(x\right) \rightarrow g\left(x + \mathit{\Delta} x\right) \tag{1}$$

A place cell code would make such updates very easy. The combination of the currently-active place cell and your movement specify the next representation: the place cell displaced by the movement, fig. 2A. However, this requires a place cell for every potential position, limiting how many positions you can encode.

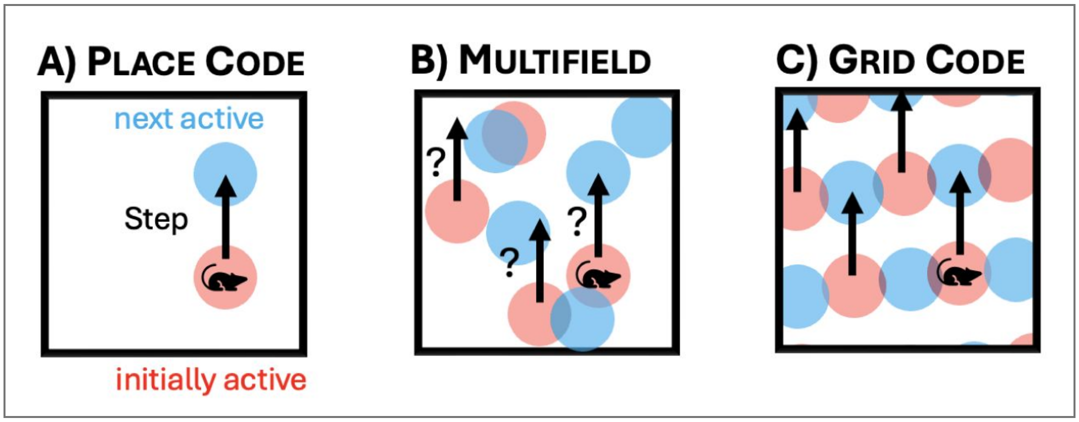

Figure 2. Path-integration with different codes A: Path-integrating with a place cell code is easy, current cell plus step uniquely determines next cell, but it is limited by the number of cells. **B**: Multifield cells improve the coding capacity but make path-integration more challenging, instead resources must be devoted to learning a mapping between unique combinations of cells. **C**: Within a grid moudle, current cell plus movement again uniquely determines the next cell: no matter which firing field of a grid cell you are in, thanks to the translational symmetry, you always know which cell to activate after a step. As such, grid cells elegantly combine the easy path-integration of place cells, with the higher capacity coding of multifield cells, and the path-integration mechanism generalises across space.

Instead imagine a cell that activates in multiple positions—a multifield place cell, fig. 2B. These neurons can improve your encoding of position: rather than giving each position a unique cell, they are given a unique combination of cells, of which there are many more, improving the capacity of the code. However, this implies a more complex path-integration mechanism: knowing that a neuron is active and which movement you make is not enough; you need to know the full set of currently active neurons, and, upon stepping north, must have a mechanism to map each combination to its neighbour one step north. This, while possible, is much more complex and specific only to the particular arrangement of firing fields.

Modules of grid cells combine the coding quality of multifield cells with simple path-integrability (Kubie and Fenton). Each position is encoded by a combination of neurons, one in each module, leading to a more informative multifield-like code. Crucially, however, the path-integration problem is separated by modules, and within each module it is simple. Knowing that one neuron in a module is active and that you make a movement north uniquely determines which neuron in that module should be active next—the one with a receptive field translated one step north, fig. 2C. By baking translational symmetry into the multifield pattern path-integration is made easy.

In short, these are the functional insights that underlie the grid cell code: a dense multifield code for position combined with easy module-wise path-integration. Indeed, in the final section, we conclude by outlining how a combination of these two functional goals with simple biological considerations (nonnegative small firing rates) leads to grid cells. For now we turn to attempts to model grid cells without reference to path-integration.

3 Grid Cells are *not* the most Efficient Code for Space

In the previous section we outlined the links between path-integration and grid cells, in particular their modular translated receptive-field structure, P2. In contrast, in this section we review what we term 'efficient coding' theories of grid cells. These normative models posit that grid cells are the most efficient encoding of position, without mentioning path-integration. We labour on these models as many have become prevalent, yet they lack the key computational feature that defines entorhinal cortex—path-integration—and do not match many critical aspects of grid cell data. We will begin by showing instantiations of efficient coding that do not generate hexagonal tuning curves. We will then discuss efficient coding that do generate hexagonal tuning curves, P1, but will show that in each case they do not capture the translated receptive fields, P2, a symptom of dropping path-integration.

3.1 Context: Many Efficient Coding Models do not generate Grid Cells

Most efficient coding theories can be decomposed into two parts. The first measures the quality of the encoding, for example, how well can a linear decoder predict where you are from your representation. The second measures or enforces the efficiency or biological plausibility of the code, for example via low nonnegative firing rates. Combinations of the two lead to some of the famous results in theoretical neuroscience, such as histogram equalisation via the fly eye's nonlinearity (Laughlin), whitening via centre-surround in retinal ganglion cells (Atick and Redlich), or sparsifcation of natural images via the V1 gabor code (Olshausen and Field).

Before studying efficient coding theories that generate grid cells, we make a useful counterpoint: very natural instantiations of efficient coding of space do not produce grids. Comparing between these theories clarifies the choices that lead to grids. Sengupta et al. use the similarity matching objective: given two inputs (e.g. positions), $\mathbf{x}$ and $\boldsymbol{x}'$, and their neural encodings, $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x}')$, this objective encourages the dot-product similarity of the representation, $\boldsymbol{g}(\boldsymbol{x})^T \boldsymbol{g}(\boldsymbol{x}')$, to match that of the input similarity structure, $\boldsymbol{x}^T\boldsymbol{x}$, through maximising the following loss:

$$\begin{gathered}\mathscr{L} = \iint \underbrace{\left(x^T x' - \alpha\right)}_{\text{Input Similarity}} \underbrace{g(x)^T g\left(x'\right)}_{\text{Representation Similarity}} \underbrace{dp\left(x\right) dp\left(x'\right)}_{\text{Occupancy probability}} \\ \text{subject to} \quad \underbrace{\|g\left(x\right)\| = 1,}_{\text{Normed activity}} \underbrace{\|g\left(x\right)\| \geq 0}_{\text{Nonnegativity}}\end{gathered} \tag{2}$$

Sengupta et al. take inputs from a compact continuous space, such as angles on a ring, and (reasonably) assume that the input similarity, $\boldsymbol{x}^T\boldsymbol{x}$, decays with distance: nearby points are similar, distant are dissimilar. From this they analytically derive that, with infinitely many neurons, place cells are the optimal nonnegative representation. This is not specific to this loss: recent work has drawn similar conclusions from an information theoretic measure of coding quality (Deighton et al.). This is somewhat natural, place cells are a very informative code, and a much simpler one than multifield codes. When there are enough neurons such that a place cell code can tile the space with sufficient resolution, these works present evidence that some efficient coding approaches prefer place cells (in section 5 we also show that place cells are preferred even with few neurons).

As such, it seems difficult for efficient coding of space alone to produce grid cells. To modify an efficient coding theory we can either change how coding quality is measured or the efficiency constraints. Many efficient coding models can be described in this way and succeed in generating hexagonal lattice tuning curves, P1. They are, however, unable to account for each module's axis-aligned translated receptive field structure, P2. We conceptually cluster these approaches into two groups, nonnegative bandpass filter models, which we review next, and clustering models, which we review in appendix A.

3.2 Grid Cells via Nonnegative Bandpass Filtering

We now review nonnegative efficient coding grid cell models that generate hexagonal lattices via nonnegative Fourier combinations, and in particular, a bandpass filter effect. These include nonnegative PCA models (Dordek et al.; Sorscher et al.; Sorscher et al.) and metric encoding models (Pettersen et al.).

Nonnegative PCA of difference-of-Gaussian Place Cells

The first set of models use an encoding objective that rewards the representation for containing high power at a critical spatial frequency, then use nonnegativity to produce a hexagonal lattice. The pivotal link in these arguments was first described by Dordek et al. who modelled grid cells as the nonnegative PCA of difference-of-Gaussian place cells, producing hexagonal receptive fields. This link is neat, but, in brief, it suffers from two major flaws. First, it relies on the use of difference-of-Gaussian place cells which are not observed; second, it fails to produce modules of translationally-symmetric grid cells.

The similarities to the approaches in section 3.1 are large; the largest difference is the choice of target, $\boldsymbol{x}$. rather than something like Gaussian place cells, whose similarity structure decays with distance, they use difference-of-Gaussian cells. Dordek et al. (later paralleled by Sorscher et al. and Sorscher et al.) nicely explain the effect of this substitution: difference-of-Gaussian cells lead to a bandpass covariance structure peaked at a particular frequency band fig. 3A, leading the optimal linearly-decodable representation to highly encode this frequency. Combining this with a lattice discretisation effect from the finite room leads to square grid cells (Dordek et al.). Finally, enforcing nonnegative firing rates changes the optimal solution from square to hexagonal grids, justified either through a triplet interaction effect (Sorscher et al.; Sorscher et al.), or the efficiency in positivising the code (Dordek et al.).

This approach has been influential with many papers using the nonnegative PCA of difference-of-Gaussian place cells (Dordek et al.; Sorscher et al.; Sorscher et al.; Schøyen et al.; Tang, Barron, and Bogacz). It has also been controversial, prompting a rebuttal (Schaeffer, Khona, and Fiete), a rebuttal to the rebuttal (Sorscher et al.), and two further rebuttals cubed (Schaeffer et al.; Schaeffer et al.). One point of disagreement lay in the finetuning of parameters required to produce grid cells: an interesting point, but clearly not fatal since the brain could simply use these parameters. A more existential threat comes from the choice of difference-of-Gaussian tuning curves. These fit hippocampal place cells less well than Gaussian curves, but, as the theoretical analysis states, are clearly vital for the production of hexagonal grid cells. Many more realistic choices of place cells don't produce grid cells in this framework (Schaeffer et al.), since they don't generate the required bandpass filter. This could be an interesting prediction about the relationship between place and grid coding, but currently there's no evidence this particular link exists.

Second, and fundamentally, these approaches do not capture the translated receptive field structure of grid modules. Instead, they produce grid cells whose orientations cluster into two groups offset at 30 degrees (Pettersen et al.) fig. 3B, a pattern that is not observed experimentally. Further, when they do produce multiple modules, the intermodule relationship

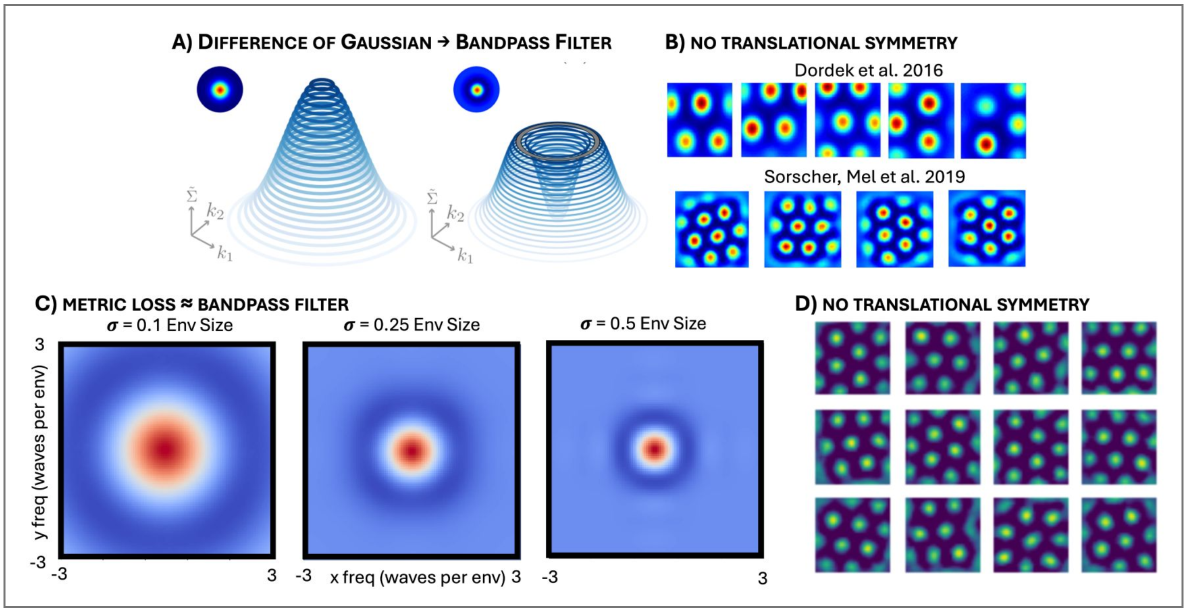

Figure 3. **Grid Cells via Bandpass Filtering.**

A: A Gaussian place cell code has a covariance whose frequency content is a smoothly-decaying Gaussian, left, but a difference-of-Gaussian code has covariance whose frequency content peaks at a non-zero frequency, figure from Sorscher et al. **B**: The grid cells that result from nonnegative PCA on difference-of-Gaussian place cells are not translationally symmetric, each population contains grid cells whose axes are rotated relative to one another (for example, the left and rightmost grid cells from dordek have lattices rotated 30° relative to one another), figures from Dordek et al. and Sorscher et al. **C**: We create a representation, $\boldsymbol{g}(\boldsymbol{x})$, that contains a single frequency, and plot the conformal loss, eq. (3), as a function of this single frequency for a few σ values. This loss is minimised (dark blue) at an intermediate value of frequency: a bandpass filtering effect. **D**: Metric encoding also produces a population of grid cells that are rotated relative to one another, figure from (Pettersen et al.).

appears to be worryingly governed by numerical discretisation effects (Sorscher et al.), nor does the framework offer an explanation of conjunctive cells, P4. Only when combined with a path-integrating task (for example by training an RNN to both path-integrate and linearly project to difference-of-Gaussian place cells) do you get axis-aligned grid cells, a topic we'll return to. Hence, this theory appears to be, at best, part of the solution.

Metric Encoding

A seemingly-distinct class of theories study a loss that encourages the 'neural metric' to match the metric of space. We will show that we can understand these as performing a similar bandpassing effect as discussed.

A metric is a function that measures distances between points. Matching a particular metric means that the distance between two points, $\mathbf{x}$ and $\Delta\boldsymbol{x}$, is preserved in the distance between the representation of those points, $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{g}(\Delta\boldsymbol{x})$, at least for a small region of space (small $\Delta\boldsymbol{x}$):

$$\| g\left(x+\Delta x\right)-g\left(x\right)\| = s\|\Delta x\| + \mathcal{O}\left(\Delta x^2\right)$$

where s is a scaling factor. Normative approaches including losses like these are common routes to grid cells often in combination with path-integration (Gao et al.; Gao et al.; Xu et al.; Pettersen et al.). Here we focus on the findings of Pettersen et al.: optimising a nonnegative unit-norm representation to preserve distances while penalising the L1 norm of the firing rates is sufficient to generate hexagonal firing fields without path-integration. The loss used is:

$$\mathcal{L} = \overbrace{\alpha\mathbb{E}_{x,x'}\left[e^{-\frac{1}{2\sigma^2}\|x-x'\|_2^2}\left(\|x-x'\|_2-\|g(x)-g(x')\|_2\right)^2\right]}^{\text{Conformal Isometry}\approx\text{Similarity Matching}} - \overbrace{(1-\alpha)\mathbb{E}\left[||g(x)||_1\right]}^{L1\ \text{Capacity Loss}} \tag{3}$$

$$\text{subject to}\quad \|g\left(x\right)\| = 1,\quad \|g\left(x\right)\| \geq 0 \tag{4}$$

The first term, called the conformal loss, forces the neural distance, $\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}')\|_2$, to match the separation in space, but only when $\mathbf{x}$ and $\boldsymbol{x}'$ are close, via the $e^{-\frac{1}{2\sigma^2}\|x-x'\|_2^2}$ weighting term. As such, it is conceptually close to similarity matching, section 3.1. In particular, the weighting sets a lengthscale, a, on the local region in which similarity matching has to occur. If σ is much larger than the environment, $e^{-\frac{1}{2\sigma^2}\|x-x'\|_2^2} \approx 1$, the loss becomes a similarity matching one, and place cells are again the optimal representation with many neurons, as in Sengupta et al., fig. 6.

When σ is smaller this loss generates hexagonal grids. We now show that this can also understood as a Fourier bandpass effect. The loss contains two biases, one that penalises high frequencies, another low frequencies, that together create a bandpass filter. The local region, encapsulated by σ, sets a lower bound on the frequency content of the code: if your code contains a component oscillating slower than $\sim\frac{1}{\sigma}$ it won't have varied meaningfully within the regions you care about, so won't decrease the loss. Conversely the similarity matching part, $(\|\mathbf{x} - \boldsymbol{x}'\|_2 - \|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}')\|_2)^2$, sets a high-frequency cutoff: the code should contain low frequencies so that nearby points are similar, and distant ones are different. We illustrate this for a neural code containing a single frequency by plotting the loss as a function of this frequency fig. 3D. The loss is minimised at a particular frequency ring (shown in dark blue) whose radius scales with the inverse of a. This is exactly the same bandpass filter of (Sorscher et al.).

Having established the bandpass filter, similar arguments to the previous section can then be used to justify how positivity and capacity constraints might lead to grid cells. Indeed, hexagonal grid cells with a single lengthscale emerge from this optimisation, with the lengthscale controlled by σ (Pettersen et al.). This is not a complete picture: for example, it is an interesting mathematical puzzle that combining this loss with an L1 capacity constraint, but not an L2, leads to hexagonal grids (Pettersen et al.). Regardless, these grid cells still suffer from the same shortcoming of other efficient coding only approaches: the grids are not aligned within the same module, rather, they feature the same loose 30° alignment as the Fourier approaches, fig. 3E. Only by adding path-integration is this effect removed.

Summary

Nonnegative combinations of Fourier components can generate hexagonal grid cells. In addition to some plausibility concerns (place cells are not well modelled by difference-of-Gaussians), without path-integration, these models are unable to reproduce the translationally symmetric modular structure that is vital for path-integration.

3.3 Conclusion: Inefficiency of Axis-Aligned Grid Cells

From this large body of work (see also clustering models in appendix A) we conclude that grid cells, despite clearly being a good code, are not the optimal efficient code of 2D space. In natural instantiations of the efficient coding problem the optimal solution are place cells (with either one or multiple fields depending on the problem, section 5). This matches unpublished findings from Tzushuan Ma's PhD thesis (Ma), and recent work that shows multifield place cells, as in the hippocampus, are a very good code (Rich, Liaw, and Lee; Harland et al.; Eliav et al.). Changing the problem in various ways can make hexagonal-lattice receptive fields optimal, either through a bandpass filter, section 3.2, or a dense packing argument, appendix A. However, it never recovers translational symmetry. This is intuitive: the grid-cell code has some glaring design flaws from a pure efficient coding perspective. The periodicity of grid cells means they identically encode points separated by the lattice symmetry, rendering a single cell unable to distinguish them. The translational symmetry within a module means that rather than helping each other to decode new points, points that are indistinguishable to one neuron are also indistinguishable to all neurons in the module! Breaking the symmetry, either by rotating and scaling the grid lattices of different neurons or removing the lattice entirely, usually improves the coding quality. As such, translated receptive fields, P2, are a key symptom of grid cells' role in path-integration, and very hard to justify from an efficient coding perspective.

4 Path-integration + Position Encoding = A Module of Grid/Place cells

In section 2.2, we outlined how grid modules' translational symmetry forms an ideal substrate for path-integration, something that purely efficient coding approaches are unable to capture. Here, we review various models that combine path-integration with an encoding loss and recover a single module of axis aligned grid cells.

4.1 Path-Integrating Models of Grid Cells

Path-Integrable Efficient Codes

Dorrell et al., similarly to unpublished work (Ma), use mathematical analysis to combine path integration with the earlier efficient coding approaches. Identically to an efficient coding approach, the representation is asked to encode space subject to some efficiency constraints. However, crucially, the code is also asked to permit path-integration: $\boldsymbol{g}(\mathbf{x} + \Delta\boldsymbol{x}) = \mathrm{f}(\boldsymbol{g}(\boldsymbol{x}), \Delta\boldsymbol{x})$ predicting next representation, $\boldsymbol{g}(\mathbf{x} + \Delta\boldsymbol{x})$, from the current representation, $\boldsymbol{g}(\boldsymbol{x})$, and velocity, $\Delta\boldsymbol{x}$. For mathematical analysis, this constraint is enforced using action-dependent weight matrices: each weight matrix has to correctly implement all transformations of the code for a given action, independent of the animal's current position:

$$g\left(x+\Delta x\right)=W\left(\Delta x\right)g\left(x\right)\quad\forall x \tag{5}$$

This constraint ensures that if the agent is at a position $\boldsymbol{x}$, it can use $\mathbf{W}(\Delta\boldsymbol{x})$ to predict where it will reach next, permitting path-integration. Further, it can be mathematically derived that this constraint forces the code to contain a small number of Fourier features, providing a basis for further analysis. Combining this with an efficient coding loss leads to either one or multiple modules depending on the choice of loss (Dorrell et al.). It does not directly explain the conjunctive grid coding, nor are action dependent weight matrices particularly biologically plausible. Both of these problems can be alleviated through action gating, a plausible scheme to implement action-dependent weight matrices as seen in other models (Logiaco, Abbott, and Escola).

Efficient Coding of Trajectories

Rebecca et al., following similar work by Waniek, formulate grid cells in a reversed manner: rather than requiring velocity to update the encoding from one timestep to the next, they instead predict velocity from each current and next encoding. From this approach, and a small number of assumptions, they show that a single hexagonal grid module is optimal for predicting velocity. While elegant, this argument suffers from using binary neurons and a discretisation of space, and struggles to naturally encapsulate multiple modules. Regardless, this alternate formulation of path-integration makes some useful novel predictions, such as how a 2D module should encode a 1D sequence.

Grids as Eigendecomposition of Transition Matrices

A set of models have formalised spatial coding via transitions on 2D graphs. For example, Stachenfeld, Botvinick, and Gershman argue that the hippocampus encodes a successor representation (a simple function of a transition matrix) of space, and that the thresholded-nonnegative eigenvectors of the successor representation (and thus the transition matrix)—which are periodic—correspond to grid cells. Later Yu, Behrens, and Burgess generalised this approach, showing that directed, rather than diffusive, transitions matrices can be used to path-integrate. However, the grid cells that emerge from eigende-composition of such transition matrices are unlike real grid cells. They exist in modules of only two neurons, many of which are not hexagonal grids but instead form bands or amorphous blobs, fig. 3C, especially in non-square rooms (Stachenfeld, Botvinick, and Gershman). Further, while one of the selling points of the successor representation theory is its sensitivity to transition statistics, pure grid cells only emerge with a diffusive policy, whereas real grid cells are more robustly hexagonal (Stensola et al.; Vollan et al.). Thus, while these models are an elegant mathematical framing, they leave several unanswered questions: why only some eigenvectors match grid behaviour; why each modelled grid module has only 2 neurons per module; why empirical grid cells are not so dramatically affected by transition statistics; and how this model could account for conjunctive grid cells.

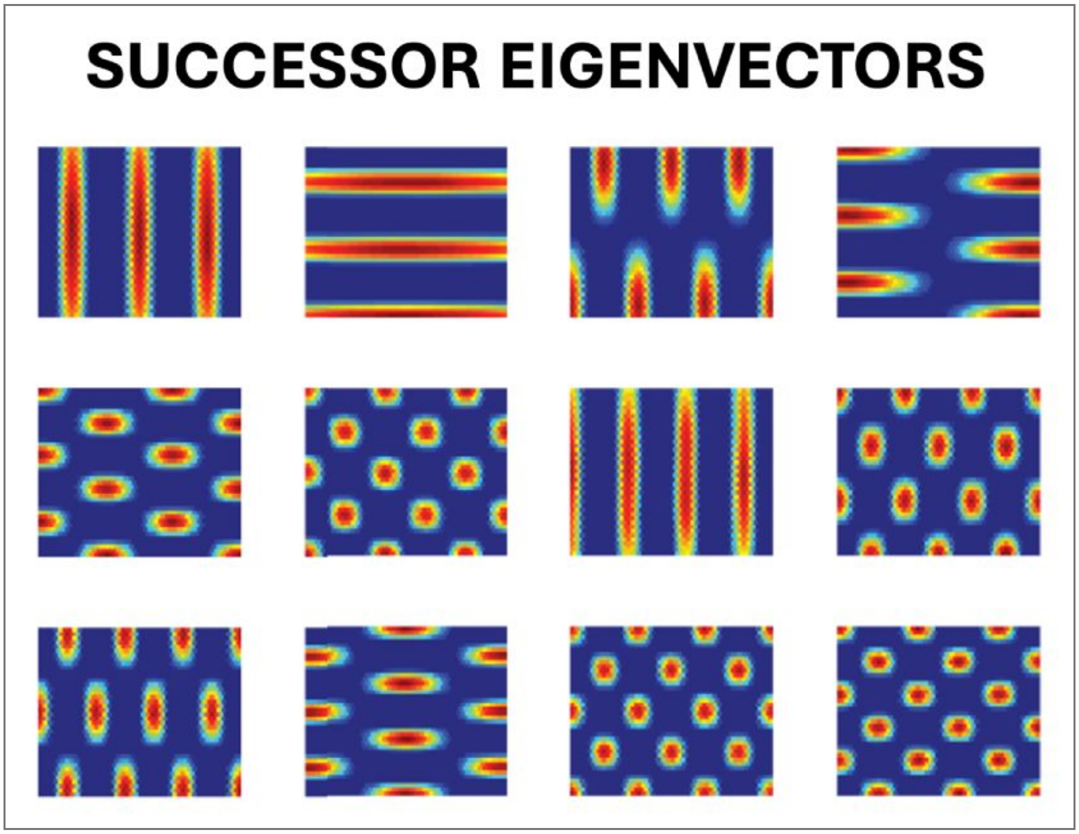

Figure 4. **Successor representation eigenvectors are poor models of grid cells, figure from Stachenfeld, Botvinick, and Gershman.**

Neural Network Models

The most common path-integration approach is to train recurrent neural networks (RNN) to path-integrate, and then to use the learnt internal representation as a model of grid cells. In its simplest instantiation, RNNs are provided a sequence of actions, and required to output the corresponding sequence of positions. This captures all three aspects of the efficient path-integrating code above: the code must path-integrate, it must distinguish different points so they can be decoded, and it must do efficiently; with low weights (if using regularisations) and with nonnegative activities (if using ReLU nonlinearities). However, the precise design choices, and the results, have varied considerably.

- Some models provide the action as a standard input to the RNN, ***a****(t)*:

$$g\left(t+1\right)=\sigma\left(Wg\left(t\right)+Ia\left(t\right)+b\right) \tag{6}$$

  while others learn a mapping between the action and the recurrent weight matrix, similar to the normative models above:

$$g\left(t+1\right)=\sigma\left(W\left(a\left(t\right)\right)g\left(t\right)\ +b\right) \tag{7}$$

- Some networks predict (*x,y*) coordinates, others Gaussian place cells or difference-of-Guassian place cells.
- Some networks use a ReLU nonlinearity, enforcing nonnegativity, others use tanh.
- Weight or activity is often constrained, either through a regularisor, or through a unit norm constraint.
- Other regularisors might be added, most often the conformal isometry loss, section 3.2.

An early pair of results suggested that path-integrating RNNs could model grid cells. Cueva and Wei trained standard RNNs to path-integrate and found grid and band-like neurons, though these grids were often square rather than hexagonal. Key choices included the use of tanh rather than ReLU nonlinearity, meaning the activities were both positive and negative, and reading out (x,y) coordinates rather than a place cell code. Concurrently, Banino et al. trained a large reinforcement learning model and showed that a feedforward layer in the network, heavily regularised by dropout, learnt somewhat griddy neurons, though there are concerns that these 'grid cells' are as grid-cell-like as low-pass filtered noise (Sorscher et al.).

Since then, the class of models that learn an action-dependent weight matrix, eq. (7), have been very successful. First studied by Issa and Zhang, who derived conditions for such a model to work, these were then used as part of a larger model of the hippocampal-entorhinal system by Whittington et al. and Whittington et al., who trained sub-networks to path-integrate, and found hexagonal modules of grid cells, though they baked the modular structure into the network. Another vein of work used the conformal isometry losses and a difference-of-Gaussian place cell readout to learn a single module of hexagonal grid cells (Gao et al.; Gao et al.; Xu et al.). Finally, Schaeffer et al. showed that training the action-dependent matrices in a ReLU RNN with a unit-norm constraint, an activity loss to reduce network capacity, a conformal loss, and a separation loss, led to multiple modules of axis aligned grid cells. Since these models do not explicitly capture the way velocity is coded by neurons, instead embedding it in the changing weight matrix, this architecture will never capture the conjunctive grid cells. Despite this, they present a ringing endorsement for the idea that optimising for a good, efficient, path-integrating code for position is sufficient for recovering grid-cells.

Path-integrating in more standard RNNs, eq. (6), can also lead to grid cells. Sorscher et al. and Sorscher et al. trained such an RNN to predict difference-of-Gaussian place cells and found a single axis-aligned module of grid cells, later supported by Tang, Barron, and Bogacz. A similar story was seen in Pettersen et al., who showed that a metric approach combined with path-integration led to a single module of axis-aligned hexagonal grid cells. Finally, Xu et al. show that a standard RNN formulation with a unit-norm, positivity, and conformal constraint is sufficient to generate a single module of grid cells, matching theoretical work (Schøyen et al.). Each of these approaches highlight a move from efficient coding-only approaches to path-integration: the coding losses alone produce hexagonal grid cells, but the axes of these grid cells are not aligned, section 3.2. Additionally asking for path-integration aligns the axes.

Each of these models demonstrates that RNNs trained to path-integrate naturally generate a module of grid cells. We will focus on two further points of discrepancies. In section 5, we will discuss how many of these models are limited to a single module. First, however, no model has reported the path-integration mechanism using conjunctive grid cells, P4, as in purely mechanistic models (Burak and Fiete), a discrepancy we will discuss next.

4.2 A Velocity Update Puzzle

In this section we review an ongoing puzzle regarding the precise grid cell velocity-update mechanism. In section 2.1 we discussed the how the pre-eminent mechanistic models, CANNs, use conjunctive neurons to path-integrate, matching connectivity measurements (Vollan et al.). Here, we outline a discrepancy between this and normative models.

Of the path-integrating theories listed in section 4.1, most do not comment on velocity-update mechanism. They either abstract away from this part of the model, or use an action-dependent weight matrix that muddies how such dependence arises. The only models which do include such effects are RNNs with standard updates, eq. (6). Surprisingly, Schøyen et al. and Pettersen et al. found that such networks learn a population of band-like cells, and that these are the neurons that seem to do the work of performing path-integration—the network can path-integrate without the grid cells! This is in contrast to a CANN model in which the grid cells are vital for the path-integration. Chu et al. elegantly explain this finding: in task-optimised RNNs the two-dimensional path-integration problem is effectively broken down into two one-dimensional problems. Along two directions a population of cells integrates motion using a standard ring attractor architecture and, due to their focus on one dimension, these cell's tuning curves resemble band cells. Then, since they are using a bandpass filter loss which specifically encourages the formation of grid cells section 3.2, a module of axis-aligned grid cells is generated from the band cells.

As such, it seems that the brain and task-optimised RNNs with standard architectural choices use fundamentally different path-integration mechanisms. Resolving this discrepancy remains an open question.

4.3 Conclusion: Path-Integration and Axis-Aligned Grid Cells

Overall, it seems well established that RNNs optimised to perform a task that includes (1) path-integration, (2) encoding of position, and (3) biological constraints (mainly nonnegativity and low firing rates) robustly learn grid cells. However, as yet the precise structure of the set of necessary constraints is unclear, especially when using a more standard RNN architecture, and the discrepancy between velocity-update mechanisms remains puzzling.

5 Only with Nonlinear Encoding are Multimodular/Combinatorial Solutions Optimal

By encoding each position with a unique combination of cells, combinatorial codes achieve higher capacity than unimodal codes, section 2.2. However, this comes at a trade-off in ease of decoding position from such a code. In particular, here we outline how 'linear' approaches cannot make use of multi-field codes and instead prefer either place cells or one module of grid cells; only with more powerful 'nonlinear' approaches do combinatorial multifield place or multimodular grid representations become optimal. Lastly, we provide a cohesive summary of the conditions in which grid cells are optimal positional representations—nonlinear efficient codes of path-integration—and review successes at predicting the optimal size and alignment of grid modules.

5.1 Combinatorial Codes Require Nonlinearity

Consider a population of N binary neurons; assigning each position its own disjoint set of cells can encode at most N positions, one per neuron. Alternatively, a combinatorial scheme which assigns each position a unique but overlapping set of cells can produce up to 2^N unique codes, enormously expanding the set of encodable positions. It is this basic fact that makes combinatorial positional codes, be that the apparently random multi-scale code in the hippocampus (Eliav et al.) or the multimodular structure of grid cells, more effective.

Yet, using such a combinatorial code requires nonlinear processing. Imagine trying to decode whether or not you are in position $\boldsymbol{x}$. In a simple place cell code this can be done linearly: simply check whether the place cell uniquely corresponding to $\boldsymbol{x}$ is on or off. It's similarly easy to decode position in a rotation of a place cell code. But in a combinatorial code, $\boldsymbol{x}$ corresponds to many place cells, and each place cell corresponds to many $\boldsymbol{x}$. Decoding $\boldsymbol{x}$ from a combinatorial code thus requires responding to a specific conjunction of place cells, and this is not something that a linear decoder can do. It requires nonlinearity.

'Functionally linear' losses prefer single grid modules

Losses that rely on linear decoding of place cells, PCA of place cells, or linear similarity objectives, such as eq. (2), struggle to profit from multimodularity. Indeed in our previous work we demonstrated that losses that are a linear function of similarity, such as eq. (2), exhibit a failure mode: they encourage further distinguishing already well distinguished positions rather than those that are poorly distinguished. This representational pressure leads to place cells or single modules of grid cells, rather than a combinatorial code (Dorrell et al.). This finding reflects a broader pattern: all prior works that use metric encoding or nonnegative PCA of difference-of-Gaussian place cells is similarly 'functionally linear', and to the best of our knowledge, all works that combine such losses with path-integration lead to a single module (Sorscher et al.; Sorscher et al.; Tang, Barron, and Bogacz; Schøyen et al.; Pettersen et al.). We note that while some models do report multiple modules using these losses, they only do so by baking a multiple modular structure into the code to begin with (Gao et al.; Gao et al.; Xu et al.), i.e. multiple modules do not emerge as the optimal code.

‘Functional nonlinearity’ profits from multiple modules

This failure model of linear losses motivated us to introduce the following ‘nonlinear’ similarity matching objective (Dorrell et al.):

$$\mathscr{L} = \iint \underbrace{X(x, x')}_{\text{Input Similarity}} \underbrace{e^{-\frac{\|g(x)-g(x')\|_2^2}{2\sigma^2}}}_{\text{Nonlinear Representational Similarity}} \underbrace{dp(x)\, dp(x')}_{\text{Occupancy probability}} \tag{8}$$

In this loss, if the representations of two points are already well distinguished ($\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x}')$ are already further apart than σ), no further gain is achieved by distinguishing them further. Instead, the code focuses its efforts on distinguishing poorly distinguished points. This encourages the formation of combinatorial codes, which make best use of the available neurons. Indeed, we know of only two normative models that derive multiple translationally symmetric modules as the optimal solution, ours (Dorrell et al.) and Schaeffer et al. Both use the nonlinear similarity matching objective we proposed, eq. (8).

In sum, we suggest that this division between ‘functionally nonlinear or linear’ losses—which correspond to linear or nonlinear decodability of position—can neatly explain which approaches generate single or multiple modules, depending on whether the loss is flexible enough to take full advantage from a combinatorial code.

5.2 The Interplay of Path-Integration, Nonlinear Decoders, and Resource Constraints

We are now in a position to summarise the optimality of different spatial representations as a function of a small number of key modelling choices: linear versus nonlinear loss functions, whether path integration is required, and neural resource constraints (i.e., the number of neurons; throughout, we assume nonnegative neural activity with unit norm).

One initially surprising finding is that, when many neurons are available, place cells are optimal independent of other considerations. In section 3.1 we related how place cells are the optimal nonnegative similarity matching code when there are more neurons than positions to be distinguished. We find that the same is true with a nonlinear similarity matching loss, and/or with an additional path-integration constraint (for example, by enforcing actionability, eq. (5), Dorrell et al.). We suggest this is because when there are enough neurons, simple place cell codes can tile the space at sufficient resolution.

When the number of neurons are scarce, under linear losses place cells are optimal without a path-integration requirement and a single module of grid cells when path-integration is required. Both these codes are *not* combinatorial as linear losses do not profit from combinatorial codes, fig. 5 top. On the other hand, with a nonlinear loss multifield (combinatorial) place cells are optimal without a path-integration requirement, while multiple modules of axis-aligned grid cells are optimal when path-integration is required, fig. 5 bottom.

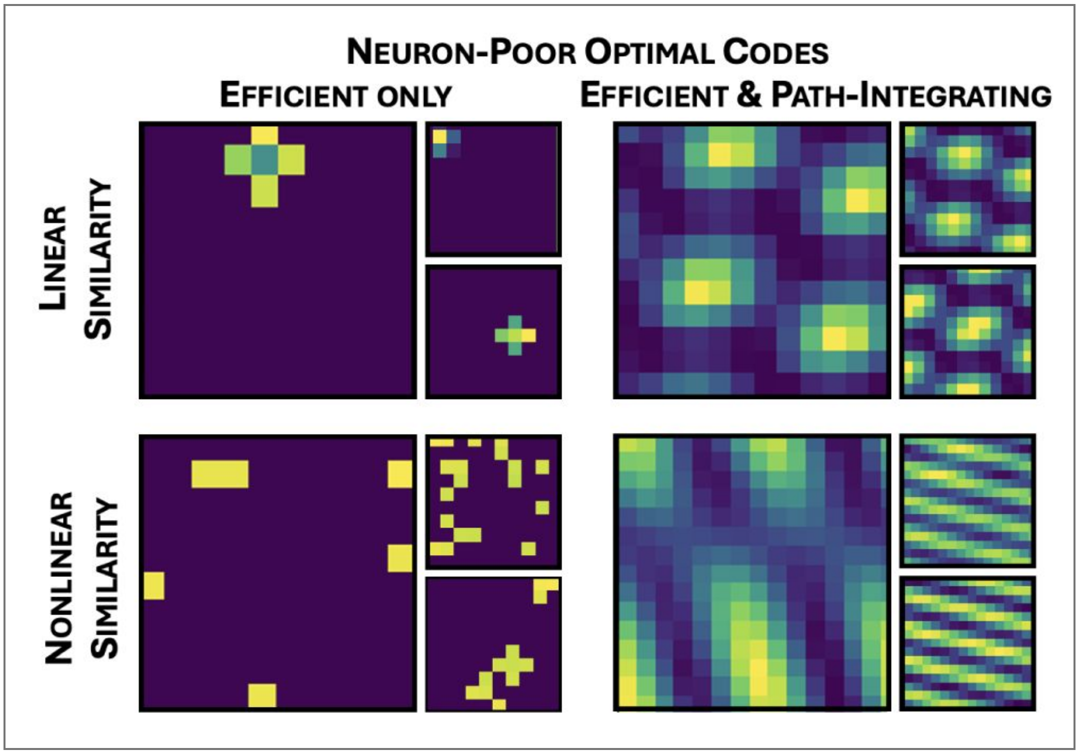

Figure 5. **A Space of Optimal Codes.** We optimise a nonnegative, unit-norm representation of position to minimise a similarity matching objective either linear, eq. (2), or nonlinear, eq. (8), with or without a path-integrating constraint, eq. (5). With more neurons than positions all choices lead to place cells (not shown). With few neurons and no path-integration (left column) we get place cells with a linear objective, and random multifields with a nonlinear objective (see also fig 15C, Dorrell et al.). Adding a path-integration constraint leads to either one grid module for the linear similarity loss, or multiple under the nonlinear loss (for more discussion, see Dorrell et al.).

5.3 Efficient Coding using Multimodular Codes

We have discussed how combining low nonnegative firing rates with a sufficiently flexible nonlinear decoding and path-integration leads to multiple modules of translationally symmetric grid cells. We now consider one final normative question: how should these modules actually be structured? What lattice should they use (e.g. square or hexagon)? What should the relative size and orientation between modules be? And how many neurons per module?

The first forays in tackling these question assumed a multimodular structure and then optimised the remaining parameters to maximise the mutual information between neural activity and position, through proxies such as the Fisher information. Having demonstrated that a multimodular grid code encodes space with a higher accuracy than a place cell code (Sreenivasan and Fiete; Mathis, Herz, and Stemmler), it was found that, of all lattice choices, hexagonal lattices were optimal (Mathis, Herz, and Stemmler; Mathis, Herz, and Stemmler). Subsequent related works derived similar results(Stemmler, Mathis, and Herz; Wei, Prentice, and Balasubramanian) and emphasised the effect of independent per-module noise (Towse et al.). Further, the same set of ideas have been used to suggest that fewer neurons are required in grid modules with longer lengthscales (Mosheiff et al.).

Much work then analysed the optimal choice of ratio between the lattice lengthscales of successive grid modules. Early experimental work suggested a geometric progression of lengthscales with a constant ratio of between 1.4 and 1.7 (Stensola et al.; Barry et al.), findings that were matched by multiple theoretical accounts (Wei, Prentice, and Balasubramanian; Mathis, Herz, and Stemmler). However, it remains unclear whether a geometric progression model is actually well-matched to data, especially as measuring multiple modules simultaneously is technically difficult. Indeed, recent models based on developmental arguments predict non-geometric ratios that also appear to match measurements well (Khona, Chandra, and Fiete), while our own work which suggests that grid modules should be related by non harmonic ratios (Dorrell et al.).

Grid modules are not only defined by their lengthscale, but also the relative orientation to other modules. To understand these relative orientations, we used the same efficient coding arguments (that show multiple modules of grid cells are optimal) to predict that successive grid modules should be oriented at small angles relative to one another (Dorrell et al.), matching measurements (Stensola et al.; Lykken et al.). Finally, encoding arguments have also proved useful at understanding how grid cells code 1D space (Rebecca et al.), the alignment of grid axes to square rooms (Rebecca et al.), and the changing of grid lattice parameters to different room shapes (Stensola et al.; Dorrell et al.).

In sum, having arrived at a multimodular structure, efficient coding is a useful framework for understanding the details of the multimodular arrangement.

6 Discussion

Over a decade of normative grid cell theorising points to a core claim: grid cells form a (1) high-fidelity, (2) pathintegrating, (3) biologically-plausible code for space. In contrast, normative attempts to explain grid cells without path-integration cannot match their translational symmetry, section 3; and theories using 'overly linear' measures of coding capacity struggle to explain multimodular structure, section 5. This coheres with mechanistic and perturbative work to support a compelling narrative regarding the grid cell code.

There remain puzzles. While models based on action dependent weight matrices recover the multi-modular axis-aligned structure of grid cells in multiple models (Dorrell et al.; Schaeffer et al.), these models are unable to model the conjunctive grid cells. Models using standard RNNs can make statements about precise velocity update mechanisms (Sorscher et al.; Schøyen et al.; Chu et al.), but do so in ways that don't match biology (Schøyen et al.; Chu et al.), are at times badly behaved (Schaeffer, Khona, and Fiete; Schøyen et al.; Pettersen et al.), and struggle to produce multiple modules of grid cells. As such, a normative model that cohesively captures all four grid cell phenomena we began with remains at large. That said, it seems likely that a careful combination of the best parts of existing models might succeed. We now discuss two broader open questions, and a few implications of this body of work.

6.1 Future Work

Grid Cells in Other Spaces

We have focused on grid cells in 2D, a natural question is how might they behave in other spaces. Normative theories of path-integrable representations naturally generalise to other spaces, and almost always predict multiple modules densely packed lattices in other spaces (Stemmler, Mathis, and Herz; Dorrell et al.), matching similar formulations in one dimension (Aceituno, Dall'Osto, and Pisokas). However, it appears that grid cells are a bespoke 2-dimensional system: 1-dimensional maps are understood by mapping onto a slice of the grid lattice (Yoon et al.; Jacob et al.; Rebecca et al.); conversely, 3D grid cells appear to have multiple randomly scattered fields (Ginosar et al.; Grieves et al.), in contrast to either the models discussed so far, and more boutique projection models (Klukas, Lewis, and Fiete). Models have been proposed that cohesively capture some aspects of both 2D and 3D coding (Ginosar et al.), but, as reviewed, appendix A, they do a poor job at fitting 2D behaviour. Whether there is some preserved structure in the 3D recordings, or a more general model that explains how grid cells encode spaces beyond 2D remains a topic for further work.

Warping of Grid Cells to Environments or Rewards

One finding is that grid cells don't always look so… griddy. In trapezoidal environments the lattice bends along the walls (Krupic et al.), the lattice lengthscale gets smaller near boundaries (Hägglund et al.), in large environments there are often inhomogeneities (Stensola et al.;

Gutiérrez-Guzmán, Hernández-Pérez, and Dannenberg) (though these sometimes disappear with experience; Carpenter et al.), grid fields warp in response to rewards (Boccara et al.), and the grid metric stretches in inhomogeneous environments (Wen et al.). Some models have taken this at face value, and attempted to normatively explain the warped grid responses, for example as the optimal code for uncertainty (Kang, Wolpert, and Lengyel). Others have argued that the warping is the effect of an optimally mixed encoding of additional variables beyond space (Whittington et al.; Dorrell et al.). A final approach models these effects as a re-centering of the grid code in response to an external cue, such as a boundary (Ocko et al.). Since these last two approaches understand inhomogoneities through perturbations to an underlying pure grid cell code, they are consistent with existing normative theories. Indeed, the observed rate maps could represent pure grid code after a spatially dependent recentering operation, making perfect grids appear bent in some environments or towards some rewards. However, the same is not true of the first model, and, as yet, no model is able to bridge these two domains clearly.

6.2 Some Implications

How constrained are these ideas?

Across this body of work, the way in which the three ideas: ‘high-fidelity’, ‘path-integrable’, or ‘biological’, have been formalised has varied. This is a good thing, demonstrating robustness to ad hoc modelling choices. However, some recurring motifs stand-out. In all cases, the biological constraints limit the capacity of the system (e.g. by limiting the range of firing rates), and ensure the problem is not rotationally invariant, using a nonnegativity constraint either on neural firing or on weights. Similarly, path-integration always implies some mechanism for forward modelling: predicting the next encoding from your previous encoding and an action. Finally, the implementation of a high-fidelity code has relied on some form of ‘functional nonlinearity’ in the decoding loss.

Single Neurons are Pleasingly Constraining

Broadly, it is potentially unclear how much measuring a small number of single neurons can reliably guide our understanding of the brain (Whittington and Dorrell). Alternative approaches advocate for studying population-level metrics (e.g. Stringer et al.). There are only ~ 10000s grid cells in a rat (using estimates from Clark and Nolan; Gatome et al.; Diehl et al.), yet reviewing this literature we see that it has been incredibly constraining. Fitting just four high-level properties of the system has identified a core set of computational principles across models, and has proved adept at discounting alternative hypotheses. This is a ringing endorsement for the plodding progress of standard neuroscience.

RNNs as neural models

Using task-optimised neural networks as neural models is somewhat controversial; in complex tasks they are often as confusing as the brain (Banino et al.), limiting the insights we can gain from them. Yet the grid cell literature presents a compelling case for their power when coupled with clear experimentation, and thorough analysis. Task-optimised networks permit you to try a variety of hypotheses relatively quickly and flexibly. Their downside is that the signal you measure might have been caused by any number of choices made in architecture, training, or regularisation, and it is often hard to test for all of these. Simplifying the model to the point where theoretical work is possible can provide insight, allowing fine-tuning of the RNN experiments. For grid cells, iterations of this cycle seem to have nearly converged. We are optimists, and hope this will be more broadly true, suggesting a version of ‘analytic connectionism’ that pairs careful theory and network modelling. Yet, we note that in the grid cell world this has already taken a decade of intense arguments: it is not necessarily easy.

The Power of Normative Modelling

Early work demonstrated that multimodular grid cells are a much more informative code for space than place cells (Mathis, Herz, and Stemmler), leading to a view of grid cells as an efficient code for space. We hope this review has disabused you of this notion: grid cells are an efficient, but not the most efficient code for space—rather, they are the most efficient *path-integrating* code for space: random multifield place cells are the most efficient code, fig. 5. This highlights a role for normative modelling: by searching amongst all possible codes we are forced to consider all alternatives, highlighting how, if the only goal was efficiency, the best choice would never be grid cells. This null result cleanly highlights a key missing ingredient: path-integration.

6.3 Conclusion

In conclusion, the manifold structures present in the grid cell system have provided impressive constraints for normative theorising. After much work, the field has settled on a consistent set of normative theories: grid cells are a high-fidelity, path-integrable, biological (i.e. constrained and axis-dependent) code for space, agreeing with mechanistic and experimental work. In the future we hope these insights will generalise to grid cells in more complex settings, other neural systems, and provide broad lessons for successful normative theorising.

Code

A simple jupyter notebook to generate the optimal representations in fig. 5 and fig. 6 can be found at *https://github.com/WilburDoz/If_Grid_Cells_are_the_answer_what_is_the_Question.git*.

Supplementary material

A Hexagonal Lattices via Dense Packing Arguments

Hexagonal lattices are the densest packing of spheres in 2D space, or analogously, the best arrangement of sensors to minimise the average distance between all points in 2D space and the nearest sensor. One family of efficient-coding-only approaches use this idea to produce hexagonally tuned cells.

Mok and Love argue that place cells form a conceptual clustering of inputs: which place cells is active for each input corresponds to its cluster and the quality of the encoding is given by the resolution of the clustering (i.e. the best clustering would give every input its own cluster, the worst would assign all inputs to one cluster). They argue that space can be thought of as a uniform continuum of inputs to be explained, and that, thanks to dense packing, the optimal choice of a finite set of place cells (clusters) is a hexagonal grid. They then argue that grid cells are a measure of proximity between points in space and their nearest cluster—which in this model is a measure of how well fit that point is by the learnt clusters. Since the data is best explained at cluster centres this forms a hexagonal lattice.

Ginosar et al., prompted by their discovery of non-periodic encodings by grid cells of three-dimensional space (see discussion), present a parsimonious model that explains both 3D and 2D representations. They model grid fields as particles that repulse each other at short distances and attract at intermediate, the dynamics then pushes the particles towards lower energy states, and the optimal state is a dense packing. Matching neural observations, running these dynamics in 2D leads to dense packing hexagonal lattices, while in 3D it often leads to jammed sub-optimal solutions without global periodic structure.

A slightly related idea appears in Huber. In this memory model the classic roles of place and grid cells are reversed, place cells encode where a memory happens (a conjunction of a thing and a place) while grid cells encode the thing that is happening. Grid tuning curves are produced by arguing that the grid cell is encoding a variable that is uniform across space. The model then assumes that inputs that are nearby in space will be grouped into the same memory, while those beyond a critical distance will trigger a new memory. These dynamics lead to a hexagonal lattice receptive field, which can be understood via dense packing.

Despite the elegant simplicity of these approaches, simple functional questions remain non-obvious and key phenomena unexplained. Most pertinently for our current argument, no approach naturally incorporates the translational symmetry of a grid module: in Mok & Love or Huber it is not obvious why grid cells would code for a translated version of either the conceptual fit to data or a set of memories, while in Ginosar et al. some mechanism would be required to align these densely packing lattices across neurons. Similarly unclear is why there are modules with a discrete set of lengthscales or conjunctive grid cells. Finally, why we should think of grid cells as a measure of hippocampal fit, as a discretised version of a uniform variable, or as a set of repulsing particles, when more compelling narratives exist is unclear. Nonetheless, in conjunction with other ideas, dense packing does explain the choice of hexagonal lattice in many models (Stemmler, Mathis, and Herz; Dorrell et al.).

B Efficient Coding Metric Loss with Large Lengthscale Produces Place Cells

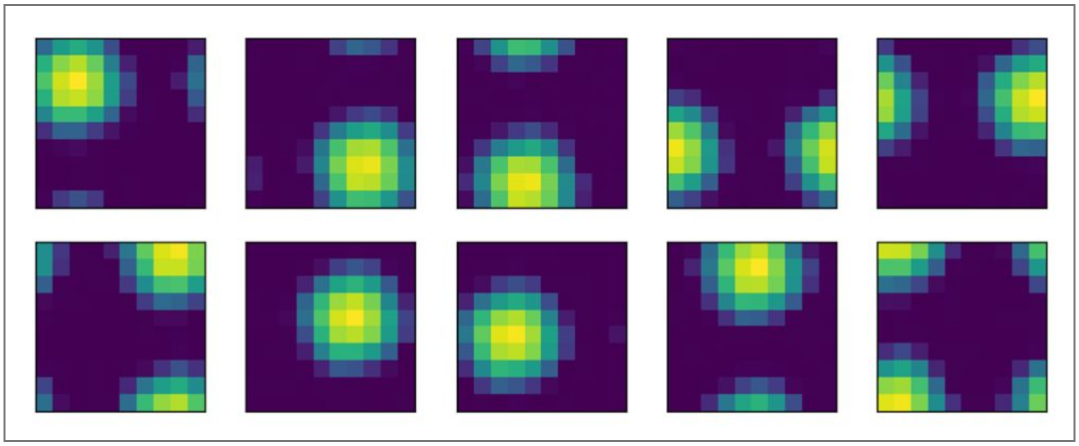

Figure 6. We optimise a metric encoding loss, eq. (3) with large σ and find the optimal representation is place cells, matching the correspondance with the similarity matching objective, section 3.1. We use a periodic environment for convenience, hence the multiple patches observed correspond to parts of the same field.

Acknowledgements

We thank Ben Sorscher, Mikhail Khona, Rylan Schaeffer, Tim Behrens, and Peter Doohan for reading earlier drafts of this work, and especially highlight Charles Burns and Markus Pettersen for their detailed and helpful comments.

We thank the following funding sources: Gatsby Charitable Foundation (GAT3755; W.D.); Sir Henry Wellcome Post-doctoral Fellowship (222817/Z/21/Z; J.C.R.W); European Research Council Starting Grant (NARFB/101222868; J.C.R.W).

Additional information

Author ORCID iDs

William Dorrell: https://orcid.org/0000-0002-6748-3401

References

Aceituno Pau Vilimelis, Dall'Osto Dominic, Pisokas Ioannis (2024) Theoretical principles explain the structure of the insect head direction circuit. *eLife* **13**:e91533 https://doi.org/10.7554/eLife.91533 | PubMed

Atick Joseph J, Redlich A Norman (1990) Towards a theory of early visual processing. *Neural computation* **2**:308-320 https://doi.org/10.1162/neco.1990.2.3.308

Banino Andrea, et al. (2018) Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**:429-433 https://doi.org/10.1038/s41586-018-0102-6 | PubMed

Barry Caswell, et al. (2007) Experience-dependent rescaling of entorhinal grids. *Nature neuroscience* **10**:682-684 https://doi.org/10.1038/nn1905 | PubMed

Boccara Charlotte N, et al. (2019) The entorhinal cognitive map is attracted to goals. *Science* **363**:1443-1447 https://doi.org/10.1126/science.aav4837 | PubMed

Burak Yoram, Fiete Ila R (2009) Accurate path integration in continuous attractor network models of grid cells. *PLoS computational biology* **5**:e1000291 https://doi.org/10.1371/journal.pcbi.1000291 | PubMed

Burgess Neil (2008) Grid cells and theta as oscillatory interference: theory and predictions. *Hippocampus* **18**:1157-1174 https://doi.org/10.1002/hipo.20518 | PubMed

Burgess Neil, Barry Caswell, O'keefe John (2007) An oscillatory interference model of grid cell firing. *Hippocampus* **17**:801-812 https://doi.org/10.1002/hipo.20327 | PubMed

Bush Daniel, Burgess Neil (2014) A hybrid oscillatory interference/continuous attractor network model of grid cell firing. *Journal of Neuroscience* **34**:5065-5079 https://doi.org/10.1523/jneurosci.4017-13.2014 | PubMed

Carpenter Francis, et al. (2015) Grid cells form a global representation of connected environments. *Current Biology* **25**:1176-1182 https://doi.org/10.1016/j.cub.2015.02.037 | PubMed

Chu Tianhao, et al. (2025) Unfolding the Black Box of Recurrent Neural Networks for Path Integration. *bioRxiv* https://doi.org/10.1101/2025.10.25.684492

Clark Harry, Nolan Matthew F (2024) Task-anchored grid cell firing is selectively associated with successful path integration-dependent behaviour. *eLife* **12**:RP89356 https://doi.org/10.7554/eLife.89356 | PubMed

Conklin John, Eliasmith Chris (2005) A controlled attractor network model of path integration in the rat. *Journal of computational neuroscience* **18**:183-203 https://doi.org/10.1007/s10827-005-6558-z | PubMed

Cueva Christopher J, Wei Xue-Xin (2018) Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. arXiv preprint arXiv:1803.07770. https://doi.org/10.48550/arxiv.1803.07770

Deighton Jared, et al. (2024) Higher-Order Spatial Information for Self-Supervised Place Cell Learning. arXiv preprint arXiv:2407.06195.

Diehl Geoffrey W, et al. (2017) Grid and nongrid cells in medial entorhinal cortex represent spatial location and environmental features with complementary coding schemes. *Neuron* **94**:83-92 https://doi.org/10.1016/j.neuron.2017.03.004 | PubMed

Dordek Yedidyah, et al. (2016) Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife* **5**:e10094 https://doi.org/10.7554/eLife.10094 | PubMed

Dorrell Will, et al. (2023) Actionable Neural Representations: Grid Cells from Minimal Constraints. The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=xfqDe72zh41

Dorrell William, et al. (2025) Range, not Independence, Drives Modularity in Biologically Inspired Representations. The Thirteenth International Conference on Learning Representations.

Eliav Tamir, et al. (2021) Multiscale representation of very large environments in the hippocampus of flying bats. *Science* **372**:eabg4020 https://doi.org/10.1126/science.abg4020 | PubMed

Fuhs Mark C, Touretzky David S (2006) A spin glass model of path integration in rat medial entorhinal cortex. *Journal of Neuroscience* **26**:4266-4276 https://doi.org/10.1523/jneurosci.4353-05.2006 | PubMed

Gao Ruiqi, et al. (2019) Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion. Internal Conference on Learning Representations.

Gao Ruiqi, et al. (2021) On Path Integration of grid cells: isotropic metric, conformal embedding and group representation. *Advances in neural information processing systems* **34**

Gatome CW, et al. (2010) Number estimates of neuronal phenotypes in layer II of the medial entorhinal cortex of rat and mouse. *Neuroscience* **170**:156-165 https://doi.org/10.1016/j.neuroscience.2010.06.048 | PubMed

Gil Mariana, et al. (2018) Impaired path integration in mice with disrupted grid cell firing. *Nature neuroscience* **21**:81-91 https://doi.org/10.1038/s41593-017-0039-3 | PubMed

Ginosar Gily, et al. (2021) Locally ordered representation of 3D space in the entorhinal cortex. *Nature* **596**:404-409 https://doi.org/10.1038/s41586-021-03783-x | PubMed

Giocomo Lisa M, Hasselmo Michael E (2008) Computation by oscillations: implications of experimental data for theoretical models of grid cells. *Hippocampus* **18**:1186-1199 https://doi.org/10.1002/hipo.20501 | PubMed

Grieves Roddy M, et al. (2021) Irregular distribution of grid cell firing fields in rats exploring a 3D volumetric space. *Nature neuroscience* **24**:1567-1573 https://doi.org/10.1038/s41593-021-00907-4 | PubMed

Guanella Alexis, Kiper Daniel, Verschure Paul (2007) A model of grid cells based on a twisted torus topology. *International journal of neural systems* **17**:231-240 https://doi.org/10.1142/s0129065707001093 | PubMed

Gutiérrez-Guzmán Blanca E, Hernández-Pérez J Jesús, Dannenberg Holger (2025) Tiling of large-scaled environments by grid cells requires experience. *bioRxiv* https://doi.org/10.1101/2025.02.16.638536 | PubMed

Hafting Torkel, et al. (2008) Hippocampus-independent phase precession in entorhinal grid cells. *Nature* **453**:1248-1252 https://doi.org/10.1038/nature06957 | PubMed

Hafting Torkel, et al. (2005) Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**:801-806 https://doi.org/10.1038/nature03721 | PubMed

Hägglund Martin, et al. (2019) Grid-cell distortion along geometric borders. *Current Biology* **29**:1047-1054 https://doi.org/10.1016/j.cub.2019.01.074 | PubMed

Harland Bruce, et al. (2021) Dorsal CA1 hippocampal place cells form a multi-scale representation of megaspace. *Current Biology* **31**:2178-2190 https://doi.org/10.1016/j.cub.2021.03.003 | PubMed

Hasselmo Michael E (2008) Grid cell mechanisms and function: contributions of entorhinal persistent spiking and phase resetting. *Hippocampus* **18**:1213-1229 https://doi.org/10.1002/hipo.20512 | PubMed

Huber David E (2025) A memory model of rodent spatial navigation in which place cells are memories arranged in a grid and grid cells are non-spatial. *eLife* **13**:RP95733 https://doi.org/10.7554/eLife.95733 | PubMed

Hulse Brad K, Jayaraman Vivek (2020) Mechanisms underlying the neural computation of head direction. *Annual review of neuroscience* **43**:31-54 https://doi.org/10.1146/annurev-neuro-072116-031516 | PubMed

Issa John B, Zhang Kechen (2012) Universal conditions for exact path integration in neural systems. *Proceedings of the National Academy of Sciences* **109**:6716-6720 https://doi.org/10.1073/pnas.1119880109 | PubMed

Jacob Pierre-Yves, et al. (2019) Path integration maintains spatial periodicity of grid cell firing in a 1D circular track. *Nature communications* **10**:840 https://doi.org/10.1038/s41467-019-08795-w | PubMed

Jun Heechul, et al. (2020) Disrupted place cell remapping and impaired grid cells in a knockin model of Alzheimer's disease. *Neuron* **107**:1095-1112 https://doi.org/10.1016/j.neuron.2020.06.023 | PubMed

Kang Louis, Balasubramanian Vijay (2019) A geometric attractor mechanism for self-organization of entorhinal grid modules. *eLife* **8**:e46687 https://doi.org/10.7554/eLife.46687 | PubMed

Kang Yul HR, Wolpert Daniel M, Lengyel Máté (2023) Spatial uncertainty and environmental geometry in navigation. *bioRxiv* https://doi.org/10.1101/2023.01.30.526278 | PubMed

Khona Mikail, Chandra Sarthak, Fiete Ila (2025) Global modules robustly emerge from local interactions and smooth gradients. *Nature* 1-10 https://doi.org/10.1038/s41586-024-08541-3 | PubMed

Kim Sung Soo, et al. (2017) Ring attractor dynamics in the Drosophila central brain. *Science* **356**:849-853 https://doi.org/10.25378/janelia.5648314.v1

Klukas Mirko, Lewis Marcus, Fiete Ila (2020) Efficient and flexible representation of higher-dimensional cognitive variables with grid cells. *PLoS computational biology* **16**:e1007796 https://doi.org/10.1371/journal.pcbi.1007796 | PubMed

Krupic Julija, et al. (2015) Grid cell symmetry is shaped by environmental geometry. *Nature* **518**:232-235 https://doi.org/10.1038/nature14153 | PubMed

Kubie John L, Fenton André A (2012) Linear look-ahead in conjunctive cells: an entorhinal mechanism for vector-based navigation. *Frontiers in neural circuits* **6**:20 https://doi.org/10.3389/fncir.2012.00020 | PubMed

Kunz Lukas, et al. (2015) Reduced grid-cell-like representations in adults at genetic risk for Alzheimer's disease. *Science* **350**:430-433 https://doi.org/10.1126/science.aac8128 | PubMed

Laughlin Simon (1981) A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c* **36**:910-912 PubMed

Logiaco Laureline, Abbott LF, Escola Sean (2021) Thalamic control of cortical dynamics in a model of flexible motor sequencing. *Cell reports* **35** https://doi.org/10.1016/j.celrep.2021.109090 | PubMed

Lykken Christine M, et al. (2025) Functional independence of entorhinal grid cell modules enables remapping in hippocampal place cells. *bioRxiv* https://doi.org/10.1101/2025.09.24.677985

Ma Tzuhsuan (2020) Towards a theory for the emergence of grid and place cell codes. PhD thesis. Massachusetts Institute of Technology.

Mathis Alexander, Herz Andreas VM, Stemmler Martin (2012) Optimal population codes for space: grid cells outperform place cells. *Neural computation* **24**:2280-2317 https://doi.org/10.1162/neco_a_00319 | PubMed

Mathis Alexander, Herz Andreas VM, Stemmler Martin B (2013) Multiscale codes in the nervous system: the problem of noise correlations and the ambiguity of periodic scales. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* **88**:022713 https://doi.org/10.1103/physreve.88.022713 | PubMed

Mathis Alexander, Herz Andreas VM, Stemmler Martin B (2012) Resolution of nested neuronal representations can be exponential in the number of neurons. *Physical review letters* **109**:018103 https://doi.org/10.1103/physrevlett.109.018103 | PubMed

McNaughton Bruce L, et al. (2006) Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience* **7**:663-678 https://doi.org/10.1038/nrn1932 | PubMed

Mok Robert M, Love Bradley C (2019) A non-spatial account of place and grid cells based on clustering models of concept learning. *Nature communications* **10**:5685 https://doi.org/10.1038/s41467-019-13760-8 | PubMed

Mosheiff Noga, et al. (2017) An efficient coding theory for a dynamic trajectory predicts non-uniform allocation of entorhinal grid cells to modules. *PLoS computational biology* **13**:e1005597 https://doi.org/10.1371/journal.pcbi.1005597 | PubMed

Ocko Samuel A, et al. (2018) Emergent elasticity in the neural code for space. *Proceedings of the National Academy of Sciences* **115**:E11798–E11806 https://doi.org/10.1073/pnas.1805959115 | PubMed

Olshausen Bruno A, Field David J (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**:607-609 https://doi.org/10.1038/381607a0 | PubMed

Pastoll Hugh, et al. (2013) Feedback inhibition enables theta-nested gamma oscillations and grid firing fields. *Neuron* **77**:141-154 https://doi.org/10.1016/j.neuron.2012.11.032 | PubMed

Pettersen Markus, et al. (2024) Self-supervised grid cells without path integration. *bioRxiv* https://doi.org/10.1101/2024.05.30.596577

Rebecca RG, et al. (2025) Spatial periodicity in grid cell firing is explained by a neural sequence code of 2-D trajectories. *eLife* **13**:RP96627 https://doi.org/10.7554/eLife.96627 | PubMed

Redish A David, Elga Adam N, Touretzky David S (1996) A coupled attractor model of the rodent head direction system. *Network: computation in neural systems* **7**:671

Reifenstein Eric T, et al. (2012) Grid cells in rat entorhinal cortex encode physical space with independent firing fields and phase precession at the single-trial level. *Proceedings of the National Academy of Sciences* **109**:6301-6306 https://doi.org/10.1073/pnas.1109599109 | PubMed

Rich P Dylan, Liaw Hua-Peng, Lee Albert K (2014) Large environments reveal the statistical structure governing hippocampal representations. *Science* **345**:814-817 https://doi.org/10.1126/science.1255635 | PubMed

Samsonovich Alexei, McNaughton Bruce L (1997) Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience* **17**:5900-5920 https://doi.org/10.1523/jneurosci.17-15-05900.1997 | PubMed

Sargolini Francesca, et al. (2006) Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* **312**:758-762 https://doi.org/10.1126/science.1125572 | PubMed

Schaeffer Rylan, Khona Mikail, Fiete Ila (2022) No Free Lunch from Deep Learning in Neuroscience: A Case Study through Models of the Entorhinal-Hippocampal Circuit. ICML 2022 Workshop AI4Science.

Schaeffer Rylan, et al. (2023) Disentangling Fact from Grid Cell Fiction in Trained Deep Path Integrators. *ArXiv* arXiv-2312 https://doi.org/10.48550/arxiv.2312.03954 | PubMed

Schaeffer Rylan, et al. (2023) Self-supervised learning of representations for space generates multi-modular grid cells. *Advances in Neural Information Processing Systems* **36**:23140-23157

Schaeffer Rylan, et al. (2023) Testing assumptions underlying a unified theory for the origin of grid cells. arXiv preprint arXiv:2311.16295. https://doi.org/10.48550/arxiv.2311.16295

Schøyen Vemund, et al. (2023) Coherently remapping toroidal cells but not grid cells are responsible for path integration in virtual agents. *Iscience* **26** https://doi.org/10.1016/j.isci.2023.108102 | PubMed

Schøyen Vemund Sigmundson, et al. (2025) Hexagons all the way down: Grid cells as a conformal isometric map of space. *PLOS Computational Biology* **21**:e1012804 https://doi.org/10.1371/journal.pcbi.1012804 | PubMed

Sengupta Anirvan, et al. (2018) Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks. *Advances in neural information processing systems* **31**

Skaggs William, et al. (1994) A model of the neural basis of the rat's sense of direction. *Advances in neural information processing systems* **7**

Sorscher Ben, et al. (2023) A unified theory for the computational and mechanistic origins of grid cells. *Neuron* **111**:121-137 https://doi.org/10.1016/j.neuron.2022.10.003 | PubMed

Sorscher Ben, et al. (2019) A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in neural information processing systems* **32**

Sorscher Ben, et al. (2022) When and why grid cells appear or not in trained path integrators. *bioRxiv* https://doi.org/10.1101/2022.11.14.516537

Sreenivasan Sameet, Fiete Ila (2011) Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature neuroscience* **14**:1330-1337 https://doi.org/10.1038/nn.2901 | PubMed

Stachenfeld Kimberly L, Botvinick Matthew M, Gershman Samuel J (2017) The hippocampus as a predictive map. *Nature neuroscience* **20**:1643-1653 https://doi.org/10.1038/nn.4650 | PubMed

Steffenach Hill-Aina, et al. (2005) Spatial memory in the rat requires the dorsolateral band of the entorhinal cortex. *Neuron* **45**:301-313 https://doi.org/10.1016/j.neuron.2004.12.044 | PubMed

Stemmler Martin, Mathis Alexander, Herz Andreas VM (2015) Connecting multiple spatial scales to decode the population activity of grid cells. *Science Advances* **1**:e1500816 https://doi.org/10.1126/science.1500816 | PubMed

Stensola Hanne, et al. (2012) The entorhinal grid map is discretized. *Nature* **492**:72-78 https://doi.org/10.1038/nature11649 | PubMed

Stensola Tor, et al. (2015) Shearing-induced asymmetry in entorhinal grid cells. *Nature* **518**:207-212 https://doi.org/10.1038/nature14151 | PubMed

Stringer Carsen, et al. (2019) High-dimensional geometry of population responses in visual cortex. *Nature* **571**:361-365 https://doi.org/10.1038/s41586-019-1346-5 | PubMed

Tang Mufeng, Barron Helen, Bogacz Rafal (2024) Learning grid cells by predictive coding. arXiv preprint arXiv:2410.01022. https://doi.org/10.48550/arxiv.2410.01022

Tolman Edward C (1948) Cognitive maps in rats and men. *Psychological review* **55** https://doi.org/10.1037/h0061626 | PubMed

Touretzky David S, Redish A David (1996) Theory of rodent navigation based on interacting representations of space. *Hippocampus* **6**:247-270 https://doi.org/10.1002/(sici)1098-1063(1996)6:3<247::aid-hipo4>3.0.co;2-k | PubMed

Towse Benjamin W, et al. (2014) Optimal configurations of spatial scale for grid cell firing under noise and uncertainty. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**:20130290 https://doi.org/10.1098/rstb.2013.0290 | PubMed

Van Cauter Tiffany, et al. (2013) Distinct roles of medial and lateral entorhinal cortex in spatial cognition. *Cerebral Cortex* **23**:451-459 https://doi.org/10.1093/cercor/bhs033 | PubMed

Vollan Abraham Z, et al. (2025) Left-right-alternating theta sweeps in entorhinal-hippocampal maps of space. *Nature* 1-11 https://doi.org/10.1038/s41586-024-08527-1 | PubMed

Waniek Nicolai (2020) Transition scale-spaces: A computational theory for the discretized entorhinal cortex. *Neural computation* **32**:330-394 https://doi.org/10.1162/neco_a_01255 | PubMed

Wei Xue-Xin, Prentice Jason, Balasubramanian Vijay (2015) A principle of economy predicts the functional architecture of grid cells. *eLife* **4**:e08362 https://doi.org/10.7554/eLife.08362 | PubMed

Wen John H, et al. (2024) One-shot entorhinal maps enable flexible navigation in novel environments. *Nature* **635**:943-950 https://doi.org/10.1038/s41586-024-08034-3 | PubMed

Whittington James, et al. (2018) Generalisation of structural knowledge in the hippocampal-entorhinal system. *Advances in neural information processing systems* **31** https://doi.org/10.48550/arxiv.1805.09042

Whittington James CR, Dorrell William (2026) How much neuroscience does a neuroscientist need to know?. arXiv preprint arXiv:2601.02063. https://doi.org/10.48550/arxiv.2601.02063

Whittington James CR, et al. (2023) Disentanglement with biological constraints: A theory of functional cell types. The Eleventh International Conference on Learning Representations.

Whittington James CR, et al. (2020) The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183**:1249-1263 https://doi.org/10.1016/j.cell.2020.10.024 | PubMed

Xu Dehong, et al. (2025) On conformal isometry of grid cells: Learning distance-preserving position embedding. The Thirteenth International Conference on Learning Representations.

Xu Yilun, et al. (2020) A Theory of Usable Information under Computational Constraints. International Conference on Learning Representations. https://doi.org/10.48550/arxiv.2002.10689

Ying Johnson, et al. (2022) Disruption of the grid cell network in a mouse model of early Alzheimer's disease. *Nature Communications* **13**:886 https://doi.org/10.1038/s41467-022-28551-x | PubMed

Yoon KiJung, et al. (2016) Grid cell responses in 1D environments assessed as slices through a 2D lattice. *Neuron* **89**:1086-1099 https://doi.org/10.1016/j.neuron.2016.01.039 | PubMed

Yu Changmin, Behrens Timothy EJ, Burgess Neil (2020) Prediction and Generalisation over Directed Actions by Grid Cells. arXiv preprint arXiv:2006.03355. https://doi.org/10.48550/arxiv.2006.03355

Peer reviews

Reviewer #1 (Public review):

Summary:

The review by Dorrell and Whittington synthesizes the progress made over the past few years with respect to a normative theory of grid cells. The core question addressed by normative frameworks of grid cells is what primary computational function grid cells serve. The review discusses evidence from mechanistic models and experimental data that point to path integration as the computational function of grid cells, consistent with results from normative models. The main goal of the review is to clarify the normative grid cell theory literature. However, the current version of the article reads at times more like a perspective or opinion article in support of the path integration hypothesis rather than a critical review of normative frameworks in the grid cell literature that contrasts the benefits and limitations, as well as pitfalls and caveats, with other modelling approaches.

Some specific comments are as follows:

(1) Abstract: "The first question quickly attracted an answer: grid cells subserve path integration ..." - I am not sure if this statement is correct. The first grid cell paper by Hafting and Fyhn in 2005 suggested that grid cells are part of a path integration-based map, and the paper emphasizes the map part. It remained unclear, and is still debated, whether grid cells are part of a system performing path integration or whether grid maps reflect the output/result of a path integration process. Other theories about the function of grid cells were brought forward as well. Although the main competing theory is discussed in this review, this review article at times appears more as a perspective or opinion article with a clear bias toward the path integration hypothesis rather than objectively discussing the evidence.

(2) Grid cells may serve multiple functions. What would be the implications for our understanding of grid cells and for interpreting the results of normative models? In general, the review could discuss some pitfalls or caveats of normative models in more detail.

(3) A normative framework can be helpful in two ways: (a) Given sufficient details on biological constraints, a normative model can help identify the computational function of grid cells. If a computational function is given and - under the given simulated biological constraints - grid cells were part of the solution, the results of the model would support the hypothesis that grid cells serve the computational function in question. (b) If a computational function were identified beyond any doubt (e.g., assume experimental data demonstrated that grid cells are necessary and sufficient for path integration), a normative model would help identify biological parameters necessary to produce grid cell firing. Unfortunately, the review falls short in making this clear distinction between (a) and (b) and in discussing important caveats regarding mixing up these two ways. E.g., the neural network model approaches by Sorscher et al. and others have been criticized because they try to achieve two things at the same time: find support for the computational function of grid cells and identify optimal parameters that result in grid cells. But doing both at the same time provides a strong bias in tweaking the parameters in exactly the way you need for the model to produce grid

cells as a solution (other solutions may be possible given other parameters), preventing strong conclusions regarding the computational function of grid cells and preventing conclusions about what the parameter choices mean for biological connectivity motifs. These caveats in setting up normative models and interpreting them could be discussed in greater detail.

(4) A common assumption underlying most grid cell models is that head direction is viewed as identical to movement direction. However, head direction can differ at times from movement direction, and entorhinal head direction cells code head direction rather than movement direction (Raudies et al., 2015; 10.1016/j.brainres.2014.10.053). This missing link in how movement direction signals reach and inform grid cells could be discussed.

(5) "Knowing that one neuron in a module is active and that you make a movement north uniquely determines which neuron in that module should be active next" - I agree that this rule follows from the fact that grid cells within one module differ in phase but share spacing and orientation. However, I am surprised that the authors do not also make the argument here for the value of a normative model. Rebecca R.G. et al. (10.7554/eLife.96627) use exactly the rule cited above as a normative function. They demonstrate that this rule begets grid cells. Isn't this a prime example of how a normative approach can contribute to scientific inquiry? First, a hypothesis about a computational function is derived from experimental data. And in turn, using a normative framework, the experimental data are derived from the computational function (under appropriate biological results). The paper is discussed later together with Nicolai Waniek's work (10.1162/neco_a_01255). However, in my opinion, their work seems to be somewhat misrepresented in that later paragraph. E.g., velocity is still required as an input to determine which neuron should be active next, neurons do not need to be binary units, and space is not discretized beyond the fact that space is encoded by neurons with spatial firing fields.

https://doi.org/10.7554/eLife.111058.1.sa3

Reviewer #2 (Public review):

Summary:

This review by Dorrell and Whittington covers a number of aspects related to normative modeling of grid cells. They begin by discussing key experimental insights on grid cell phenomenology. Then, they discuss how grid cells can be used to perform path integration and how they size up as efficient codes of space. These two sections then lead the authors to discuss how combining path integration and efficient coding objectives leads to models of axis-aligned grid cells in a single module. Discussion on non-linear objectives leading to multi-modules is presented. The review ends with several outstanding questions and an optimistic outlook of how normative models (particularly, task-optimized RNNs) can be used as tools for advancing understanding in neuroscience.

Strengths:

(1) The review is timely and covers an area that has seen a lot of recent activity. This discussion around many of the different results (and kinds of models), I think, will be generally helpful for the field.

(2) Although I think the story could be a little more coherently made (see below), in general I enjoyed the author's flow from efficient coding -> efficient coding + path integration -> efficient coding + path integration + non-linear objective. This framing supports the specific conclusion the authors arrive at.

(3) I also really liked the message that the review made of how normative modeling, despite some of its challenges/limitations, can be used effectively in neuroscience. The discussion of cycling between "experimental" modeling (e.g., vanilla RNNs) and theoretically-grounded models was nice, and I think it helps demonstrate the value of this approach.

(4) Showing how the metric loss could be seen as a bandpass filter (Figure 3C) was nice and a contribution of the review.

(5) While the focus of P4 (conjunctive HD-grid cells) felt initially a little cast aside, the discussion around "brain and task-optimised RNNs with standard architectural choices use fundamentally different path-integration mechanism" was nice and I think helpful for steering the community to an interesting open problem.

(6) Identifying how "non-linear functionality" can lead to multi-modules was nice and not something that I have seen as clearly presented before.

Weaknesses:

(1) The authors view the experimental evidence for grid cells being linked to path integration as "specific and strong" and that the " key computational feature that defines entorhinal cortex [is] path-integration". I think experimentalists (at least the ones I work with) would push back on that. First, it's hard to isolate path integration in rodent experiments. So while Gil et al. (2018) did about as good a job as you could do, there are still other interpretations of the results that are not purely path integration dependent. And second, as the authors point out later in the review, there is experimental work finding that grid cells are disrupted in large environments and 3D. Path integration certainly happens (to some extent) in these spaces, which begs the question of how it is achieved with weakened grid coding. Thus, I think reducing the claims about how strongly grid cells are experimentally linked to path integration is called for.

(2) The authors introduce the idea of efficient coding of space and discuss how grid cells are not optimal. It is later clarified (Sec. 5.3) that multi-module codes can be efficient (even if not the most optimal). I was confused reading Section 3, because in Section 2 the multiple modules are discussed, but then in Section 3, they are dropped, and only a single module is being considered. Equation 2 was also a little confusing to me. Alpha is not defined, and I would have thought that it would be x^Tx' - g(x)^T g(x') and not x^Tx' g(x)^T g(x'). Given that there is no page limit here, I think a little more detail in Section 3 would be helpful.

(3) In Section 3, the authors make use of P2 (translation invariance within a module) to rule out (or, at least, question) certain models/approaches. While this is certainly a standard assumption made in theoretical work, it is not very well supported by experimental findings. In particular, Diehl et al. (2017), Ismakov et al. (2017), and Dunn et al. (2017) all found that individual grid fields systematically vary in their peak firing rate. In addition, Redman et al. (2025) found that, within a given module, there was a small but robust diversity of grid orientations and spacings. These suggest that grid cells within a single module may actually be able to encode properties of local space and give some support to normative models that find efficient space coding with grid cells by finding non-axis-aligned grid fields. I think this is all important to mention because: a) it provides more biological nuance to the question about spatial coding; b) it provides more ways in which to test models. For instance, in Redman et al. (2025), the Sorscher et al. (2022) model was shown to produce variability in grid properties that loosely matched what was found in real data. For tests like this (e.g., how much does a model reproduce variability in grid firing field peak rates), I think it is going to be important for continuing to evaluate models.

(4) The focus of the review, I know, is grid cells, but of course, grid cells are part of the MEC and the larger hippocampal network. I totally understand, at some level, you have to make a

decision of what to model, but it seems that there are other functional classes of neurons (border cells, head direction cells) that all play an important role in path integration. And while the models the authors consider at the end of the review capture properties of grid cells really well, they do so at the cost of not modeling anything else. The authors mention this in the context of the models not capturing conjunctive grid-head direction cells, but I think the point is a deeper one, and more discussion of at what level it makes sense to consider grid cells only is important.

(5) As I mentioned in the Strengths section, I did enjoy the flow of the paper on how path integration + efficiency is needed to get grid single modules and path integration + efficiency + non-linearity is needed to get multiple grid modules. This creates the story that adding more of these theory-driven constraints helps lead to more "accurate" models of grid cells. But one alternative view is that, if path integration + efficiency is enough to get a single grid module (but only a single grid module), then maybe the utility (or need) of multiple grid modules comes from something else. That is, instead of saying "we need more constraints to get multiple modules", it could be evidence for "we need to re-think whether multiple modules might need a different theory to explain". While I understand this is a big picture question that maybe isn't entirely fair to ask of the authors, I think: 1) the authors do a nice job of positioning their review as a kind of discussion on what normative modeling can provide to neuroscience, so having this discussion on when the failure of a model to capture ALL aspects of the biological features motivates further constraints as opposed to a new approach, would be useful; 2) this question connects with the title of the paper, i.e. "what is the question?"

https://doi.org/10.7554/eLife.111058.1.sa2

Reviewer #3 (Public review):

Summary:

The authors present an extensive review of the literature on normative grid cell theory, asking what kind of cost function might be minimized by the entorhinal grid cell code. The authors show which of the main features of grid cells emerge from combinations of terms in a cost function that optimizes for spatial fidelity, biological plausibility, and path integration. They conclude by outlining potential future directions for the field.

Strengths:

The structure of the review makes it particularly useful for researchers who are familiar with grid cells but not necessarily with normative models. Equations are kept to a minimum and are usually explained conceptually.

Weaknesses:

I identified one main weakness, related to the fact that the introduction to experimental results around grid cells and what they allow us to conclude is less nuanced than the rest of the review. However, since this is not the main focus of the manuscript, I consider this a secondary limitation.

The review organizes the current literature on the subject within a coherent conceptual framework, helping to define possible paths forward for the field.

https://doi.org/10.7554/eLife.111058.1.sa1

Author response:

We thank the reviewers for their time and attention which will significantly improve the paper. Further, we are grateful for their appreciation of our goals and work. In sum, the reviewers point to our overstated discussion of experimental evidence which we will tone down, some slightly confusing points of argumentation which we will clarify, and some discussion points on the role of normative theories that we will add text to address. We believe this will improve the paper significantly and hope you agree!

Major Concern: Experimental Support for Path-Integration is not as strong as suggested

The major point raised by all reviewers (reviewer 1 comment 1, reviewer 2 comment 1, reviewer 3's only weakness) was that our presentation of the experimental perturbation evidence for path-integration is stronger than the reality. On reflection, we agree with this evaluation. We thank the reviewers for raising it; we will moderate our writing and include the sensible caveats raised. In sum, we still think that the convergence of evidence points to path-integration: first, disruptions to grid cells lead to path-integration problems, though these perturbations admittedly aren't perfectly precise; second, normative theories of path-integration lead to grid cells and predict grid cell behaviour; third, mechanistic models of path-integration match grid cell behaviour and predict connectivity subsequently measured in entorhinal cortex. However, the evidence is not as all-encompassing as we suggested.

That said, we'd like to further comment on one point. It is argued (reviewer 1, comment 1) that there are other theories of grid cell function, and that we discuss these theories. We discuss efficient-coding only models of grid cells and emphasise strongly why we reject them. We also briefly discuss oscillatory-interference models of path-integration and our reasons for not pursuing them further. As such, the reviewer is correct that our reading of literature strongly points us towards path-integration rather than other theories. We will slightly change the framing of the paper to make it clear that we are making a case. However, we are not aware of other theories the reviewer might be referring to. If the reviewer can point us to the other suggested theories that we do not address we would be happy to evaluate and include them.

We now turn to the remaining comments, and how we plan to address them.

Reviewer 1, Comment 2 – There could be multiple roles for grid cells

The reviewer is indeed right that grid cells might perform multiple functions. This could just mean that the same computational motif (e.g. path-integration) is reused across different computations though that introduces no changes to the required normative theory. A stronger claim would be that grid cells perform both path-integration and some other function. This, according to a normative perspective, would most likely change how grid cells were optimally structured. We use the fact that large parts of the grid cell code can be captured with only path-integration as an argument against additional roles for grid cells. That said, there exist properties of grid cells not well-captured by path-integration which could well be smoking guns for additional roles of grid cells. The review already discusses both discrepancies between grid cells in three and two dimensions, and inhomogeneities in the grid in complex environments, and we will add two more (heading direction and peak-to-peak/angular variability, discussed below) that we are grateful to the reviewers for raising, and we discuss each of these in detail below.

That said, whether these are necessarily arguments against purely path-integration or a reflection of interesting mappings of the core path-integration mechanism to the measurements we make remains to be seen. We would argue that both 3D grid cells (as explained below: there appear to be 2D slices in which grid cells behave as you'd expect) and spatial inhomogeneities (as explained in the paper: mappings of torus to world can introduce warping) can be explained without reference to additional computational roles of grid cells, which remain to us the most parsimonious explanation. We discuss next the slight update to path-integration only that the heading direction story suggest. But in sum, our view is that these discrepancies are likely not fatal for our path-integration-centric view of grid cells, but may well suggest some very interesting clarifications.

Reviewer 1, Comment 4 – The system has two heading signals: true & internal, why?

The reviewer is right to point to the puzzle over true vs. purely internal heading direction and which drives grid cells. We believe recent work from Abraham Vollan has effectively solved this puzzle: there appear to be two parallel circuits, one theta-modulated and following internal heading direction, another theta-unmodulated and aligning more with true heading direction. We will make sure to include discussion of this exciting work in our revised submission. This serves as a good example of an update we concede to the most austere version of the path-integration only view. Rather, it seems there are two parallel path-integrators working with different heading signals. The reasons for this remain unclear, but seem to be related to attention and planning (Vollan et al. 2026).

Reviewer 2, Comment 3: Real Grid Cells have peak-to-peak variability & Angular variability

The reviewer is right to point to the discrepancy in peak-to-peak firing rate and angles within a module that we did not adequately address. First, it is Sorscher's RNN models, not nonnegative PCA that can generate a distribution of grid angles (Redman et al. 2025), which suggests that path-integration and such variability are compatible. We emphasise this point because the non-path-integration results from nonnegative PCA produce grid cells oriented at 30 degree offsets, something not measured even when you're careful as in Redman et al. 2025. Thus, this becomes an interesting target for future work: perhaps using theories of path-integration up to an error threshold (rather than perfect) such angular diversity would be recovered. We will include this in our discussion. Further, we will include discussion of peak-to-peak variability that, as yet, has no obvious role.

Reviewer 2, Comment 1: grid cells are inhomogeneous in 3D or complex environments, doesn't that break the theory?

Disrupted grid coding in extended or 3D environments indeed deserve more discussion, which we will add. In particular, we will add recent evidence that grid cells in 3D can be understood via the correct sequence of 2D projections(Qi & Yartsev, 2026). These two phenomena seem, to us, consistent with a path-integration only view of grid cells, as discussed above, and we hope to make this position clearer.

Reviewer 2, Comment 5: Couldn't there be other reasons for multiple modules?

We have suggested a consistent normative framework in which multiple modules are explained through their role in non-linear coding. We think this elegant, and the most parsimonious current theory. We could, of course, be wrong. The discrepancies pointed to above might be good clues to follow to work out what else these modules might be doing, but currently these alternative explanations seem not to exist. We will text to clarify this.

Reviewer 1, Comment 3: The review confuses computational and parameter parts of normative theory

We disagree with the reviewer's dichotomisation of normative theory. We view a normative theory as the complete procedure that produces the predictions. Almost all such theories have parameters and hence fitting a theory to data comprises both elements (a) [computational role] and (b) [specific parameters] identified by the reviewer. Occasionally theories have no parameters in the traditional sense, e.g. Rebecca et al.; instead they have heavy assumptions that play an equivalent role. It is true that, as the reviewer says, Sorscher et al.'s work was criticised for producing grid cells only for specific parameter values. We never found this as damning as Schaeffer et al. argued: simply it says that that theory is only correct within the given parameter range. Rather, arbitrating between models, parameters, or assumptions seems the same basic process: see what they predict and keep working with models while they remain useful ways to understand measured phenomena. If a model with very specific parameter values remains useful, that seems okay. In fact, we argued extensively why we think the nonnegative PCA model is not a useful model, but this was for completely different reasons. To us this story just reinforces the importance of hygiene in normative research: perform parameter sweeps and clarify how they constrain the claims you are making, carefully arbitrate what models can capture. Indeed, that is the whole goal of this review. We might be misunderstanding and, if so, we welcome correction.

Reviewer 2, Comment 4: Normative Models of Cells Beyond Grid Cells

The reviewer is right that extending these models to other cell types is an interesting area for further work, and that other cell types do seem to be involved in aspects of navigational computations both in RNNs and the brain. We will include a discussion to this effect in the revised manuscript. That said, we think the modularity of grid cells and their tight-linking to path-integration calculations should also be appreciated as a win!

Reviewer 2, Comment 2: Multi-modularity is not cleanly explained

We thank the reviewer for the comments, we agree. We will clarify the story regarding multiple modules, and will explain the equation further.

Reviewer 1, Comment 5: the early introduction of phase-shifted Grid Cells seem the perfect place to normatively argue for Path-integration!

We agree with the reviewer that this point can be made both normatively ('oh look! If I try to do this optimally, I get translations!') or, as we did early in the paper, mechanistically ('oh look! With these cells I can do this!'). Indeed, a large part of the point of our paper is that path-integration is what is required to normatively derive phase-shifted grid modules, something discussed by Rebecca et al., our earlier work, and RNN studies, and appreciated for two decades. The earlier part of the paper does not discuss these papers as that section is aimed at giving intuition for the solution (mechanism). Later sections then heavily discuss the normative angle. We hope that division of labour makes sense.

Finally, we will refine our summary of Rebecca et al. The reviewer is right that neurons don't have to be discrete, we apologise for that error, but our understanding is that the only meaningful role of a neuron in Rebecca et al.'s work is the region in which is active, effectively making every neuron a binary unit, which seems dubious. We will clarify that by "predict velocity from each current and next encoding" we mean that the normative constraint they enforce is axiom 1: sequential activity of sets of neurons i then j can be uniquely interpreted as a trajectory, i.e. a step or velocity. Their work is elegant, and we will try to do more justice to it in the revision.

To conclude, we thank the reviewers for their extensive comments, and look forward to releasing a version that addresses their concerns.

https://doi.org/10.7554/eLife.111058.1.sa0