No Free Lunch from Deep Learning in Neuroscience: A Case Study through Models of the Entorhinal-Hippocampal Circuit

Anonymous Author(s)
Affiliation
Address
email

Abstract

Research in Neuroscience, as in many scientific disciplines, is undergoing a renaissance based on deep learning. Unique to Neuroscience, deep learning models can be used not only as a tool but interpreted as models of the brain. The central claims of recent deep learning-based models of brain circuits are that they make novel predictions about neural phenomena or shed light on the fundamental functions being optimized. We show, through the case-study of grid cells in the entorhinal-hippocampal circuit, that one may get neither. We begin by reviewing the principles of grid cell mechanism and function obtained from first-principles modeling efforts, then rigorously examine the claims of deep learning models of grid cells. Using large-scale hyperparameter sweeps and theory-driven experimentation, we demonstrate that the results of such models may be more strongly driven by particular, non-fundamental, and post-hoc implementation choices than fundamental truths about neural circuits or the loss function(s) they might optimize. We discuss why these models cannot be expected to produce accurate models of the brain without the addition of substantial amounts of inductive bias, an informal No Free Lunch result for Neuroscience. Based on first principles work, we provide hypotheses for what additional loss functions will produce grid cells more robustly. In conclusion, caution and consideration, together with biological knowledge, are warranted in building and interpreting deep learning models in Neuroscience.

1 Introduction

Over the past decade, deep learning (DL) has underpinned nearly every success story in machine learning, e.g., [44, 4] and increasingly many advances in fundamental science research, e.g., [26]. In neuroscience, deep learning is similarly gaining widespread adoption as a useful method for behavioral and neural data analysis [41, 39, 21, 33, 30, 36].

But DL offers a unique contribution to neuroscience that goes beyond its role in other fields, in that deep networks can be viewed as models of the brain. The success of DL in matching or surpassing human performance means it is now possible to construct models of circuits that may underlie human intelligence. As a recent review wrote, “researchers are excited by the possibility that deep neural networks may offer theories of perception, cognition and action for biological brains. This approach has the potential to radically reshape our approach to understanding neural systems" [42]. Further, DL is a democratizing force for building neural circuit models of complex function.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

The essential claims (and promises) of DL-based models of the brain are that 1) Because the models are trained on a specific optimization problem, if the resulting representations match what has been observed in the brain, then they reveal which optimization problem(s) the brain is solving, or 2) These models, when trained on sensible optimization problems, should generate novel predictions about the brain's representations and behavior.

However, given the nascent nature of such approaches and the excitement accompanying some claims, we should examine them carefully. In DL and deep reinforcement learning, some successes attributed to novel algorithms have been shown to instead stem from seemingly minor or unstated implementation choices [52, 15, 25]. In this paper, we ask whether Neuroscientists should similarly be cautious that DL-based models of neural circuits that make specific claims about revealing the brain's optimizaton functions or that predict specific neural tuning curves may tell us less about fundamental scientific truths and more about programmers' particular implementation choices, and might as a result be more post hoc than predictive.

To explore these questions and point toward a systematic and circumspect use of DL for Neuroscience, we evaluate recent DL-based models of grid cells in the entorhinal-hippocampal circuit. The medial entorhinal cortex (MEC) and hippocampus (HPC) are part of the hippocampal formation, a critical brain structure for learning and memory. In a pair of Nobel-prize winning discoveries, HPC was shown to contain **place cells** [38] and MEC, its cortical input, was shown to contain **grid cells** [22]. Place cells fire at one or several seemingly random locations in small and large environments [40], respectively, while grid cells fire in a spatially periodic pattern in all two-dimensional environments, whenever the animal is at the vertices of a hexagonal lattice [22]. Over five decades, the hippocampal formation has been central to understanding how the brain organizes spatial and episodic memory, for experimentalists and theorists alike, with many mysteries remaining. A recent series of DL-based models of the circuit [12, 2, 46, 54, 37]) present a story that training neural circuits on **path integration (PI)** (i.e., the task of estimating one's spatial position in an environment by integrating velocity estimates) results in the emergence of grid cells.

We use public code from prior publications to demonstrate these results are due to implementation details that tell us more about those choices than they do about MEC. By leveraging theoretically-guided large-scale hyperparameter exploration and hypothesis-driven experimentation, we show:

1. Networks trained on path integration tasks almost always learn to optimally encode position, but almost never learn grid-like representations to do so.
2. The emergence of grid-like representations depends wholly on a specifically chosen encoding of the supervised target, not on the task itself, and these choices may be unrealistic.
3. The chosen encoding requires many other sensitive hyperparameter choices, such that small alterations result in loss of grid-like representations.
4. Grid periods and period ratios depend on hyperparameter choices and are not set by the task.
5. Multiple grid modules, a fundamental characteristic of the grid cell system, do not emerge.

DL produces grid cells and some attendant properties only after making many specific design choices and searching hyperparameter space to obtain such representations, effectively baking grid cells into the task-trained networks. **It is highly improbable that DL models of path integration would have produced grid cells as a novel prediction from task-training, had grid cells not already been known to exist.** Moreover, it is unclear what interpretability or understanding these models contribute, beyond or even up to what has already been shown for these particular circuits. These results challenge the notion that deep networks offer a free lunch for Neuroscience in terms of discovering the brain's optimization problems or generating novel a priori predictions about single-neuron representations, and warn that caution is needed when building and interpreting such models.

Our work benefited greatly from previous publications that published their code, for which the authors should be commended. By building on their code, we have been able to present novel insights that

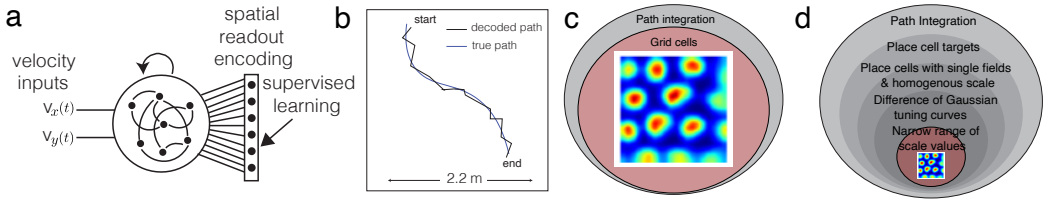

Figure 1: (a-b) Schematic of approach of training recurrent networks to predict (different possible encodings of) 2D position from 2D velocity, in a supervised manner. (b-c) Grid cells are shown to emerge in recent DL papers as a consequence of optimizing a path integration objective [12, 2, 46, 54, 37], with the suggestion that path integration implies grid cells. (d) In the current work, we show how most ANNs trained to path integrate can do so, but only a very small set of output encoding functions (vanishingly small in the function space) and very small fraction of hyperparameter space yields grid cells, leading to the conclusion that grid cell emergence results in ANNs are post hoc: they result from tuning hyperparameters to bake grid cells into networks.

we hope will contribute to a clearer understanding of the potential risks and rewards of using and interpreting DL models in Neuroscience. To facilitate further research, we similarly release code[1].

2 Background: grid cells

Grid cells [22] are found in the medial entorhinal cortex of mammals and are tuned to represent the spatial location of the animal as it traverses 2D space. Each cell fires at every vertex of a triangular lattice that tiles the explored space, regardless of the speed and direction of movement through the space. As a population, grid cells exhibit several striking properties that provide support for a specialized and modular circuit. Grid cells form discrete modules (clusters), such that all cells within a module share a common period and orientation, while different modules express discretly different spatial periods [49]. The period ratios of successive modules have values in the range of 1.2-1.5.

The mechanism underlying grid cells is widely supported to be through attractor dynamics: Translation-invariant lateral connectivity within the grid cell network results in a linear Turing instability and pattern formation [7, 17, 5]. These models explain how grid cells can convert velocity inputs into updated spatial estimates, and make several predictions that have been confirmed in experiments, including most centrally the stability of low-dimensional cell-cell relationships regardless of environment and behavioral state, that define a toroidal attractor dynamics [18, 55, 51, 20, 19].

3 Experimental approach

Path integration (PI) is the task of using self-velocity estimates to track one's spatial position, a crucial component of spatial navigation. The central message of existing DL models of grid cells is that training ANNs to path integrate causes the networks to learn grid cells [12, 2, 46, 54, 37].

We follow the setup used by many previous papers: a 2.2m x 2.2m arena is created, then, spatial trajectories (i.e. sequences of positions and velocities) are sampled. Networks receive as inputs the initial position and velocities, and are trained to output (a possible encoding of) the positions (Fig. 1ab) in a supervised manner. There are multiple possible encodings of position, and as we will show, this choice is critical. Two simple encodings are Cartesian [12] or polar [1]. Another encoding scheme is via bump functions in 2D space, with each output assigned a single different position that tiles the space and all outputs assigned identical tuning curves [2, 46, 46, 37]. This encoding has been equated with place cells, even though place fields tend to be heterogeneous in size and shape [40, 13], as well as in number [40], and unlike the choice of a difference-of-Gaussians (DoG) readout tuning in ANN models, do not exhibit any inhibitory surround. See Appendix A for position encoding details. For all encodings, supervised learning is used to train the network via backpropagation through time.

[1] GitHub link (currently redacted to preserve anonymity for the NeurIPS review process)

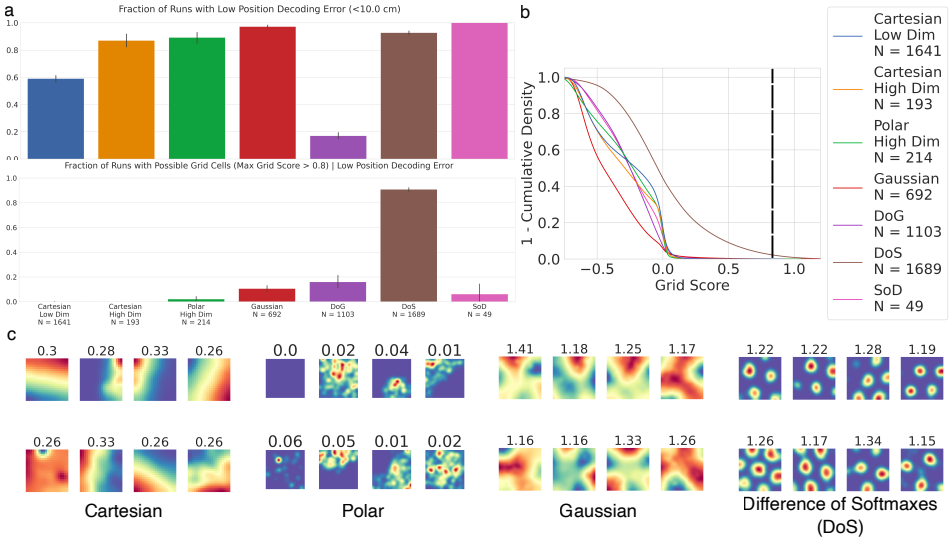

Figure 2: (a) Top: Across readout encodings, almost all networks learn to optimally encode position. Bottom: Few networks display possible grid-like representations (grid score threshold = 0.8). (b) Kernel density estimates of grid scores per readout encoding. (c) Rate maps of high grid-scoring units in deep networks trained on i) Cartesian, ii) Polar, iii) Gaussian, iv) specifically selected (tuned) Difference-of-Softmaxes (DoS) readouts. i)-iii) do not learn any grid cells. (b) Only networks trained on DoS readouts display grid-like cells. Numbers above rate maps are grid scores.

Spatial tuning assessments The spatial tuning of hidden units in the networks, visualized as ratemaps, is the primary method for comparison with the brain's grid cells. To compute ratemaps, a trained network is evaluated on long trajectories that cover the 2D environment, then each hidden unit's activity is binned against the true position in the environment, and counts per bin are normalized by the number of visits to that bin. Ratemaps of units are compared with grid cells through the gridness score. **We are extremely lenient with classifying a training run with particular hyperparameters a success: if even a *single* hidden unit has a grid score above a certain threshold, we say the model possesses grid cells.** The grid score is not perfect since cells classified by grid scores represent only an upper bound on the total number of grid cells; for details, see Appendices B and C.

4 Networks trained on path integration tasks learn to estimate position, but rarely learn grid cells

As [46] wrote in their section titled "Optimally encoding position yields diverse grid-like patterns," "Why do these diverse architectures, across diverse tasks (both navigation and autoencoding), all converge to a grid-like solution, and what governs the lattice structure of this solution?" We demonstrate, in contrast, that most networks do not converge to a grid-like solution, instead requiring very specific readout tuning functions and hyperparameter choices. Grid-like representation emergence is determined by choices made by the programmer that, rather than relating to the path integration objective or biologically realistic place cells, is designed post hoc to produce grid cells.

We ran large-scale hyperparameter sweeps across common implementation choices: 1) Architectures: RNN [14]; LSTM [24]; GRU [10]; UGRNN [11]; 2) Activation: Sigmoid; Tanh; ReLU; Linear; 3) Optimizers: SGD, Adam [31]; RMSProp [23] 4) Supervised Targets: Cartesian; Polar; high-dimensional bump-like readout population code with Gaussian [2], **Difference-of-Softmaxes (DoS)** [46, 47, 37] or Difference-of-Gaussians (DoG) tuning curves. 5) Loss: mean squared error on the agent's Cartesian position [27, 12]; geodesic distance on the agent's polar position [1]; cross entropy on a high-dimensional population of bump-like readout units [2, 46, 37] 6) Miscellaneous: recurrent & readout dropout, initialization, regularization, seed.

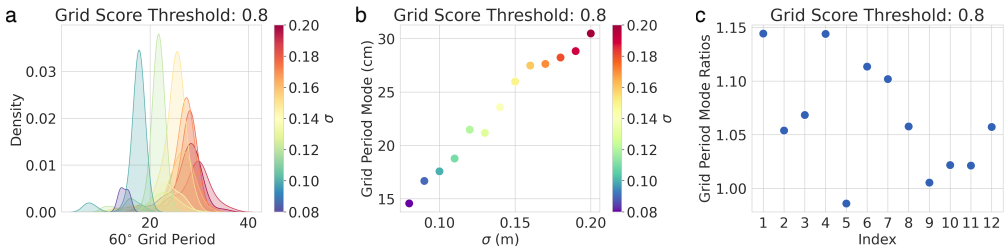

Figure 3: **Grid periods are hyperparameter-dependent and multiple modules do not emerge.** (a) Over a wide sweep of DoS (the most likely to produce grid cells, Fig. 2a bottom) target field widths σ_E, the distribution of grid periods is unimodal (each color is distribution of periods obtained from 3 runs with a σ_E value; different periods are only obtained by varying σ_E), meaning multiple grid modules do not emerge, in contrast to real grid cells. (b) The chosen target field width σ_E determines the grid period mode, meaning that hand-designed hyperparameter choices, not an intrinsic emergent property, sets the grid period. (c) If we use grid periods obtained from smoothly sweeping σ_E as a proxy for different modules, the period ratios are closer to 1 than the experimental ratios of ~ 1.4.

For networks with bump-like population readouts, we additionally swept: 1) Field width σ i.e. standard deviation of the Gaussian tuning curve (often denoted $\sigma_E \stackrel{\text{def}}{=} \sigma$ in the literature); 2) Whether the bump-like readouts have homogeneous or heterogeneous field widths; 3) In the case of readouts with DoG or DoS tuning curves, the surround scale s i.e. the ratio between the inhibitory and excitatory Gaussian's standard deviations ($s \stackrel{\text{def}}{=} \sigma_I/\sigma_E$); 4) Number of fields per readout unit. Evaluating the entire hyperparameter volume is computationally prohibitive, so we evaluated a subvolume most consistent with previous papers, focusing on perturbations from hyperparameters that did produce grid cells. In this sense, **our search was biased toward configurations shown to produce grid cell emergence and thus our findings about the fragility of these solutions conservatively favored these solutions as much as possible.** All sweeps are provided in Appendix E.

To evaluate whether a network learns to optimally encode its position, we measured its position decoding error using previous papers' method [2, 46, 37]: we computed position decoding error using the network's output Cartesian positions (if trained on Cartesian targets) or by decoding position from the network's outputs. Any network with error < 10 cm was considered to have achieved optimal position encoding; this threshold was chosen based on noise inherent in the decoding algorithm.

In total, we trained > 5500 networks and found that most hyperparameter configurations succeed in learning to path integrate (Fig. 2a, Top), but few learn grid cell representations (Fig. 2a, Bottom). This is consistent with earlier work [27, 1] demonstrating that networks can learn to path integrate and solve other navigational problems (e.g. estimating which of several environments correspond to the current location) without lattice cells emerging as a solution.

5 Lattice cell emergence requires a highly specific choice of supervised target encoding

We next sought to characterize when grid cells are learnt under different encodings of the supervised target i.e., 2D position. We tested multiple encodings: i) Cartesian, ii) Polar, iii) Gaussian, iv) a very specifically selected, Difference-of-Gaussian (DoG/DoS) readout and more. We found that grid cells do not emerge from Cartesian, Polar or Gaussian encodings (Fig. 2abc), consistent with earlier work [27]. Only by choosing a DoG readout tuning curve did grid cells sometimes emerge.

Grid periods are parameter-dependent and multiple modules do not emerge

Next, a prominent feature of grid cells that is critical for their unambiguous encoding of position over large scales is the existence of a discrete set of grid periodicities, which tend to scale by a rough factor of 1.4 between adjacent scales [49]. We asked whether ANN models generate multiple periods and when they did so, what hyperparameters and other choices the formed periods depended on.

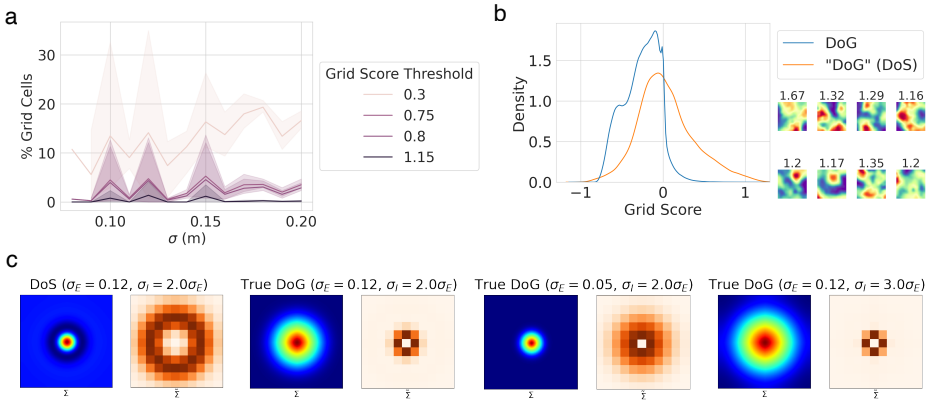

Figure 4: **Other changes to Difference-of-Softmaxes (DoS) affect the formation of grid cells:** (a) The existence of grid solutions even with DoG readout encodings is highly sensitive to parameters such as the target function receptive field width σ_E, and small alterations result in the disappearance of grid cells regardless of grid score threshold. (b) (left) Comparison of grid scores of trained networks with DoG shows that the Difference-of-Softmaxes (DoS), as used by [46, 37, 47], is critical for high grid scores, (right) Rate maps of highest scoring units from DoG networks achieving low position decoding error. (c) Computing the Fourier transform (right of each pair) of the readout second-moment matrix (left of each pair) explains that DoS places Fourier power on an annulus with sufficiently large radius to produce grid cells, showing why the particular DoS choice is critical.

To ensure we would obtain at least some grid cells, we fixed the readouts to be DoSs, and swept over different scales of the DoS. We found that almost all runs had a unimodal distribution of grid periods (Fig. 3a), meaning the networks learnt only one module of grid cells.

Further, we found that the period of the formed grid-like representation is completely determined by the scale σ_E of the externally imposed readout DoG (Fig. 3). The period of the grid-like responses in every run increased monotonically with the width of the DoG readout (Fig. 3b). Since the models did not result in multiple modules in a single network, we used the somewhat discrete distribution of peaks of the single module formed when sweeping the DoG parameter more continuously to compute grid period ratios. These period ratios from adjacent peaks led to non-biological values (Fig. 3c).

Fourier analysis of Turing instability explains the preceding empirical results Why do only Difference-of-Softmaxes (DoS) readouts produce grid cells? We restate the essence of previous analyses of first-principles models [7, 29] here, and explain why they might explain our empirical findings.

In a recurrent network with dynamics $\dot{r}(x) = -r(x) + g(W \star r)$, where x designates the neural index (in a continuum approximation for neurons), $W \star r$ designates the total (integrated) inputs from the network to the neuron at index x, and g is the non-linearity, if the recurrent weight interaction is translationally invariant, then $W(x, x') = W(x - x') = W(\Delta x)$. Under DoG interactions:

$$W(\Delta x) \equiv f(\Delta x) = \alpha_E \exp\left(-\frac{(\Delta x)^2}{2\sigma_E^2}\right) - \alpha_I \exp\left(-\frac{(\Delta x)^2}{2\sigma_I}\right) \tag{1}$$

where Δx refers to the difference of indices between the neural pair linked by the weights. The evolution of activity can be decomposed into the growth and decay of Fourier components of the rate vector, which is fully determined by the Fourier transform of W, which is given by:

$$\tilde{f}(k) = \int_{\mathcal{R}} d(\Delta x) f(\Delta x) e^{ik\Delta x} = \alpha_E \sigma_E \exp\left(-\frac{\sigma_E^2 k^2}{2}\right) - \alpha_I \sigma_I \exp\left(-\frac{\sigma_I^2 k^2}{2}\right) \tag{2}$$

Here α_E (α_I) denotes the strength and σ_E (σ_I) denotes the scale of excitation (inhibition). For linearized dynamics that approximate $\dot{r}(x) \sim -r(x) + f(\Delta x) \star r$ (i.e. g has been linearized),

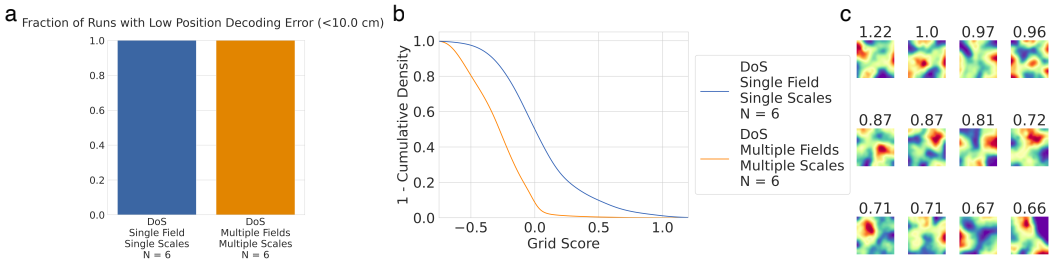

Figure 5: **Adding place cell-like heterogeneity to the readouts prevents the formation of grid cells.** To maximally favor the emergence of grid representations, we selected DoS encodings with the best hyperparameters for grid cell emergence (RNN or UGRNN, ReLU, $\sigma_E = 0.12$ cm, $s = 2.0$, 3 seeds), then tested the effect of multiple fields per place cell ($\sim 1 + \text{Pois}(3.0)$) and multiple scales (receptive field width $\sigma_E \sim \text{Unif}(0.06, 1.0)$ m and surround scale $s \sim \text{Unif}(1.25, 4.5)$). (a) Networks trained on single-field single-scale DoS and multi-scale multi-field DoS (more similar to biological place cells) readouts all obtain low position decoding error. (b) Multi-scale multi-field DoS readouts do not learn grid cells. (c) Highest-scoring rate maps from multi-field multi-scale networks.

the solution will be periodic if the maxima of $\tilde{f}(k)$, given by $[k^*]^2 = \frac{2}{\sigma_E^2 - \sigma_I^2} \log\left(\alpha_E \sigma_E^3 / \alpha_I \sigma_I^3\right)$, contains sufficient power and if $k^* \neq 0$. Specifically, the condition for pattern formation is $\tilde{f}(k^*) > 1$ [8, 29]. In particular, the inhibitory surround contained in $f(\Delta x)$, with strength σ_I, is key to pattern formation; if $\sigma_I \to \infty$ or $\alpha_I \to 0$, the maximum is at the origin ($k^* = 0$), causing no pattern formation. Therefore, a Gaussian interaction cannot produce periodic patterns.

The connection of this theory to ANNs trained with supervised readout target P was made in [46] through the observation that dynamics with an MSE loss $||P - W_{\text{readout}} r||^2$ can be approximated as $\dot{r} = -\lambda r + \Sigma r$, where $\Sigma = PP^T$ is the readout correlation matrix and λ is a regularization parameter. This dynamics now looks identical to the first dynamics with recurrent interaction matrix $W(x, x') = \Sigma_{xx'}$. For identical, unimodal target readout functions with DoG tuning curves, this readout correlation matrix is also a difference of Gaussians (Appendix D). By the same theory outlined above, Gaussian tuning curves should not produce periodic patterns.

Unmentioned implementation details matter We discovered a seemingly minor implementation detail in several preceding papers [46, 47, 37] that proved important for the emergence of grid cells, that (to the best of our knowledge) is unmentioned in the main texts and supplements. While these papers nominally refer to a Difference-of-Gaussian (DoG) readout target function (Eqn. 1), their code instead uses a Difference-of-Softmaxes (DoS) target function. When we trained ideal grid-forming ReLU networks on place-cell-like target encodings with DoG tuning curves, sweeping the receptive field σ and surround scale s, DoG readouts did not result in grid cells (Fig. 4b), while DoS readouts did. The preceding Fourier analysis shows why DoS tuning, though still more unrealistic as a place cell model than DoG, matters (Fig. 4c).

6 Grid cells disappear with more biologically-realistic readout functions

Place cells, to which grid cells project, differ significantly from the highly simplified single-scale, single-field ANN readouts. Place cells have heterogeneously distributed field widths, many with multiple fields [40, 13] (and are not DoS-like functions). This naturally leads to the question: Will readouts targets with multiple scales and multiple fields per place cell still produce grid cells?

We found that networks trained with multiple-field multiple-scale bump encodings per readout unit achieve comparably low position decoding error as good single-field single-scale encodings (Fig. 5a), but do not learn grid cells (Fig. 5bc). This further demonstrates that accurate position encoding and integration does not require grid representations.

7 Why path-integrating ANNs might achieve high predictivity of MEC data

We conclude by introducing a puzzle. A recent NeurIPS spotlight [37] notes that networks trained on single-field single-scale DoS readout encodings explain variance in mouse MEC neural activity at nearly 100% of variance explained by other mice. In contrast, our results demonstrate that these networks learn few grid cells and produce unimodal grid period distributions inconsistent with biological grid cells, and require readout encodings inconsistent with biological place cells to do so. How are these networks able to predict mouse MEC neural activity so well?

The data preprocessing and analysis code is not public, so we are unable to investigate this puzzle in detail. However, we offer a conjecture with preliminary supporting evidence. The analysis of [37] linearly regressed rate maps from one agent (mouse or network) onto rate maps from another mouse, and used Pearson correlation as a measure of "neural predictivity." We conjecture that different architecture-activation pairs achieve different neural predictivity scores because different pairs learn different intrinsic dimensionalities that then provide richer or poorer bases for linear regressions.

To explore our conjecture, we trained [37]'s 5 architectures (RNN, LSTM, GRU, UGRNN, SRNN) with ReLU activations, with 5 seeds per architecture-activation pair. For each trained network, we extracted rate maps and computed a standard linear measure of dimensionality called participation ratio [32], then plotted each pair's average participation ratio against the published neural predictivity. We chose a linear measure of intrinsic dimensionality since linear regression cannot fit nonlinear patterns. We found that networks with higher (lower) dimensional rate maps have higher (lower) neural predictivity (Fig. 6).

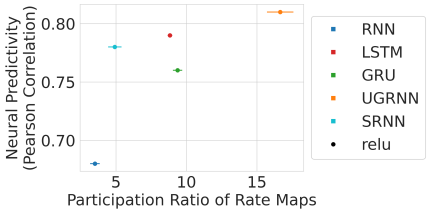

Figure 6: **Networks with higher (lower) dimensional rate maps display higher (lower) "neural predictivity" of mouse MEC rate maps.** Each dot is an architecture-activation pair; participation ratio averaged over 5 seeds. Neural predictivity from [37].

We caution that this correlation between rate map dimensionality and neural predictivity is not strong evidence, but we are unable to investigate further without data preprocessing and analysis code. If our conjecture is correct, it raises an important future research question of whether linear regression-based comparisons of ANNs with neural data might produce better matches to biology more because ANNs have higher dimensional representations than competitors, rather than because of a detailed similarity.

8 Discussion

For research that uses deep networks as models of the brain, there is a fundamental obstacle to making the claim that a given optimization problem is what the brain is solving: If we know the responses of a significant fraction of units from biological networks performing a certain task, we cannot infer the loss function that the brain is optimizing since in principle, numerous different loss functions can have the same/similar minima (Fig. 7 top). In other words, there is typically a many-to-one mapping between loss functions and some point in state space where the functions have a minimum. Some of the different grid models from deep learning and first principles show that this is possible [12, 2, 46, 54, 7]. Conversely, given a reasonable optimization problem that we select based on an organism's ecological niche, we cannot infer a single solution (and thus build truly predictive single-cell tuning models), since there exist several minima to that optimization problem (Fig. 7 bottom). In other words, there is typically a one-to-many mapping from a loss function to its set of solutions. To break this degeneracy of multiple minima and arrive at truly predictive models or a better understanding of the brain's optimization problems, we must acknowledge, understand, and model the specific inductive biases and constraints present in the biological system we are trying to model. It is untenable to expect success without doing so. This is what we refer to as an informal neural 'no-free-lunch'. In these cases, these constraints and biases are as informative as other design principles, such as the loss function.

What can we learn from DL models about brain circuits without considering and studying biological inductive biases? Population-wide low-dimensional latent representations and dynamics that arise as necessary for solving difficult problems are possibly sufficiently robust and abstracted to be predictive of population dynamics in a neural circuit without addition of detailed biological constraints. This can explain successes of the population-level analyses of the visual pathway [28, 3], as well as the population-level low-dimensional dynamics of circuits solving inference tasks [50, 43, 1, 53]. In these cases, the emergent solutions are fundamental features of any systems that solve the task, and by construction need not be specific to brains.

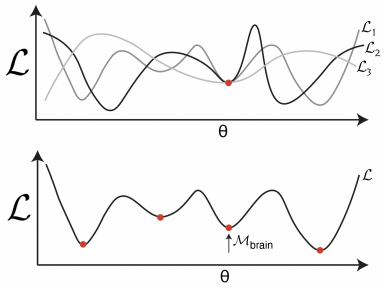

Figure 7: **Challenges in achieving the two central claims of recent DL models of neuroscience**: Top: Building a model that replicates observed neural responses does not guarantee that the loss function used is the brain's objective, as multiple objectives can share a solution. Bottom: Training a network on a plausible loss function or even the correct loss function need not yield the solution the brain has selected because the loss function may have multiple minima, of which the brain selects one based on its constraints, while an ANN selects another, based on the optimization technique used.

Returning to grid cells, we have conclusively shown that they do not generically arise in networks trained to path-integrate. This raises the question of what different additional architectural, hyperparameter, and constraint choices had to be made in previous papers [2, 46, 47] to obtain grid-like tuning, given that they used a path integration objective with Gaussian place cell targets, requirements that are reasonable but not sufficient to obtain grid-like tuning.

Theoretical work on grid cell representations [16, 48, 35] suggests that the following two factors are important features of the grid cell code: a very large coding range and the related property of robustness/intrinsic error correcting coding. Adding these properties to the loss are likely to be a more principled way to obtain grid-like fields rather than by hand-designing a biologically unrealistic DoG or DoS readout. Consistent with this, the very large amount of dropout or other noise required in [2, 12] suggests that the coding-theoretic insight on intrinsic error-correction properties of grid cells and their related large capacity may indeed be central ingredients to produce grid cells.

In fact, there are a number of key properties of the grid cell code elucidated earlier by theoretical arguments, including but not limited to path integration, that we hypothesize may form a sufficient and biologically relevant set for the emergence of grid cells: 1) non-negative activations; 2) equivariant population responses (i.e., the population response always lies on a hypersphere); 3) a path integrating (PI) code or in other words, translation invariant representations [17, 7]; 4) high representational capacity [6, 48, 34]; 5) intrinsic error-correcting capabilities [6, 48]; and finally, 6) uniformly distributed (whitened) and low spatial information per cell, such that the total spatial information of the grid code should be equally distributed across all cells.

We suggest that several of these properties (e.g. excluding the more specific property 3)) are extremely general features of neural codes, and incorporating them into ANN models of neural circuits could increase their ability to make de novo rather than post hoc predictions, and remove the need to use synthetic and unrealistic choices with fine hyperparameter tuning (such as DOG and DOS readouts).

In sum, the findings here argue that the central contribution of scientific interest when building and studying ANN models of the brain that reproduce specific tuning curves is not that the model produced the curves (after all, given the expressive power of deep networks, it is no surprise that training them to effectively generate a given tuning will in fact succeed), but rather to explore and carefully characterize conditions and constraints under which the particular tuning does and does not emerge, to consider whether the constraints align with biological constraints, and to consider what principles can be extracted from them. When not predicting specific tuning curves, but rather population-level response dynamics, the results are expected to be more robust to specific implementational choices and thus more durable.

References

[1] Emergence of dynamically reconfigurable hippocampal responses by learning to perform probabilistic spatial reasoning. *biorxiv*.

[2] Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J. Chadwick, Thomas Degris, Joseph Modayil, Greg Wayne, Hubert Soyer, Fabio Viola, Brian Zhang, Ross Goroshin, Neil Rabinowitz, Razvan Pascanu, Charlie Beattie, Stig Petersen, Amir Sadik, Stephen Gaffney, Helen King, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, and Dharshan Kumaran. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, May 2018.

[3] Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, May 2019. Publisher: American Association for the Advancement of Science.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July 2020. arXiv: 2005.14165.

[5] Yoram Burak and Ila Fiete. Do We Understand the Emergent Dynamics of Grid Cell Activity? *Journal of Neuroscience*, 26(37):9352–9354, September 2006. Publisher: Society for Neuroscience Section: Journal Club.

[6] Yoram Burak and Ila R Fiete. Unpublished observations. 2008.

[7] Yoram Burak and Ila R. Fiete. Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. *PLOS Computational Biology*, 5(2):e1000291, February 2009. Publisher: Public Library of Science.

[8] Yoram Burak and Ila R Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS Comput Biol*, 5(2):e1000291, Feb 2009.

[9] Malcolm G. Campbell, Samuel A. Ocko, Caitlin S. Mallory, Isabel I. C. Low, Surya Ganguli, and Lisa M. Giocomo. Principles governing the integration of landmark and self-motion cues in entorhinal cortical codes for navigation. *Nature Neuroscience*, 21(8):1096–1106, August 2018. Number: 8 Publisher: Nature Publishing Group.

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *NIPS Workshop Deep Learning and Representation Learning*, December 2014. Number: arXiv:1412.3555 arXiv:1412.3555 [cs].

[11] Jasmine Collins, Jascha Sohl-Dickstein, and David Sussillo. Capacity and Trainability in Recurrent Neural Networks. *International Conference on Learning Representations*, March 2017. Number: arXiv:1611.09913 arXiv:1611.09913 [cs, stat].

[12] Christopher J Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *International Conference on Learning Representations*, page 19, 2018.

[13] Tamir Eliav, Shir R. Maimon, Johnatan Aljadeff, Misha Tsodyks, Gily Ginosar, Liora Las, and Nachum Ulanovsky. Multiscale representation of very large environments in the hippocampus of flying bats. *Science*, 372(6545):eabg4020, May 2021. Publisher: American Association for the Advancement of Science.

[14] Jeffrey L. Elman. Finding Structure in Time. *Cognitive Science*, 14(2):179–211, 1990. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1402_1.

[15] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO. *arXiv:2005.12729 [cs, stat]*, May 2020. arXiv: 2005.12729.

[16] Ila R Fiete, Yoram Burak, and Ted Brookings. What grid cells convey about rat location. *J Neurosci*, 28(27):6858–71, Jul 2008.

[17] Mark C Fuhs and David S Touretzky. A spin glass model of path integration in rat medial entorhinal cortex. *J Neurosci*, 26(16):4266–4276, 2006.

[18] Marianne Fyhn, Torkel Hafting, Alessandro Treves, May-Britt Moser, and Edvard I Moser. Hippocampal remapping and grid realignment in entorhinal cortex. *Nature*, 446(7132):190–194, 2007.

[19] Richard J. Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A. Baas, Benjamin A. Dunn, May-Britt Moser, and Edvard I. Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895):123–128, February 2022. Number: 7895 Publisher: Nature Publishing Group.

[20] Richard J. Gardner, Li Lu, Tanja Wernle, May-Britt Moser, and Edvard I. Moser. Correlation structure of grid cells is preserved during sleep. *Nature Neuroscience*, 22(4):598–608, April 2019. Number: 4 Publisher: Nature Publishing Group.

[21] Joshua I. Glaser, Ari S. Benjamin, Raeed H. Chowdhury, Matthew G. Perich, Lee E. Miller, and Konrad P. Kording. Machine Learning for Neural Decoding. *eNeuro*, 7(4), July 2020. Publisher: Society for Neuroscience Section: Research Article: Methods/New Tools.

[22] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, August 2005. Number: 7052 Publisher: Nature Publishing Group.

[23] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Lecture 6e-RMSProp.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[25] Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A Closer Look at Deep Policy Gradients. *arXiv:1811.02553 [cs, stat]*, May 2020. arXiv: 1811.02553.

[26] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. Number: 7873 Publisher: Nature Publishing Group.

[27] I. Kanitscheider and I. R. Fiete. Training recurrent networks to generate hypotheses about how the brain solves hard navigation problems. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[28] Alexander J. E. Kell, Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, May 2018.

[29] Mikail Khona, Sarthak Chandra, and Ila R. Fiete. From smooth cortical gradients to discrete modules: spontaneous and topologically robust emergence of modularity in grid cells. *bioRxiv*, page 2021.10.28.466284, January 2022.

[30] Timothy D. Kim, Thomas Z. Luo, Jonathan W. Pillow, and Carlos D. Brody. Inferring Latent Dynamics Underlying Neural Population Activity via Neural Differential Equations. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5551–5561. PMLR, July 2021. ISSN: 2640-3498.

[31] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, January 2017. Number: arXiv:1412.6980 arXiv:1412.6980 [cs].

[32] Ashok Litwin-Kumar, Kameron Decker Harris, Richard Axel, Haim Sompolinsky, and L. F. Abbott. Optimal Degrees of Synaptic Connectivity. *Neuron*, 93(5):1153–1164.e7, March 2017.

[33] Jesse A. Livezey, Kristofer E. Bouchard, and Edward F. Chang. Deep learning as a tool for neural data analysis: Speech classification and cross-frequency coupling in human sensorimotor cortex. *PLOS Computational Biology*, 15(9):e1007091, September 2019. Publisher: Public Library of Science.

[34] A. Mathis, A. Herz, and M. Stemmler. Optimal population codes for space: grid cells outperform place cells. *Neural Comp.*, 24:2280–2317, 2012.

[35] Alexander Mathis, Andreas V. M. Herz, and Martin Stemmler. Optimal Population Codes for Space: Grid Cells Outperform Place Cells. *Neural Computation*, 24(9):2280–2317, September 2012.

[36] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, September 2018. Number: 9 Publisher: Nature Publishing Group.

[37] Aran Nayebi, Alexander Attinger, Malcolm Campbell, Kiah Hardcastle, Isabel Low, Caitlin S Mallory, Gabriel Mel, Ben Sorscher, Alex H Williams, Surya Ganguli, Lisa Giocomo, and Dan Yamins. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 12167–12179. Curran Associates, Inc., 2021.

[38] J. O'Keefe and J. Dostrovsky. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34:171–175, 1971. Place: Netherlands Publisher: Elsevier Science.

[39] Talmo D. Pereira, Diego E. Aldarondo, Lindsay Willmore, Mikhail Kislin, Samuel S.-H. Wang, Mala Murthy, and Joshua W. Shaevitz. Fast animal pose estimation using deep neural networks. *Nature Methods*, 16(1):117–125, January 2019. Number: 1 Publisher: Nature Publishing Group.

[40] P. D. Rich, H.-P. Liaw, and A. K. Lee. Large environments reveal the statistical structure governing hippocampal representations. *Science*, 345(6198):814–817, August 2014.

[41] Muhammad Saif-ur Rehman, Robin Lienkämper, Yaroslav Parpaley, Jörg Wellmer, Charles Liu, Brian Lee, Spencer Kellis, Richard Andersen, Ioannis Iossifidis, Tobias Glasmachers, and Christian Klaes. SpikeDeeptector: a deep-learning based method for detection of neural spiking activity. *Journal of Neural Engineering*, 16(5):056003, July 2019. Publisher: IOP Publishing.

[42] Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67, January 2021. Number: 1 Publisher: Nature Publishing Group.

[43] Rylan Schaeffer, Mikail Khona, Leenoy Meshulam, Brain Laboratory International, and Ila Fiete. Reverse-engineering recurrent neural network solutions to a hierarchical inference task for mice. *Advances in Neural Information Processing Systems*, 33:4584–4596, 2020.

[44] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016. Number: 7587 Publisher: Nature Publishing Group.

[45] Trygve Solstad, Charlotte N. Boccara, Emilio Kropff, May-Britt Moser, and Edvard I. Moser. Representation of Geometric Borders in the Entorhinal Cortex. *Science*, 322(5909):1865–1868, December 2008. Publisher: American Association for the Advancement of Science.

[46] Ben Sorscher, Gabriel C Mel, Surya Ganguli, and Samuel A Ocko. A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in Neural Information Processing Systems*, page 18, 2019.

[47] Ben Sorscher, Gabriel C. Mel, Samuel A. Ocko, Lisa Giocomo, and Surya Ganguli. A unified theory for the computational and mechanistic origins of grid cells. Technical report, bioRxiv, December 2020. Section: New Results Type: article.

[48] Sameet Sreenivasan and Ila Fiete. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nat Neurosci*, 14(10):1330–7, Sep 2011.

[49] Hanne Stensola, Tor Stensola, Trygve Solstad, Kristian Frøland, May-Britt Moser, and Edvard I. Moser. The entorhinal grid map is discretized. *Nature*, 492(7427):72–78, December 2012. Number: 7427 Publisher: Nature Publishing Group.

[50] David Sussillo and Omri Barak. Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation*, 25(3):626–649, March 2013.

[51] S.G. Trettel, J.B. Trimper, E. Hwaun, I.R. Fiete, and L.L. Colgin. Grid cell co-activity patterns during sleep reflect spatial overlap of grid fields during active behaviors. *Nat Neurosci*, 22(4):609–617, 04 2019.

[52] George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard E. Turner, Zoubin Ghahramani, and Sergey Levine. The Mirage of Action-Dependent Baselines in Reinforcement Learning. *arXiv:1802.10031 [cs, stat]*, November 2018. arXiv: 1802.10031.

[53] Jakob Voigts, Ingmar Kanitscheider, Nicholas J. Miller, Enrique H. S. Toloza, Jonathan P. Newman, Ila R. Fiete, and Mark T. Harnett. Spatial reasoning via recurrent neural dynamics in mouse retrosplenial cortex. *biorxiv*, April 2022.

[54] James C. R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E. J. Behrens. The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, 183(5):1249–1263.e23, November 2020.

[55] K.J. Yoon, M.A. Buice, R. Barry, C.and Hayman, N. Burgess, and I.R. Fiete. Specific evidence of low-dimensional continuous attractor dynamics in grid cells. *Nat Neurosci*, 16(8):1077–84, Aug 2013.

9 Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not feel our paper has potential negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We haven't yet had time to collect this information, but we will add this information to the final version if accepted.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]