ARE GRID CELLS AN EFFICIENT ENCODING OF SPACE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Since their discovery over 20 years ago, grid cells have attracted considerable theoretical interest. Although it is widely agreed they subserve path integration (the tracking of position from self-motion cues), the precise reason for their distinctive activity remains debated. Previous work has generated grid-like representations by optimising objectives motivated from both the efficient coding of spatial position and path integration, often encouraging properties such as conformal isometry, translation-invariant dynamics or high coding capacity. These studies have identified important conditions under which grid-like codes emerge. However, whether grid cells emerge from directly optimising for efficient coding alone, across a broad class of possible representations, remains an open question. To address this question, we consider a natural task-level formulation of efficient coding: under biologically plausible constraints, we seek representations that minimise the mean squared error in the presence of noise. Specifically, we consider an encoder that receives noisy displacement updates and whose activity itself is corrupted by noise, deriving conditions under which this path-based problem reduces to the classical efficient-coding problem of encoding a source distribution. To analyse this problem, we employ the Ziv–Zakai lower bound, an information-theoretic tool that captures both local decoding precision and catastrophic errors arising from periodic spatial codes. Along the way, we see that several properties previously found to produce grid cells are theoretically advantageous within our coding framework. Grid-cell-like representations then emerge from the optimisation, providing evidence that grid cells constitute a highly efficient spatial code. More broadly, our results support the idea that the problem of efficient spatial coding is tightly related to the one of path integration, this suggests that task-relevant compression can favour representations that reflect the symmetries and compositional structure of the task.

1 INTRODUCTION

Following their Nobel Prize winning discovery in 2005 Hafting et al. (2005), grid cells have attracted the attention of many theorists trying to identify their normative role. Given the influence of efficient coding as a normative theory in neuroscience, a natural starting point has been the efficient coding hypothesis, first proposed over 60 years ago by Barlow and Attneave Barlow et al. (1961); Attneave (1954). Building on the earlier development of information theory by Claude Shannon Shannon (1948), Barlow proposed that sensory neurons should encode stimuli so as to maximise the information they convey, compared to the theoretical limit (defined through the channel capacity Cover et al. (1991)). Since then the efficient coding hypothesis has taken various forms, depending on ones interpretation on what it means to be 'efficient', this includes (though is not limited to) mutual-information maximisation Linsker (1988); Laughlin et al. (1981) and minimisation of mean squared decoding error Doi & Lewicki (2014); Olshausen & Field (1996), paired with resource constraints such as limited neuron number, bounded response ranges, or metabolic cost Blättler & Hahnloser (2011); Schreiber et al. (2002).

The first theoretical considerations of grid cells from an efficient coding perspective identified several coding advantages of grid-like representations, such as the efficiency of hexagonal lattices through packing arguments and the use of multiple modules as a residue-number-like coding system Sreenivasan & Fiete (2011); Wei et al. (2015); Mathis et al. (2012). These works particularly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

focused on these advantages when compared to place cells. However, these analyses generally assumed a modular code rather than giving an exact problem statement from which grid cells emerge as being optimal from all or a broad class of solutions. Thus, although many coding advantages of grid-cell representations were identified, it remains unclear whether grid-like codes would emerge from efficient coding principles alone.

This gap has become particularly salient following studies showing that grid-like representations can emerge in neural networks trained for path integration Banino et al. (2018); Cueva & Wei (2018). Building on these findings, recent work has sought minimal assumptions or constraints that give rise to grid-like representations, including conformal isometry, translational covariance, high coding capacity, and path invariance Dorrell et al. (2022); Schaeffer et al. (2023); Schøyen et al. (2025). These properties are typically motivated by their mechanistic role in path integration or by their coding advantages, rather than by efficient-coding principles alone. This success, together with the limited success of purely efficient-coding models, has suggested that incorporating the role of grid cells in path integration may be necessary to reproduce grid-like representations in simulations. This has led some authors to conclude that grid cells do not provide the most efficient spatial code, and that place cells or multi-field place cells are instead optimal Dorrell & Whittington (2026); Sengupta et al. (2018). However, this conclusion conflicts with the aforementioned theoretical work and other numerical investigations Dietrich et al. (2024); Stemmler et al. (2015). Thus, we explore this tension addressing the gap by asking: what is the most efficient neural code for two-dimensional space?

We do so through a natural task-level formulation of the efficient-coding hypothesis: minimising mean squared error (MSE) in the presence of noise. Whereby, the encoder representing a neuronal population, receives noisy displacement updates, and its activity is further corrupted before reaching a downstream decoder. We additionally impose biologically plausible non-negativity and energy constraints on the encoder. Rather than only considering path-independent encoders of space, we derive conditions under which the problem can be restricted to path-invariant codes without loss of performance. We then see when this reduces our original path-integration problem set-up to the more familiar efficient coding of a source distribution. Linked to recent observations that Fisher information may not adequately characterise coding performance Kriegeskorte & Wei (2021); Malerba et al. (2025), we use the Ziv–Zakai lower bound (ZZLB) Ziv & Zakai (1969) as a tractable and informative proxy for MSE. This provides a geometric interpretation of efficient coding as a balance between stretching the coding manifold and avoiding castestrophic errors between distant locations, particularly important when considering grid cells due to their characteristic lattice tuning curves. In one dimension, this trade-off can be intuitively visualised as a rope-packing problem.

Crucially, many of the aforementioned properties linked to successfully producing grid-like representations in simulations Xu et al. (2025); Pettersen et al. (2024) also emerge as theoretically advantageous in our efficient coding set-up. This both provides a task level normative motivation as well as a new solely efficient coding perspective. Furthermore, the emergence of grid-like representations in our simulations suggests that grid cells do constitute an efficient, even if not uniquely optimal, spatial code. More broadly, linking efficient coding to the task of path integration provides an example of how compression may give rise to useful structural abstractions. An efficient encoding should seek to discard task-irrelevant variation Tishby et al. (2000) and exploit symmetries Wolff (2019), thus efficient representations may also support compositional generalisation Whittington et al. (2020); Riveland et al. (2026); Conklin et al. (2026); Hinton & Van Camp (1993). Or more succinctly, compression is comprehension Chaitin (2006).

2 Problem Formulation

We consider an animal navigating two-dimensional space, beginning at a known starting position $\boldsymbol{x}_0 \in \mathbb{R}^2$, established, for example, via landmark recalibration. Over discrete time intervals $t \in \{1, \dots, T\}$, it undergoes true displacement increments $\Delta\boldsymbol{x}_t \in \mathbb{R}^2$ (drawn from a distribution $p(\Delta\boldsymbol{x}_t)$ such as a gaussian random walk), yielding the true final position:

$$\boldsymbol{x}_T = \boldsymbol{x}_0 + \sum_{t=1}^{T} \Delta\boldsymbol{x}_t. \tag{1}$$

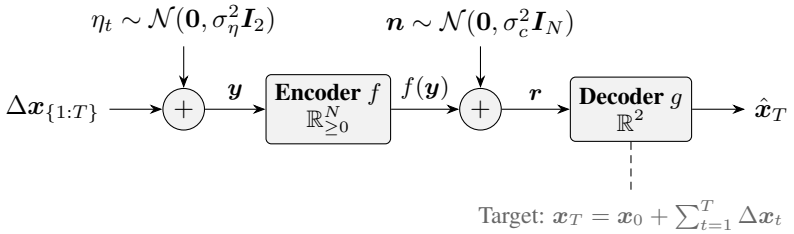


Figure 1: Schematic of the noisy path integration coding model. Noisy self-motion cues $\boldsymbol{y}_t$ are mapped by population encoder f, later corrupted by channel noise $\boldsymbol{n}$, and decoded by g to form the final position estimate $\hat{\boldsymbol{x}}_T$.

Upstream Path Integration. At each time step $t \in \{1, \ldots, T\}$, self-motion velocity cues are corrupted by independent additive Gaussian noise, these are the received observations:

$$\boldsymbol{y}_t = \Delta \boldsymbol{x}_t + \eta_t, \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}\left(\mathbf{0}, \sigma_\eta^2 \boldsymbol{I}_2\right). \tag{2}$$

The complete observation history across all T steps is collected into the stacked column vector $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_T^\top]^\top \in \mathbb{R}^{2T}$.

Neural Encoding and Downstream Channel. The trajectory history is encoded into the steady-state firing rates of an N-neuron population via the encoding $f : \mathbb{R}^{2T} \to \mathbb{R}^N_{\geq 0}$. The neural response is then corrupted by noise, for example due to a noisy synaptic channel; this is modelled as an independent additive isotropic Gaussian channel:

$$\boldsymbol{r} = f(\boldsymbol{y}) + \boldsymbol{n}, \qquad \boldsymbol{n} \sim \mathcal{N}\left(\mathbf{0}, \sigma_c^2 \boldsymbol{I}_N\right), \tag{3}$$

The decoder $g : \mathbb{R}^N \to \mathbb{R}^2$ then forms an estimate of the true endpoint $\hat{\boldsymbol{x}}_T = g(\boldsymbol{r})$ from the noisy neural code. We seek an encoder f and decoder g that minimise the mean squared error (MSE), subject to biologically plausible non-negativity and energy constraints:

$$\boxed{\begin{aligned} MMSE = \min_{f,\, g} \quad & \mathbb{E}\Big[\|\boldsymbol{x}_T - g(f(\boldsymbol{y}) + \boldsymbol{n})\|_2^2\Big] \\ \text{s.t.} \quad & 0 \leq f(\boldsymbol{y}), \quad \|f(\boldsymbol{y})\|_2^2 \leq C, \ \forall \boldsymbol{y} \in \mathbb{R}^{2T}, \end{aligned}} \tag{4}$$

When the means and variances are finite, the MMSE estimator is uniquely defined by the posterior mean. Therefore, for any fixed encoder f, the MSE-optimal decoder is the posterior mean:

$$g(f(\boldsymbol{y}) + \boldsymbol{n}) \triangleq \mathbb{E}[\boldsymbol{x}_T \mid f(\boldsymbol{y}) + \boldsymbol{n}]. \tag{5}$$

3 Path Invariance

So far, we have not imposed that the encoder $f(\boldsymbol{y})$ produces a self-consistent mapping of position, $\boldsymbol{x}$. As perhaps the optimal encoder preserves information about the trajectory taken to reach a given point, thereby mapping identical spatial coordinates to different neural encodings due to this information being useful to the downstream decoder. Below, we will prove when in our setup a path independent encoder is optimal.

To formalise this, define the cumulative noisy path-integrated position $\boldsymbol{q}_T \in \mathbb{R}^2$ and the residual vector $\boldsymbol{d} = [\boldsymbol{d}_1^\top, \ldots, \boldsymbol{d}_T^\top]^\top$ by

$$\boldsymbol{q}_T \triangleq \boldsymbol{x}_0 + \sum_{t=1}^T \boldsymbol{y}_t = \boldsymbol{x}_T + \boldsymbol{z}_T, \qquad \boldsymbol{z}_T \triangleq \sum_{t=1}^T \eta_t \sim \mathcal{N}\left(\mathbf{0}, T\sigma_\eta^2 \boldsymbol{I}_2\right), \tag{6}$$

and

$$\boldsymbol{d}_t \triangleq \boldsymbol{y}_t - \frac{1}{T}\sum_{s=1}^T \boldsymbol{y}_s = \boldsymbol{y}_t - \frac{\boldsymbol{q}_T - \boldsymbol{x}_0}{T}, \qquad t = 1, \ldots, T. \tag{7}$$

Thus, $\boldsymbol{q}_T$ gives the total noisy cumulative position, while $\boldsymbol{d}_t$ gives the deviation of the t^{th} displacement update from the average. Collecting the residuals into $\boldsymbol{d} = [\boldsymbol{d}_1^\top, \ldots, \boldsymbol{d}_T^\top]^\top$, the mapping between the raw observation history and the pair of summary statistics, for a fixed known $\boldsymbol{x}_0$, is bijective and invertible:

$$\boldsymbol{y} \longleftrightarrow (\boldsymbol{q}_T, \boldsymbol{d}). \tag{8}$$

Proposition 1. *$\boldsymbol{q}_T$ is a sufficient statistic for the true endpoint, $\boldsymbol{x}_T$, given observations, $\boldsymbol{y}$, when:*

$$\boxed{\boldsymbol{d} \perp\!\!\!\perp \boldsymbol{x}_T \mid \boldsymbol{q}_T}. \tag{9}$$

Proof. Factoring the joint conditional distribution of the equivalent representation $(\boldsymbol{q}_T, \boldsymbol{d})$ given $\boldsymbol{x}_T$ yields

$$p(\boldsymbol{y} \mid \boldsymbol{x}_T) = p(\boldsymbol{q}_T, \boldsymbol{d} \mid \boldsymbol{x}_T) = p(\boldsymbol{q}_T \mid \boldsymbol{x}_T)\, p(\boldsymbol{d} \mid \boldsymbol{q}_T, \boldsymbol{x}_T). \tag{10}$$

If condition equation 9 holds, then $p(\boldsymbol{d} \mid \boldsymbol{q}_T, \boldsymbol{x}_T) = p(\boldsymbol{d} \mid \boldsymbol{q}_T)$, which does not depend on the target parameter $\boldsymbol{x}_T$. Hence,

$$p(\boldsymbol{y} \mid \boldsymbol{x}_T) = p(\boldsymbol{q}_T \mid \boldsymbol{x}_T)\, p(\boldsymbol{d} \mid \boldsymbol{q}_T), \tag{11}$$

Thus by the Fisher–Neyman factorisation theorem, $\boldsymbol{q}_T$ is a sufficient statistic for $\boldsymbol{x}_T$.

In appendix A we prove Eq.9 holds for the example of a Gaussian random walk. □

Path-invariant encoding. The Blackwell-Rao Theorem Blackwell (1947); Rao et al. (1945) tells us that the MSE of any estimator cannot worsen when variations are averaged out by conditioning on a sufficient statistic. Thus when Eq. 9 holds, it is hard to justify why the encoder should include path-dependent variations from $\boldsymbol{d}_T$, as doing so would require finite neural energy, while the Blackwell-Rao Theorem implies that averaging out such variations for a given $\boldsymbol{q}_T$ cannot worsen and may improve decoding performance. While intuitive this argument is not precise as the decoder does not have direct access to $\boldsymbol{y}$. Thus in Appendix B we consider this more carefully, explicitly proving that the MSE achievable by an encoder allowed to depend only on $\boldsymbol{q}_T$ is as low as that achievable when the encoder is free to depend on both $\boldsymbol{q}_T$ and $\boldsymbol{d}_T$.

4 Relation to Efficient Coding

For the subclass of path-integration problems satisfying Eq. 9, we may restrict attention to encoders based only on $\boldsymbol{q}_T$. Defining the MMSE estimator as $h(\boldsymbol{q}) \triangleq \mathbb{E}[\boldsymbol{x}_T \mid \boldsymbol{q}_T = \boldsymbol{q}]$, we write:

$$\mathbb{E}\big[\|\widehat{\boldsymbol{x}}_T - \boldsymbol{x}_T\|_2^2\big] = \mathbb{E}\big[\|\boldsymbol{x}_T - h(\boldsymbol{q}_T)\|_2^2\big] + \mathbb{E}\big[\|\widehat{\boldsymbol{x}}_T - h(\boldsymbol{q}_T)\|_2^2\big]\,, \tag{12}$$

since $\mathbb{E}[\boldsymbol{x}_T - h(\boldsymbol{q}_T) \mid \boldsymbol{q}_T] = \mathbf{0}$, and hence the cross term vanishes.

Tweedie's formula Robbins (1992) gives $h(\boldsymbol{q}) = \boldsymbol{q} + \sigma_z^2 \nabla_{\boldsymbol{q}} \log p_{\boldsymbol{q}_T}(\boldsymbol{q})$, where $\sigma_z^2 = T\sigma_\eta^2$. Thus, $h(\boldsymbol{q}) \approx \boldsymbol{q}$ whenever the correction term is negligible. This occurs, for example, when the marginal density $p_{\boldsymbol{q}_T}$ is locally uniform, so that $\nabla_{\boldsymbol{q}} \log p_{\boldsymbol{q}_T}(\boldsymbol{q}) \approx \mathbf{0}$, or when the sensory noise is sufficiently small. In the Gaussian random-walk model, this corresponds to $\sigma_\eta^2/(\sigma_\eta^2 + \sigma_v^2) \approx 0$, equivalently $\sigma_\eta^2 \ll \sigma_v^2$. Since $\boldsymbol{x}_T = \boldsymbol{q}_T - \boldsymbol{z}_T$, it follows that

$$\mathbb{E}\big[\|\boldsymbol{x}_T - h(\boldsymbol{q}_T)\|_2^2\big] = \mathbb{E}\big[\|\boldsymbol{q}_T - \boldsymbol{z}_T - h(\boldsymbol{q}_T)\|_2^2\big] \approx \mathbb{E}\big[\|\boldsymbol{z}_T\|_2^2\big] = 2T\sigma_\eta^2. \tag{13}$$

Consequently, the total path-integration MSE is approximately

$$\mathbb{E}\big[\|\widehat{\boldsymbol{x}}_T - \boldsymbol{x}_T\|_2^2\big] \approx \underbrace{\mathbb{E}\big[\|\widehat{\boldsymbol{x}}_T - \boldsymbol{q}_T\|_2^2\big]}_{\text{downstream channel MSE}} + \underbrace{2T\sigma_\eta^2}_{\text{sensory-drift MSE}}\,. \tag{14}$$

Therefore, reducing the path-integration problem reduces to the efficient coding of $\boldsymbol{q}_T$, plus a constant sensory-drift term. Holding when the sensory noise is small compared to the variations in the expected occupancy distribution and Eq. 9 is satisfied.

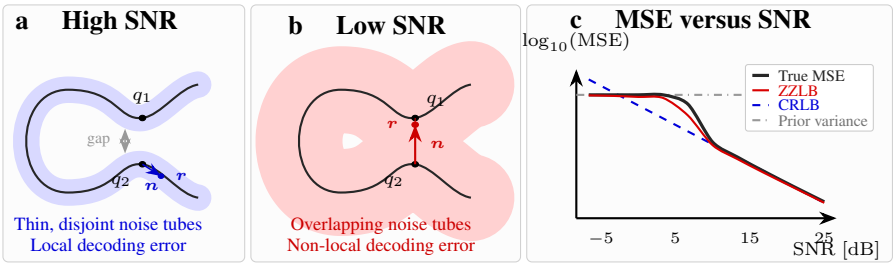


Figure 2: **Local and global errors in a folded neural code. (a)** At high SNR, large errors are rare and the CRLB is informative. **(b)** At low SNR, noise can cross between nearby folds, causing large non-local errors, we see a sudden decrease in performance known as the threshold effect. **(c)** Unlike the CRLB the ZZLB captures the threshold effect.

5 Efficient Coding in the Presence of Global Errors

Having established the conditions under which the general path-integration problem reduces to the efficient coding of $\boldsymbol{q}_T$, we now focus on this reduced problem. The variable $\boldsymbol{q}_T$ is the cumulative sum of the noisy self-motion cues and, when sensory noise is small relative to the variation in the prior position, this closely approximates the best internal estimate of position $h(\boldsymbol{q})$. As a starting point, we assume that $\boldsymbol{q}_T$ is uniformly distributed over a disc. This corresponds to an animal assigning equal probability, and hence equal coding importance, to all locations within a large, empty circular enclosure. In this case the previous reduction is not exact: it is most accurate in the bulk of the enclosure and becomes less accurate within a distance of order σ_z from its boundary. As, for example, the animal should not interpret a noisy value of $\boldsymbol{q}_T$ lying beyond the enclosure as evidence that it managed to pass through the wall. However, interestingly boundaries, small enclosures, and reward locations are known to distort grid-cell firing patterns Butler et al. (2019); Hägglund et al. (2019). Moreover, the entorhinal cortex contains border cells as well as grid cells Solstad et al. (2008). Thus although now for simplicity we disregard such boundary effects, in the future we would like to explore if a more careful consideration may explain these observed features. Finally, proposed links between grid cells and memory suggest a need for high-capacity, noise-robust representations Chandra et al. (2025). A uniform source model, which assigns equal importance to every location, therefore seems reasonable from this perspective.

The Threshold Effect. Immediately following his foundational paper on information theory (Shannon, 1948), Shannon studied communication in the presence of noise (Shannon, 1949), offering a geometric explanation for the "threshold effect" first observed in FM radio. Shannon visualised the channel input as a continuous mapping that embeds the source space into a higher-dimensional signal space. At high signal-to-noise ratios (SNR), decoding error is dominated by small, local perturbation tangential to the encoding manifold. However, as noise increases past a critical threshold, the decoder may confuse an entirely different fold of the manifold, causing the confusion of points widely separated in the source space. This transition from small local errors to large global errors allowed him to explain the threshold effect, which we illustrate in Figure 2.

Despite this many classical normative neuroscience models largely fail to capture it. Many efficient coding accounts quantify fidelity using Fisher information connected via the Cramér–Rao lower bound (CRLB). Because Fisher information is a purely local metric, it is inherently blind to the global geometry of the encoding manifold and thus non-local self-proximity (Kriegeskorte & Wei, 2021). As a result, the CRLB cannot detect when distant points of the source space are mapped dangerously close together in neural space, severely underestimating the error in the presence of globally ambiguous representations. Grid cells through their quintessential lattice encoding structure provide a poignant example. By looking at a single module of grid cells, one cannot distinguish distantly spatially seperated points, with the lattice encoding folding the encoding back upon itself. Thus, to accurately reflect the coding performance of grid cells we motivate the need of a method capable of balancing the effect of both local and global errors. Thus we introduce the Ziv Zakai Lower Bound (ZZLB) (Ziv & Zakai, 1969) to this problem, providing to our knowledge, the first application of the ZZLB to the efficient coding of neural representations.

Ziv–Zakai Lower Bound By using the ZZLB as a more tractable surrogate for the MSE. We can complete the reduction of the original path-integration objective Eq. 4 with the efficient coding of a source uniformly distributed over the disc $\mathcal{D}_R \triangleq \{\boldsymbol{q} \in \mathbb{R}^2 : \|\boldsymbol{q}\|_2 \le R\}$, such that $\boldsymbol{q}_T \sim \mathrm{Unif}(\mathcal{D}_R)$, as:

$$\boxed{\begin{aligned} MZZLB = \min_{f} \quad & \frac{1}{\pi^2 R^2} \sum_{i=1}^{2} \int_0^{\pi} \int_0^{2R} h \int_{\Omega_{h,i}^{\phi}} \mathcal{Q}\left(\frac{d_i^{\phi}(\boldsymbol{q}, h)}{2\sigma_c} \right) d\boldsymbol{q}\, dh\, d\phi \\ \text{s.t.} \quad & f_j(\boldsymbol{q}) \ge 0, \qquad \|f(\boldsymbol{q})\|_2^2 \le C, \quad \boldsymbol{q} \in \mathcal{D}_R, \end{aligned}} \tag{15}$$

Here,

$$\begin{gathered} \boldsymbol{e}_1^{\phi} \triangleq (\cos\phi, \sin\phi)^{\top}, \quad \boldsymbol{e}_2^{\phi} \triangleq (-\sin\phi, \cos\phi)^{\top}, \quad \mathcal{Q}(z) \triangleq \frac{1}{\sqrt{2\pi}} \int_z^{\infty} e^{-t^2/2}\, dt, \\ \Omega_{h,i}^{\phi} \triangleq \mathcal{D}_R \cap (\mathcal{D}_R - h\boldsymbol{e}_i^{\phi}), \quad d_i^{\phi}(\boldsymbol{q}, h) \triangleq \|f(\boldsymbol{q} + h\boldsymbol{e}_i^{\phi}) - f(\boldsymbol{q})\|_2. \end{gathered} \tag{16}$$

The ZZLB can be understood as lower-bounding the MSE by considering binary hypothesis tests between pairs of source points $\boldsymbol{q}$ and $\boldsymbol{q} + h\boldsymbol{e}_i^{\phi}$. The term $\mathcal{Q}\Big(d_i^{\phi}(\boldsymbol{q}, h)/(2\sigma_c)\Big)$ is the probability of confusing their noisy neural encodings, while the factor h weights this error by their separation in source space. The bound then averages over all valid pairs. Its full derivation is given in Appendix C.

Jensen's Inequality We will now try to understand what properties our efficient coding objective favours. To do so we perform Jensens Inequality onto the ZZLB to provide some intuition on how we can minimise this bound. A more intuitive example (which served as our inspiration) can be seen when we perform the same trick on the CRLB for a 1D version of our problem as in Appendix D, this allowed us to understand the CRLB as either being lowered by increasing the length of the line in the encoding space or by having a more uniform 'speed'. This represents a simplified version of the application of Hölders inequality to the CRLB in Wang et al. (2013), for the case of uniform distributions. We define:

$$\Phi(y) \triangleq \mathcal{Q}\left(\frac{\sqrt{y}}{2\sigma_c} \right), \qquad y \ge 0. \tag{17}$$

This function is decreasing and strictly convex for $y > 0$; see Appendix E. For each direction $\boldsymbol{e}_i^{\phi}$ and separation h, define the mean squared neural separation:

$$\overline{d_{i,\phi}^2}(h) \triangleq \frac{1}{A_{\cap}(h)} \int_{\Omega_{h,i}^{\phi}} \left[d_i^{\phi}(\boldsymbol{q}, h) \right]^2 d\boldsymbol{q}, \tag{18}$$

Where $A_{\cap}(h)$ is the area of $\Omega_{h,i}^{\phi}$. Thus by Jensen's inequality the ZZLB, Eq.15, can be lower bounded by:

$$\mathrm{MZZLB} \ge \frac{1}{\pi^2 R^2} \sum_{i=1}^{2} \int_0^{\pi} \int_0^{2R} h A_{\cap}(h) \mathcal{Q}\left(\frac{\sqrt{\overline{d_{i,\phi}^2}(h)}}{2\sigma_c} \right) dh\, d\phi. \tag{19}$$

Equality holds if and only if $\left[d_i^{\phi}(\boldsymbol{q}, h)\right]^2 = \overline{d_{i,\phi}^2}(h)$, for all $\boldsymbol{q} \in \Omega_{h,i}^{\phi}$. We can interpret this result as follows: for a fixed $\overline{d_{i,\phi}^2}(h)$, if one wishes to minimise the ZZLB optimisation problem eq.15, it is advantageous to minimise variations in $\boldsymbol{q}$ as much as possible. That is, the distance between any two neural encodings should ideally be independent of the absolute positions they encode, depending only on their relative separation. In Appendix F we apply the same argument across displacement directions we see this favours isotropic codes.

From our geometric perspective of efficient coding Fig. 2, squared neural separations form a resource that protects against noisy confusions. As seen through Jensens inequality, convexity favours distributing this resource uniformly where feasible, plus clearly to reduce confusion we should increase its overall amount. We next relate this resource to the centred neural energy.

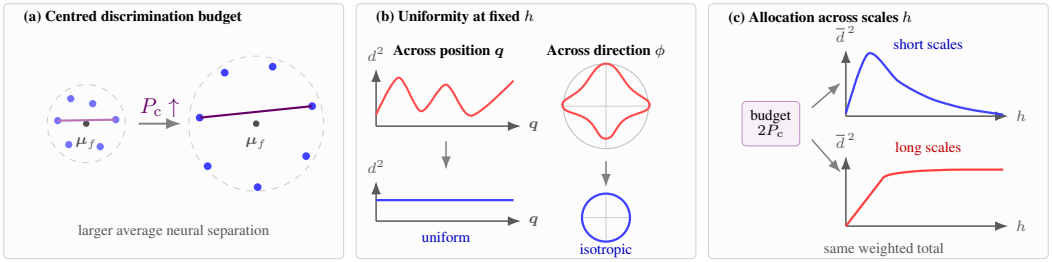


Figure 3: (a) Increasing the centred neural energy P_c enlarges the total coding budget. (b) For each displacement scale h, symmetry and convexity favour neural separations that vary as little as possible across absolute source position $\boldsymbol{q}$ and displacement direction ϕ. (c) The fixed total budget must also be allocated optimally across source separations h.

Centred Energy as a Global Coding Budget We define the centred neural energy as

$$P_c \triangleq \mathbb{E}\left[\|f(\boldsymbol{q}) - \boldsymbol{\mu}_f\|_2^2\right] = P_f - \|\boldsymbol{\mu}_f\|_2^2, \tag{20}$$

where $\boldsymbol{\mu}_f \triangleq \mathbb{E}[f(\boldsymbol{q})]$ and $P_f \triangleq \mathbb{E}\left[\|f(\boldsymbol{q})\|_2^2\right]$ denote the mean neural activity and mean neural energy, respectively. For independent and identically distributed position $\boldsymbol{q}, \boldsymbol{q}'$, we have

$$\boxed{\mathbb{E}\left[\|f(\boldsymbol{q}') - f(\boldsymbol{q})\|_2^2\right] = 2P_c} \tag{21}$$

Thus, the centred neural energy is one half of the mean squared neural chord length between two independently sampled source positions. Letting $H = \|\boldsymbol{q}' - \boldsymbol{q}\|_2$ and $\Theta \in [0, \pi)$ denote the orientation of the displacement $\boldsymbol{q}' - \boldsymbol{q}$ modulo π. Then we see our 'coding resource':

$$\overline{d_\phi^2}(h) \triangleq \mathbb{E}\left[\|f(\boldsymbol{q}') - f(\boldsymbol{q})\|_2^2, \,\middle|\, H = h,\, \Theta = \phi\right]. \tag{22}$$

For $\boldsymbol{q}$ and $\boldsymbol{q}'$ sampled uniformly from a disc of radius R, it can be shown that:

$$p_{H,\Theta}(h, \phi) = \frac{2hA_\cap(h)}{(\pi R^2)^2}, \qquad 0 \le h \le 2R, \quad 0 \le \phi < \pi, \tag{23}$$

Where $p_{H,\Theta}(h, \phi)$ represents the probability that $\boldsymbol{q}$ and $\boldsymbol{q}'$, when drawn iid from $\mathrm{Unif}(\mathcal{D}_R)$,, are separated by a vector with $H = h$ and $\Theta = \phi$. Thus from Eq.21, we may write:

$$2P_c = \int_0^\pi \int_0^{2R} p_{H,\Theta}(h, \phi)\, \overline{d_\phi^2}(h)\, dh\, d\phi. \le 2C \tag{24}$$

Where the final inequality follows from $P_c \le P_f$, due to our energy constraint $P_f \le C$.

We find this analysis very enlightening. Equation 21 identifies centred neural energy ($2P_c$) as the relevant global coding budget and the quantity $\overline{d_\phi^2}(h)$ describes how this budget is distributed across source separations and orientations. Then building on the preceding applications of Jensen's inequality (for $\boldsymbol{q}$ and ϕ), a final application to the ZZLB makes the role of this budget explicit:

$$\boxed{\mathrm{MZZLB} \ge R^2 \mathcal{Q}\left(\frac{\sqrt{2P_c}}{2\sigma_c}\right).}$$

Increasing P_c lowers this bound. At fixed P_c, moving the encoder towards isotropy, translational covariance and an optimal displacement allocation $\overline{d_\phi^2}(h)$ tightens the bound and therefore lowers the intended ZZLB objective.

Equality would require squared neural separations to equal $2P_c$ for all source pairs. This would correspond to a spatially uniform, isotropic encoding with constant squared neural separation across all displacement scales. Other than for the trivial constant encoding, this cannot be achieved by a self consistent continuous encoding. This motivates a qualitative conjecture: an efficient code should produce separations that rise from zero at short source distances and become as uniform as possible at larger distances. The red curve in Fig. 3 illustrates this.

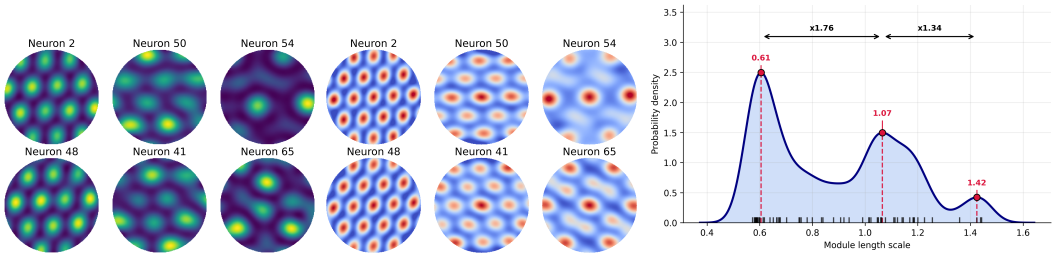


Figure 4: From left to right: Rate maps and spatial autocorrelograms for 6 selected neurons and the length scale distribution of the autocorrelogram lattices. The rate maps for all 73 neurons can be found in Appendix L.

Together, these considerations motivate four guiding principles for reducing the ZZLB objective in Eq. 15, illustrated in Fig. 3:

1. **Coding budget maximisation:** maximise P_c to increase the total coding budget;
2. **Scale allocation:** distribute this budget optimally across displacement scales h;
3. **Spatial uniformity:** distribute the allocation at each (h, ϕ) uniformly across source positions $\boldsymbol{q}$, thereby tightening the Jensen bound in Eq. 19.
4. **Isotropy:** distribute the budget uniformly across ϕ at a given h, as shown in Appendix F.

Each of these principles can be linked to features that have previously been shown to produce grid-like representations. The global coding budget, or centred neural energy, is closely related to the capacity losses used in Schaeffer et al. (2023); Pettersen et al. (2024); we examine this connection in Appendix G. Plus exact correspondence with the closely related work Whittington et al. (2022). The allocation of this resource across separations h underlies objectives relating neural separation to source separation Schaeffer et al. (2023); Dorrell et al. (2022). Spatial uniformity is closely related to translation covariance, while isotropy requires equivalent treatment of all directions. Together, these two properties can be connected to conformal isometry Xu et al. (2025); Schøyen et al. (2025), as explored in Appendix H.

Thus, many features previously found to promote grid-cell-like representations emerge naturally from a single task-level objective: the efficient coding of a uniform spatial source.

6 Numerical Experiments

Building upon our previous theoretical intuitions, we will now try to optimise our problem statement Eq. 15. Unlike, for example, the channel-capacity or rate-distortion problems, zero-delay joint source–channel coding (JSCC) is generally non-convex, and a closed form solution for the optimal encoder is only known for the trivial case of the scalar gaussian source channel pair Akyol et al. (2014). Numerical approaches to the Zero-Delay JSCC problem broadly follow three routes: optimising the parameters of a structured encoder Hekland et al. (2009); Hu et al. (2011), considered a discretised version of the problem using power-constrained channel-optimised vector quantisation Floor et al. (2007); Karlsson & Skoglund (2010), and recent deep learning based approaches Bourtsoulatze et al. (2019).

We adopt the first approach, parameterising the encoder by a continuous translation-covariant code. This equates to when Eq. 19 tight, thus when the third of our four principles of an efficient code **Spatial uniformity** is exactly met. Such codes on $\mathbb{R}^2$ have the form:

$$f(\boldsymbol{q}) = \boldsymbol{c} + \sum_{m=1}^{\frac{N-1}{2}} a_m \left[\boldsymbol{u}_m \cos(\boldsymbol{k}_m^\top \boldsymbol{q}) + \boldsymbol{v}_m \sin(\boldsymbol{k}_m^\top \boldsymbol{q}) \right], \tag{25}$$

where $\boldsymbol{k}_m \in \mathbb{R}^2$ are wavevectors, $a_m \geq 0$, and the vectors $\{\boldsymbol{u}_m, \boldsymbol{v}_m\}_{m=1}^M$ are orthonormal and orthogonal to $\boldsymbol{c}$, we enforce non-negativity separately. The proof follows from Bochners Theorem

Rudin et al. (1962) and can be found in Appendix I. Interestingly, the version of this encoding for a 1D source has been studied extensively in the JSCC literature Buvarp et al. (2023); Costa et al. (2004), motivated by its performance due to its constant curvature. This aligns with our geometric intuition: sharp bends can bring otherwise distant source points close together in neural space, making confusion more likely. Another interesting observation is that these curves are geodesics on flat tori Taylor et al. (2013), an intriguing connection to the toroidal structure observed in grid-cell population activity Gardner et al. (2022).

Given the parametric form of Eq. 25, we then use Adam Kingma & Ba (2014) to optimise the constant component $\boldsymbol{c}$, the amplitudes $\{a_m\}_{m=1}^M$, the wavevectors $\{\boldsymbol{k}_m\}_{m=1}^M$, and the orthonormal directions $\{\boldsymbol{u}_m, \boldsymbol{v}_m\}_{m=1}^M$; full details are provided in Appendix J. We see the results of our optimisation in Fig. 4, all 73 rate maps for this example and the table of the parameters used can be found in Appendix L . As demonstrated, we successfully recover both hexagonal lattices and multiple modules of varying scales, separated by a factor of approximately 1.5, in agreement with the experimental data Hafting et al. (2005).

7 Discussion and Limitations

Summary In this work, we revisited the efficient coding hypothesis as an explanation for grid cells. Within a general model for coding self-motion cues, we first established when we can restrict ourselves to path-invariant encoders, and further when the problem reduces to the classical efficient coding of a source distribution. Guided by our geometric intuition of coding, we then introduced and motivated the Ziv–Zakai lower bound (ZZLB). This allowed us to understand the efficient coding of a uniform source through four principles: isotropy, translation covariance, high centred neural power, and the allocation of neural separation across displacement scales. Many closely related features have previously been found to promote grid-cell representations. Thus, a key contribution is to connect these properties to a single task-level normative motivation, rather than treating them as imposed design choices. It also connects features such as path invariance and translation covariance to efficient coding, which previously have only been motivated by their mechanistic usefulness for path integration and route planning. Finally, our simulations produce multiple modules of spatially shifted hexagonal grid cells, showing that such structure can arise within a purely efficient-coding formulation.

Limitations The most significant limitation in our approach is that we impose spatial uniformity exactly to simplify the optimisation. Although theoretically motivated, this restriction may exclude more efficient codes that trade some spatial uniformity for improvements in the other three principles. Future work could test broader encoder classes, for example using deep JSCC Bourtsoulatze et al. (2019). Thus, while encouraging, our results do not definitively prove that grid cells are the optimal encoding; particularly since a similar parametrisation was previously motivated by path integration Dorrell et al. (2022), implying path integration may still be a necessary component. Second, although theoretically motivated, the encoding which minimises the ZZLB need not also minimise the MSE, which represents the most natural quantification of efficient coding performance. We explore this further through a numerical comparison in Appendix K . Furthermore, as discussed in Section 5, our reduction neglects enclosure-boundary effects. Modelling these effects could help explain border cells and the grid-lattice distortions observed in small or irregular enclosures. The choice of source distribution also merits discussion. A uniform source provides a natural baseline when no location is preferred, and it is unclear to us which distribution would be most biologically appropriate. We argue that the success of this choice in producing grid-like codes is insightful in itself and may point, for example, to links between grid cells and memory Huber (2025). Another key limitation is our rate-based model. The precise timing of spikes also carries information Reifenstein et al. (2012), particularly given the theta phase precession observed in grid cells Hafting et al. (2008). Extending the framework by considering the codes in time, would then be an interesting direction.

Overall, we hope these results encourage a reconsideration of the role of efficient coding in the search for a normative explanation for grid cells. Beyond grid cells, the parallels between the properties favoured by efficient coding and those motivated by structural abstraction support the idea that the most efficient representation is also the one that exploits the symmetries and invariances of the inputs.

8 AI DISCLOSURE

AI tools were utilised to propose alternative phrasing during drafting and to provide code for figures and schematics. Additionally, AI contributed for mathematical proofs in particular making the authors aware of the noise-outsourcing trick (Austin (2015)) used in Appendix B. The authors reviewed, verified, and take full responsibility for all final text, figures, and mathematical derivations.

REFERENCES

Emrah Akyol, Kumar B Viswanatha, Kenneth Rose, and Tor A Ramstad. On zero-delay source-channel coding. *IEEE Transactions on Information Theory*, 60(12):7473–7489, 2014.

Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.

Tim Austin. Exchangeable random measures. In *Annales de l'IHP Probabilités et statistiques*, volume 51, pp. 842–861, 2015.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233, 1961.

David Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, pp. 105–110, 1947.

Florian Blättler and Richard HR Hahnloser. An efficient coding hypothesis links sparsity and selectivity of neural responses. *PloS one*, 6(10):e25506, 2011.

Eirina Bourtsoulatze, David Burth Kurka, and Deniz Gündüz. Deep joint source-channel coding for wireless image transmission. *IEEE Transactions on Cognitive Communications and Networking*, 5(3):567–579, 2019.

William N Butler, Kiah Hardcastle, and Lisa M Giocomo. Remembered reward locations restructure entorhinal spatial maps. *Science*, 363(6434):1447–1452, 2019.

Anders M Buvarp, Robert M Taylor, Kumar Vijay Mishra, Lamine Mili, and Amir I Zaghloul. Constant curvature curve tube codes for low-latency analog error correction. *IEEE Transactions on Information Theory*, 69(12):7738–7754, 2023.

Gregory Chaitin. The limits of reason. *Scientific American*, 294(3):74–81, 2006.

Sarthak Chandra, Sugandha Sharma, Rishidev Chaudhuri, and Ila Fiete. Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature*, 638(8051):739–751, 2025.

Henry C Conklin, Tom Hosking, Tan Yi-Chern, Julian Gold, Jonathan D Cohen, Thomas L Griffiths, Max Bartolo, and Seraphina Goldfarb-Tarrant. Learning is forgetting: Llm training as lossy compression. *arXiv preprint arXiv:2604.07569*, 2026.

Sueli I Rodrigues Costa, Marcelo Muniz, Edson Agustini, and Reginaldo Palazzo. Graphs, tessellations, and perfect codes on flat tori. *IEEE Transactions on Information Theory*, 50(10): 2363–2377, 2004.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

Christopher J Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *arXiv preprint arXiv:1803.07770*, 2018.

Robin Dietrich, Nicolai Waniek, Martin Stemmler, and Alois Knoll. Grid codes vs. multi-scale, multi-field place codes for space. *Frontiers in Computational Neuroscience*, 18:1276292, 2024.

Eizaburo Doi and Michael S Lewicki. A simple model of optimal population coding for sensory systems. *PLoS computational biology*, 10(8):e1003761, 2014.

William Dorrell and James CR Whittington. If grid cells are the answer, what is the question? a review of normative grid cell theory. *arXiv preprint arXiv:2601.12424*, 2026.

William Dorrell, Peter E Latham, Timothy EJ Behrens, and James CR Whittington. Actionable neural representations: Grid cells from minimal constraints. *arXiv preprint arXiv:2209.15563*, 2022.

Pal Anders Floor, Tor A Ramstad, and Niklas Wernersson. Power constrained channel optimized vector quantizers used for bandwidth expansion. In *2007 4th International Symposium on Wireless Communication Systems*, pp. 667–671. IEEE, 2007.

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, 2022.

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, 2005.

Torkel Hafting, Marianne Fyhn, Tora Bonnevie, May-Britt Moser, and Edvard I Moser. Hippocampus-independent phase precession in entorhinal grid cells. *Nature*, 453(7199):1248–1252, 2008.

Martin Hägglund, Maria Mørreaunet, May-Britt Moser, and Edvard I Moser. Grid-cell distortion along geometric borders. *Current Biology*, 29(6):1047–1054, 2019.

Fredrik Hekland, Pal Anders Floor, and Tor A Ramstad. Shannon-kotel-nikov mappings in joint source-channel coding. *IEEE Transactions on Communications*, 57(1):94–105, 2009.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13, 1993.

Yichuan Hu, Javier Garcia-Frias, and Meritxell Lamarca. Analog joint source-channel coding using non-linear curves and mmse decoding. *IEEE Transactions on Communications*, 59(11):3016–3026, 2011.

David E Huber. 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, 2025.

Johannes Karlsson and Mikael Skoglund. Optimized low-delay source-channel-relay mappings. *IEEE Transactions on Communications*, 58(5):1397–1404, 2010.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature reviews neuroscience*, 22(11):703–718, 2021.

Simon Laughlin et al. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch*, 36(910-912):51, 1981.

Ralph Linsker. An application of the principle of maximum information preservation to linear systems. *Advances in neural information processing systems*, 1, 1988.

Simone Blanco Malerba, Mirko Pieropan, Yoram Burak, and Rava Azeredo da Silveira. Random compressed coding with neurons. *Cell Reports*, 44(3), 2025.

Alexander Mathis, Andreas VM Herz, and Martin Stemmler. Optimal population codes for space: grid cells outperform place cells. *Neural computation*, 24(9):2280–2317, 2012.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Markus Pettersen, Vemund Sigmundson Schøyen, Mattis Dalsætra Østby, Anders Malthe-Sørenssen, and Mikkel Elle Lepperød. Self-supervised grid cells without path integration. *bioRxiv*, pp. 2024–05, 2024.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

C Radhakrishna Rao et al. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*, 37(3):81–91, 1945.

Eric T Reifenstein, Richard Kempter, Susanne Schreiber, Martin B Stemmler, and Andreas VM Herz. 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 (16):6301–6306, 2012.

Danilo Jimenez Rezende and Fabio Viola. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.

Reidar Riveland, Alexandre Pouget, and Laura Driscoll. The compositionality continuum as a principle for studying the neural basis of intelligence. *Nature neuroscience*, pp. 1–14, 2026.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pp. 388–394. Springer, 1992.

Walter Rudin et al. *Fourier analysis on groups*, volume 12. Interscience publishers New York, 1962.

Rylan Schaeffer, Mikail Khona, Tzuhsuan Ma, Cristobal Eyzaguirre, Sanmi Koyejo, and Ila Fiete. Self-supervised learning of representations for space generates multi-modular grid cells. *Advances in Neural Information Processing Systems*, 36:23140–23157, 2023.

Vemund Sigmundson Schøyen, Kosio Beshkov, Markus Borud Pettersen, Erik Hermansen, Konstantin Holzhausen, Anders Malthe-Sørenssen, Marianne Fyhn, and Mikkel Elle Lepperød. Hexagons all the way down: Grid cells as a conformal isometric map of space. *PLOS Computational Biology*, 21(2):e1012804, 2025.

Susanne Schreiber, Christian K Machens, Andreas VM Herz, and Simon B Laughlin. Energy-efficient coding with discrete stochastic events. *Neural computation*, 14(6):1323–1346, 2002.

Anirvan Sengupta, Cengiz Pehlevan, Mariano Tepper, Alexander Genkin, and Dmitri Chklovskii. Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks. *Advances in neural information processing systems*, 31, 2018.

Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

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, 2008.

Sameet Sreenivasan and Ila Fiete. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature neuroscience*, 14(10):1330–1337, 2011.

Martin Stemmler, Alexander Mathis, and Andreas VM Herz. Connecting multiple spatial scales to decode the population activity of grid cells. *Science Advances*, 1(11):e1500816, 2015.

Robert M Taylor, Lamine Mili, and Amir Zaghloul. Packing tubes on tori: An efficient method for low snr analog error correction. In *2013 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2013.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Zhuo Wang, Alan A Stocker, and Daniel D Lee. Optimal neural population codes for high-dimensional stimulus variables. *Advances in neural information processing systems*, 26, 2013.

Xue-Xin Wei, Jason Prentice, and Vijay Balasubramanian. A principle of economy predicts the functional architecture of grid cells. *Elife*, 4:e08362, 2015.

James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

James CR Whittington, Will Dorrell, Surya Ganguli, and Timothy EJ Behrens. Disentanglement with biological constraints: A theory of functional cell types. *arXiv preprint arXiv:2210.01768*, 2022.

J Gerard Wolff. Information compression as a unifying principle in human learning, perception, and cognition. *Complexity*, 2019(1):1879746, 2019.

Dehong Xu, Ruiqi Gao, Wenhao Zhang, Xue-Xin Wei, and Yingnian Wu. On conformal isometry of grid cells: Learning distance-preserving position embedding. In *International Conference on Learning Representations*, volume 2025, pp. 30739–30761, 2025.

Jacob Ziv and Moshe Zakai. Some lower bounds on signal parameter estimation. *IEEE transactions on Information Theory*, 15(3):386–391, 1969.

A Sufficiency for the Gaussian Random Walk

We show that the cumulative noisy position $\boldsymbol{q}_T$ is sufficient for the true endpoint $\boldsymbol{x}_T$ under a Gaussian random walk.

Lemma 1. *Suppose*

$$\Delta \boldsymbol{x}_t \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v), \qquad \boldsymbol{\eta}_t \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma_\eta^2 \boldsymbol{I}_2), \tag{26}$$

where the displacement and noise sequences are mutually independent. Define

$$\boldsymbol{y}_t = \Delta \boldsymbol{x}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{d}_t = \boldsymbol{y}_t - \frac{1}{T}\sum_{k=1}^{T} \boldsymbol{y}_k, \tag{27}$$

and let $\boldsymbol{d} = [\boldsymbol{d}_1^\top, \dots, \boldsymbol{d}_T^\top]^\top$. *Then,*

$$\boldsymbol{d} \perp\!\!\!\perp (\boldsymbol{x}_T, \boldsymbol{q}_T), \tag{28}$$

and hence

$$\boldsymbol{d} \perp\!\!\!\perp \boldsymbol{x}_T \mid \boldsymbol{q}_T. \tag{29}$$

Consequently, $\boldsymbol{q}_T$ *is sufficient for* $\boldsymbol{x}_T$ *given* $\boldsymbol{y}$.

Proof. Define

$$\boldsymbol{w} \triangleq \begin{bmatrix} \Delta \boldsymbol{x}_1 \\ \vdots \\ \Delta \boldsymbol{x}_T \\ \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_T \end{bmatrix}. \tag{30}$$

By independence, $\boldsymbol{w}$ is multivariate Gaussian. Since $\boldsymbol{d}$, $\boldsymbol{x}_T$ and $\boldsymbol{q}_T$ are linear transformations of $\boldsymbol{w}$, they are jointly Gaussian. For jointly Gaussian random variables zero correlation implies independence, hence we will consider the relevant covariances below:

Let

$$\boldsymbol{\Sigma}_y \triangleq \boldsymbol{\Sigma}_v + \sigma_\eta^2 \boldsymbol{I}_2. \tag{31}$$

Then it is easy to see that:

$$\mathrm{Cov}(\boldsymbol{y}_j, \boldsymbol{y}_k) = \delta_{jk} \boldsymbol{\Sigma}_y. \tag{32}$$

Since

$$\boldsymbol{q}_T = \boldsymbol{x}_0 + \sum_{j=1}^{T} \boldsymbol{y}_j, \tag{33}$$

we have, for each t,

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{d}_t, \boldsymbol{q}_T) &= \mathrm{Cov}\left(\boldsymbol{y}_t - \frac{1}{T}\sum_{k=1}^{T}\boldsymbol{y}_k, \sum_{j=1}^{T}\boldsymbol{y}_j\right) \\ &= \boldsymbol{\Sigma}_y - \frac{1}{T}\left(T\boldsymbol{\Sigma}_y\right) \\ &= \mathbf{0}.\end{aligned} \tag{34}$$

Similarly, since

$$\boldsymbol{x}_T = \boldsymbol{x}_0 + \sum_{j=1}^{T}\Delta\boldsymbol{x}_j, \tag{35}$$

and the noise is independent of the displacements,

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{d}_t, \boldsymbol{x}_T) &= \mathrm{Cov}\left(\boldsymbol{y}_t - \frac{1}{T}\sum_{k=1}^{T}\boldsymbol{y}_k, \sum_{j=1}^{T}\Delta\boldsymbol{x}_j\right) \\ &= \mathrm{Cov}\left(\Delta\boldsymbol{x}_t - \frac{1}{T}\sum_{k=1}^{T}\Delta\boldsymbol{x}_k, \sum_{j=1}^{T}\Delta\boldsymbol{x}_j\right) \\ &= \boldsymbol{\Sigma}_v - \frac{1}{T}\left(T\boldsymbol{\Sigma}_v\right) \\ &= \mathbf{0}.\end{aligned} \tag{36}$$

These identities hold for every t, so

$$\mathrm{Cov}\left(\boldsymbol{d}, \begin{bmatrix}\boldsymbol{x}_T \\ \boldsymbol{q}_T\end{bmatrix}\right) = \mathbf{0}. \tag{37}$$

As noted the variables are jointly Gaussian, so zero correlation implies independence. Hence

$$\boldsymbol{d} \perp\!\!\!\perp (\boldsymbol{x}_T, \boldsymbol{q}_T). \tag{38}$$

Thus

$$\boldsymbol{d} \perp\!\!\!\perp \boldsymbol{x}_T \mid \boldsymbol{q}_T. \tag{39}$$

The sufficiency of $\boldsymbol{q}_T$ follows from Proposition 1. □

B Path Invariance optimality proof

Lemma 2 (Path invariance under a pointwise power constraint)**.** *When $\boldsymbol{q}_T$ is a sufficient statistic for $\boldsymbol{x}_T$, then:*

$$\begin{aligned}&\inf_{\substack{f(\boldsymbol{q}_T, \boldsymbol{d}_T),\, g \\ \|f(\boldsymbol{q}_T, \boldsymbol{d}_T)\|_2^2 \le P_0 \text{ a.s.}}} \mathbb{E}\Big[\|\boldsymbol{x}_T - g(f(\boldsymbol{q}_T, \boldsymbol{d}_T) + \boldsymbol{\eta})\|_2^2\Big] \\ &= \inf_{\substack{\widetilde{f}(\boldsymbol{q}_T),\, \widetilde{g} \\ \|\widetilde{f}(\boldsymbol{q}_T)\|_2^2 \le P_0 \text{ a.s.}}} \mathbb{E}\left[\left\|\boldsymbol{x}_T - \widetilde{g}\Big(\widetilde{f}(\boldsymbol{q}_T) + \boldsymbol{\eta}\Big)\right\|_2^2\right].\end{aligned} \tag{40}$$

Proof. We use the noise-outsourcing lemma (Austin, 2015, Lemma 3.1), which allows us to re-express $\boldsymbol{d}_T$ as a deterministic function of $\boldsymbol{q}_T$ and an independent uniform "noise", while preserving the same conditional distribution. Hence, define

$$U \sim \mathrm{Unif}[0, 1], \tag{41}$$

where U is independent of both $\boldsymbol{x}_T$ and $\boldsymbol{q}_T$. The noise-outsourcing lemma then allows us to write

$$\widetilde{\boldsymbol{d}}_T = \tau(\boldsymbol{q}_T, U), \tag{42}$$

whereby $\widetilde{\boldsymbol{d}}_T$ has the same conditional distribution $p(\boldsymbol{d}_T \mid \boldsymbol{q}_T)$ as $\boldsymbol{d}_T$.

As U is independent of both $\boldsymbol{q}_T$ and $\boldsymbol{x}_T$, and $\boldsymbol{q}_T$ is a sufficient statistic for $\boldsymbol{x}_T$, we have

$$\begin{aligned} p(\boldsymbol{x}_T, \boldsymbol{q}_T, \widetilde{\boldsymbol{d}}_T) &= p(\boldsymbol{x}_T, \boldsymbol{q}_T)\, p(\widetilde{\boldsymbol{d}}_T \mid \boldsymbol{q}_T) \\ &= p(\boldsymbol{x}_T, \boldsymbol{q}_T)\, p(\boldsymbol{d}_T \mid \boldsymbol{q}_T) \\ &= p(\boldsymbol{x}_T, \boldsymbol{q}_T, \boldsymbol{d}_T). \end{aligned} \tag{43}$$

Hence,

$$(\boldsymbol{x}_T, \boldsymbol{q}_T, \widetilde{\boldsymbol{d}}_T) \overset{d}{=} (\boldsymbol{x}_T, \boldsymbol{q}_T, \boldsymbol{d}_T). \tag{44}$$

For each fixed $u \in [0, 1]$, define the deterministic path-invariant encoder

$$f_u(\boldsymbol{q}) = f(\boldsymbol{q}, \tau(\boldsymbol{q}, u))\,. \tag{45}$$

We first note that revealing the value of U to the decoder can only reduce the minimum MSE. Therefore,

$$\begin{aligned} \inf_{g} \mathbb{E}\Big[\|\boldsymbol{x}_T - g(f_U(\boldsymbol{q}_T) + \boldsymbol{\eta})\|_2^2\Big] &\geq \mathbb{E}_U\left[\inf_{g_u} \mathbb{E}\Big[\|\boldsymbol{x}_T - g_u(f_u(\boldsymbol{q}_T) + \boldsymbol{\eta})\|_2^2 \,\Big|\, U = u\Big]\right] \\ &\geq \inf_{u\in[0,1]} \inf_{g_u} \mathbb{E}\Big[\|\boldsymbol{x}_T - g_u(f_u(\boldsymbol{q}_T) + \boldsymbol{\eta})\|_2^2\Big]\,. \end{aligned} \tag{46}$$

Whereby, the second inequality comes by selecting the best-performing value of u. If u is fixed and known to the decoder, then this just equates to using a deterministic encoder based upon $\boldsymbol{q}_T$ alone.

Finally, path-invariant encoders $\widetilde{f}(\boldsymbol{q}_T)$ are clearly a special case of the more general encoders $f(\boldsymbol{q}_T, \boldsymbol{d}_T)$. Therefore,

$$\begin{aligned} &\inf_{\substack{f(\boldsymbol{q}_T,\boldsymbol{d}_T),\, g \\ \|f(\boldsymbol{q}_T,\boldsymbol{d}_T)\|_2^2 \leq P_0 \text{ a.s.}}} \mathbb{E}\Big[\|\boldsymbol{x}_T - g(f(\boldsymbol{q}_T, \boldsymbol{d}_T) + \boldsymbol{\eta})\|_2^2\Big] \\ &\leq \inf_{\substack{\widetilde{f}(\boldsymbol{q}_T),\, \widetilde{g} \\ \|\widetilde{f}(\boldsymbol{q}_T)\|_2^2 \leq P_0 \text{ a.s.}}} \mathbb{E}\left[\left\|\boldsymbol{x}_T - \widetilde{g}\Big(\widetilde{f}(\boldsymbol{q}_T) + \boldsymbol{\eta}\Big)\right\|_2^2\right]. \end{aligned} \tag{47}$$

Hence the two reverse inequality produce equality, completing the proof. □

C Ziv–Zakai Lower Bound Derivation

The Ziv–Zakai lower bound provides a global lower bound on the MSE, $\mathbb{E}[\|\hat{\boldsymbol{q}} - \boldsymbol{q}\|_2^2]$. We follow the standard ZZLB derivation Ziv & Zakai (1969), adapted to our setting.

Model Let the source coordinate be uniformly distributed over the disc,

$$\boldsymbol{q} \sim \mathrm{Unif}(\mathcal{D}_R), \qquad \mathcal{D}_R \triangleq \left\{\boldsymbol{q} \in \mathbb{R}^2 : \|\boldsymbol{q}\|_2 \leq R\right\}, \tag{48}$$

so that

$$p_{\boldsymbol{q}}(\boldsymbol{q}) = \frac{1}{\pi R^2} 1_{\{\boldsymbol{q}\in\mathcal{D}_R\}}. \tag{49}$$

The observation model is

$$\boldsymbol{r} = f(\boldsymbol{q}) + \boldsymbol{n}, \qquad \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma_c^2 \mathbf{I}), \tag{50}$$

where $f : \mathcal{D}_R \to \mathbb{R}^N$ is the neural encoder. Let $\hat{\boldsymbol{q}} = \hat{\boldsymbol{q}}(\boldsymbol{r})$ be an arbitrary estimator.

Directional decomposition of the MSE For each $\phi \in [0, \pi)$, define the orthonormal directions

$$\boldsymbol{e}_1^{\phi} \triangleq (\cos\phi, \sin\phi)^{\top}, \qquad \boldsymbol{e}_2^{\phi} \triangleq (-\sin\phi, \cos\phi)^{\top}. \tag{51}$$

Let

$$\boldsymbol{\varepsilon} \triangleq \hat{\boldsymbol{q}} - \boldsymbol{q}, \qquad \varepsilon_i^{\phi} \triangleq (\boldsymbol{e}_i^{\phi})^{\top} \boldsymbol{\varepsilon}. \tag{52}$$

Since $\{\boldsymbol{e}_1^\phi, \boldsymbol{e}_2^\phi\}$ is an orthonormal basis,

$$\sum_{i=1}^{2} \boldsymbol{e}_i^\phi (\boldsymbol{e}_i^\phi)^\top = \mathbf{I}_2. \tag{53}$$

Therefore by using the cyclic property of the trace,

$$\begin{aligned} \mathbb{E}\big[\|\hat{\boldsymbol{q}} - \boldsymbol{q}\|_2^2\big] &= \operatorname{tr} \mathbb{E}\big[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big] \\ &= \sum_{i=1}^{2} \mathbb{E}\Big[(\varepsilon_i^\phi)^2\Big] \end{aligned} \tag{54}$$

for every ϕ. Averaging this identity over $\phi \in [0, \pi)$ gives:

$$\mathbb{E}\big[\|\hat{\boldsymbol{q}} - \boldsymbol{q}\|_2^2\big] = \frac{1}{\pi} \sum_{i=1}^{2} \int_0^\pi \mathbb{E}\Big[(\varepsilon_i^\phi)^2\Big]\, d\phi. \tag{55}$$

We consider this average as although the MSE is independent of the choice of ϕ for our orthonormal basis the ZZLB we derive is not, thus we average to consider these variations.

MSE in terms of tail probabilities For any non-negative random variable X,

$$\mathbb{E}[X] = \int_0^\infty \Pr(X \geq t)\, dt. \tag{56}$$

Taking $X = (\varepsilon_i^\phi)^2$ and substituting $t = h^2/4$ gives

$$\begin{aligned} \mathbb{E}\Big[(\varepsilon_i^\phi)^2\Big] &= \int_0^\infty \Pr\Big((\varepsilon_i^\phi)^2 \geq t\Big)\, dt \\ &= \frac{1}{2} \int_0^\infty h\, \Pr\left(|\varepsilon_i^\phi| \geq \frac{h}{2}\right) dh. \end{aligned} \tag{57}$$

The error-tail probability can be decomposed as

$$\begin{aligned} \Pr\left(|\varepsilon_i^\phi| \geq \frac{h}{2}\right) &= \Pr\left(\varepsilon_i^\phi > \frac{h}{2}\right) + \Pr\left(\varepsilon_i^\phi \leq -\frac{h}{2}\right) \\ &= \Pr\left((\boldsymbol{e}_i^\phi)^\top \hat{\boldsymbol{q}} > (\boldsymbol{e}_i^\phi)^\top \boldsymbol{q} + \frac{h}{2}\right) \\ &\quad + \Pr\left((\boldsymbol{e}_i^\phi)^\top \hat{\boldsymbol{q}} \leq (\boldsymbol{e}_i^\phi)^\top \boldsymbol{q} - \frac{h}{2}\right). \end{aligned} \tag{58}$$

Conditioning on the true source position gives

$$\begin{aligned} \Pr\left(|\varepsilon_i^\phi| \geq \frac{h}{2}\right) &= \int_{\mathbb{R}^2} p_{\boldsymbol{q}}(\boldsymbol{q}_0) \Pr\left((\boldsymbol{e}_i^\phi)^\top \hat{\boldsymbol{q}} > (\boldsymbol{e}_i^\phi)^\top \boldsymbol{q}_0 + \frac{h}{2}, \middle| \boldsymbol{q} = \boldsymbol{q}_0\right) d\boldsymbol{q}_0 \\ &\quad + \int_{\mathbb{R}^2} p_{\boldsymbol{q}}(\boldsymbol{q}_1) \Pr\left((\boldsymbol{e}_i^\phi)^\top \hat{\boldsymbol{q}} \leq (\boldsymbol{e}_i^\phi)^\top \boldsymbol{q}_1 - \frac{h}{2}, \middle| \boldsymbol{q} = \boldsymbol{q}_1\right) d\boldsymbol{q}_1. \end{aligned} \tag{59}$$

In the second integral, make the change of variables

$$\boldsymbol{q}_1 = \boldsymbol{q}_0 + h\boldsymbol{e}_i^\phi. \tag{60}$$

Relabelling $\boldsymbol{q}_0$ as $\boldsymbol{q}$ yields

$$\begin{aligned} \Pr\left(|\varepsilon_i^\phi| \geq \frac{h}{2}\right) = \int_{\mathbb{R}^2} \Bigg[& p_{\boldsymbol{q}}(\boldsymbol{q}) \Pr\left((\boldsymbol{e}_i^\phi)^\top \hat{\boldsymbol{q}} > (\boldsymbol{e}_i^\phi)^\top \boldsymbol{q} + \frac{h}{2}, \middle| \boldsymbol{q}_{\text{true}} = \boldsymbol{q}\right) \\ & + p_{\boldsymbol{q}}(\boldsymbol{q} + h\boldsymbol{e}_i^\phi) \Pr\left((\boldsymbol{e}_i^\phi)^\top \hat{\boldsymbol{q}} \leq (\boldsymbol{e}_i^\phi)^\top \boldsymbol{q} + \frac{h}{2}, \middle| \boldsymbol{q}_{\text{true}} = \boldsymbol{q} + h\boldsymbol{e}_i^\phi\right) \Bigg] d\boldsymbol{q}. \end{aligned} \tag{61}$$

Relation to binary hypothesis testing Fix a direction $\boldsymbol{e}_i^{\phi}$, a separation $h \geq 0$, and a point $\boldsymbol{q}$ such that both $\boldsymbol{q}$ and $\boldsymbol{q} + h\boldsymbol{e}_i^{\phi}$ belong to the source domain. Consider the binary hypotheses

$$\mathsf{H}_0 : \boldsymbol{q}_{\text{true}} = \boldsymbol{q}, \tag{62}$$

$$\mathsf{H}_1 : \boldsymbol{q}_{\text{true}} = \boldsymbol{q} + h\boldsymbol{e}_i^{\phi}. \tag{63}$$

We define an estimator-induced decision rule that selects H_0 when

$$(\boldsymbol{e}_i^{\phi})^{\top}\hat{\boldsymbol{q}} \leq (\boldsymbol{e}_i^{\phi})^{\top}\boldsymbol{q} + \frac{h}{2}, \tag{64}$$

and selects H_1 otherwise.

The prior probabilities assigned to the hypotheses are

$$\pi_0^{\phi}(\boldsymbol{q}, h) \triangleq \frac{p_{\boldsymbol{q}}(\boldsymbol{q})}{p_{\boldsymbol{q}}(\boldsymbol{q}) + p_{\boldsymbol{q}}(\boldsymbol{q} + h\boldsymbol{e}_i^{\phi})}, \tag{65}$$

$$\pi_1^{\phi}(\boldsymbol{q}, h) \triangleq \frac{p_{\boldsymbol{q}}(\boldsymbol{q} + h\boldsymbol{e}_i^{\phi})}{p_{\boldsymbol{q}}(\boldsymbol{q}) + p_{\boldsymbol{q}}(\boldsymbol{q} + h\boldsymbol{e}_i^{\phi})}. \tag{66}$$

The probability of error of this decision rule is

$$\begin{aligned} P_{\mathrm{e}}^{i,\phi}(\boldsymbol{q}, h) = \pi_0^{\phi}(\boldsymbol{q}, h)\Pr\left((\boldsymbol{e}_i^{\phi})^{\top}\hat{\boldsymbol{q}} > (\boldsymbol{e}_i^{\phi})^{\top}\boldsymbol{q} + \frac{h}{2}, \middle| \mathsf{H}_0\right) \\ + \pi_1^{\phi}(\boldsymbol{q}, h)\Pr\left((\boldsymbol{e}_i^{\phi})^{\top}\hat{\boldsymbol{q}} \leq (\boldsymbol{e}_i^{\phi})^{\top}\boldsymbol{q} + \frac{h}{2}, \middle| \mathsf{H}_1\right). \end{aligned} \tag{67}$$

This induced rule need not be optimal. Hence,

$$P_{\mathrm{e}}^{i,\phi}(\boldsymbol{q}, h) \geq P_{\min}^{i,\phi}(\boldsymbol{q}, h), \tag{68}$$

where $P_{\min}^{i,\phi}$ is the minimum achievable binary error probability. It follows that

$$\Pr\left(|\varepsilon_i^{\phi}| \geq \frac{h}{2}\right) \geq \int_{\mathbb{R}^2} \left[p_{\boldsymbol{q}}(\boldsymbol{q}) + p_{\boldsymbol{q}}(\boldsymbol{q} + h\boldsymbol{e}_i^{\phi})\right] P_{\min}^{i,\phi}(\boldsymbol{q}, h)\, d\boldsymbol{q}. \tag{69}$$

Restriction to the disc-overlap region Define the overlap region

$$\Omega_{h,i}^{\phi} \triangleq \left\{\boldsymbol{q} \in \mathbb{R}^2 : \boldsymbol{q} \in \mathcal{D}_R \text{ and } \boldsymbol{q} + h\boldsymbol{e}_i^{\phi} \in \mathcal{D}_R\right\} = \mathcal{D}_R \cap \left(\mathcal{D}_R - h\boldsymbol{e}_i^{\phi}\right). \tag{70}$$

For every $\boldsymbol{q} \in \Omega_{h,i}^{\phi}$,

$$p_{\boldsymbol{q}}(\boldsymbol{q}) = p_{\boldsymbol{q}}(\boldsymbol{q} + h\boldsymbol{e}_i^{\phi}) = \frac{1}{\pi R^2}. \tag{71}$$

Since the integrand in Eq. 69 is non-negative, restriction to $\Omega_{h,i}^{\phi}$ gives

$$\Pr\left(|\varepsilon_i^{\phi}| \geq \frac{h}{2}\right) \geq \frac{2}{\pi R^2} \int_{\Omega_{h,i}^{\phi}} P_{\min}^{i,\phi}(\boldsymbol{q}, h)\, d\boldsymbol{q}. \tag{72}$$

The largest possible separation between two points in $\mathcal{D}_R$ is its diameter, $2R$. Substitution into Eq. 57 therefore gives

$$\mathbb{E}\left[(\varepsilon_i^{\phi})^2\right] \geq \frac{1}{\pi R^2} \int_0^{2R} h \int_{\Omega_{h,i}^{\phi}} P_{\min}^{i,\phi}(\boldsymbol{q}, h)\, d\boldsymbol{q}\, dh. \tag{73}$$

Minimum error probability Under the two hypotheses, the observation distributions are

$$\boldsymbol{r} \mid \mathsf{H}_0 \sim \mathcal{N}\big(f(\boldsymbol{q}), \sigma_c^2\mathbf{I}\big), \tag{74}$$

$$\boldsymbol{r} \mid \mathsf{H}_1 \sim \mathcal{N}\Big(f(\boldsymbol{q} + h\boldsymbol{e}_i^{\phi}), \sigma_c^2\mathbf{I}\Big). \tag{75}$$

Define the neural separation

$$d_i^{\phi}(\boldsymbol{q}, h) \triangleq \left\| f(\boldsymbol{q} + h\boldsymbol{e}_i^{\phi}) - f(\boldsymbol{q}) \right\|_2. \tag{76}$$

For two equally likely Gaussian hypotheses with common covariance $\sigma_c^2\mathbf{I}$, the minimum error probability is

$$P_{\min}^{i,\phi}(\boldsymbol{q}, h) = \mathcal{Q}\left(\frac{d_i^{\phi}(\boldsymbol{q}, h)}{2\sigma_c} \right), \tag{77}$$

where

$$\mathcal{Q}(z) \triangleq \frac{1}{\sqrt{2\pi}} \int_z^{\infty} \exp\left(-\frac{t^2}{2} \right) dt. \tag{78}$$

Thus,

$$\mathbb{E}\Big[(\varepsilon_i^{\phi})^2 \Big] \geq \frac{1}{\pi R^2} \int_0^{2R} h \int_{\Omega_{h,i}^{\phi}} \mathcal{Q}\left(\frac{d_i^{\phi}(\boldsymbol{q}, h)}{2\sigma_c} \right) d\boldsymbol{q}\, dh. \tag{79}$$

Finally, substituting this directional bound into Eq. 55 gives

$$\boxed{\mathbb{E}\big[\|\hat{\boldsymbol{q}} - \boldsymbol{q}\|_2^2 \big] \geq \frac{1}{\pi^2 R^2} \sum_{i=1}^{2} \int_0^{\pi} \int_0^{2R} h \int_{\Omega_{h,i}^{\phi}} \mathcal{Q}\left(\frac{d_i^{\phi}(\boldsymbol{q}, h)}{2\sigma_c} \right) d\boldsymbol{q}\, dh\, d\phi} \tag{80}$$

which is the desired orientation-averaged Ziv–Zakai lower bound of eq. 15.

D THE CRLB AS A BALANCE BETWEEN LENGTH AND ITS DISTRIBUTION

Consider a scalar source $\theta \sim \mathrm{Unif}[0, 1]$ encoded by a differentiable curve $f : [0, 1] \to \mathbb{R}^N$. Its arc length in neural space is

$$L = \int_0^1 \left\| \frac{df(\theta)}{d\theta} \right\|_2 d\theta. \tag{81}$$

Writing $v(\theta) \triangleq \|df(\theta)/d\theta\|_2$ for the local rate of change of the code, we have $L = \int_0^1 v(\theta)\, d\theta$.

With independent Gaussian channel noise of variance σ_c^2 per neural coordinate, the local CRLB is $\sigma_c^2/v(\theta)^2$. Averaging over the uniform source gives

$$\overline{\mathrm{CRLB}} = \sigma_c^2 \int_0^1 \frac{1}{v(\theta)^2}\, d\theta. \tag{82}$$

Since $u \mapsto u^{-2}$ is strictly convex for $u > 0$, Jensen's inequality yields

$$\overline{\mathrm{CRLB}} \geq \frac{\sigma_c^2}{\left(\int_0^1 v(\theta)\, d\theta \right)^2} = \frac{\sigma_c^2}{L^2}. \tag{83}$$

This Jensen step is tight precisely when $v(\theta)$ is constant almost everywhere. In other words, the curve distributes its length (or Fisher information) equally across the source parameter. Therefore, coding performance improves by increasing the total curve length or by distributing that length more evenly.

E CONVEXITY OF Q-FUNCTION

Proposition 2. *Let*

$$\mathcal{Q}(z) \triangleq \frac{1}{\sqrt{2\pi}} \int_z^{\infty} \exp\left(-\frac{t^2}{2} \right) dt, \tag{84}$$

and define

$$\phi(x) \triangleq \mathcal{Q}(\sqrt{x}), \qquad x > 0. \tag{85}$$

Then ϕ is strictly convex on $(0, \infty)$.

Proof. Since

$$\mathcal{Q}'(z) = -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \tag{86}$$

the chain rule gives

$$\phi'(x) = -\frac{1}{2\sqrt{2\pi x}} \exp\left(-\frac{x}{2}\right). \tag{87}$$

Differentiating once more yields

$$\phi''(x) = \frac{1+x}{4\sqrt{2\pi}\, x^{3/2}} \exp\left(-\frac{x}{2}\right). \tag{88}$$

For every $x > 0$, all factors in this expression are strictly positive. Hence

$$\phi''(x) > 0, \qquad x > 0, \tag{89}$$

so ϕ is strictly convex on $(0, \infty)$. □

F Jensen Bound Over Orientation

We may also apply Jensen's inequality over the orientation ϕ. Recall that

$$\Phi(y) \triangleq \mathcal{Q}\left(\frac{\sqrt{y}}{2\sigma_c}\right) \tag{90}$$

is decreasing and strictly convex for $y > 0$. Define the orientation-averaged squared neural separation at displacement h by

$$\widetilde{d}^2(h) \triangleq \frac{1}{2\pi} \sum_{i=1}^{2} \int_0^{\pi} \overline{d^2_{i,\phi}}(h)\, d\phi. \tag{91}$$

Applying Jensen's inequality over the two orthogonal directions and their orientations gives

$$\frac{1}{2\pi} \sum_{i=1}^{2} \int_0^{\pi} \Phi\left(\overline{d^2_{i,\phi}}(h)\right) d\phi \geq \Phi\left(\widetilde{d}^2(h)\right). \tag{92}$$

Therefore, the Jensen-relaxed ZZLB satisfies

$$\mathrm{ZZLB}(f) \geq \frac{2}{\pi R^2} \int_0^{2R} h A_{\cap}(h) \mathcal{Q}\left(\frac{\sqrt{\widetilde{d}^2(h)}}{2\sigma_c}\right) dh. \tag{93}$$

Since Φ is strictly convex, equality in Eq. equation 92 holds if and only if

$$\overline{d^2_{i,\phi}}(h) = \widetilde{d}^2(h) \tag{94}$$

for all $i \in \{1, 2\}$ and $\phi \in [0, \pi)$.

Thus, for each displacement magnitude h, the bound is minimised when the available squared neural separation is distributed equally across orientations. In this sense, efficient coding favours an isotropic code: equally sized displacements should produce the same mean squared neural separation, independently of their direction.

G Connection to the Capacity Loss of Schaeffer et al. (2023)

Schaeffer et al. (2023) define their capacity loss as

$$\mathcal{L}_{\mathrm{Cap}} \triangleq -\left\| \frac{1}{BT} \sum_{\pi_b, t} \mathbf{g}_{\pi_b(t)} \right\|_2^2, \tag{95}$$

where B is the number of sampled path permutations, T is the number of time steps along each path, and $\mathbf{g}_{\pi_b(t)}$ is the neural representation at time t along path π_b.

Due to the path invariance of our model $\mathbf{g}_{\pi_b(t)} = f(\boldsymbol{q}_{\pi_b(t)})$. Therefore we may translate into our notation as follows:

$$\frac{1}{BT}\sum_{\pi_b,t}\mathbf{g}_{\pi_b(t)} = \frac{1}{BT}\sum_{\pi_b,t} f\big(\boldsymbol{q}_{\pi_b(t)}\big) \approx \mathbb{E}[f(\boldsymbol{q})] = \boldsymbol{\mu}_f, \tag{96}$$

where the approximation becomes exact in the asymptotic limit provided the sampled positions follow the uniform distribution as used in our analysis.

As their representations lie on a hypersphere, they satisfy a constant-energy constraint,

$$\|f(\boldsymbol{q})\|_2^2 = C. \tag{97}$$

Their capacity loss can therefore be written in our notation as

$$\mathcal{L}_{\mathrm{Cap}} = -\|\boldsymbol{\mu}_f\|_2^2 = P_{\mathrm{c}} - C. \tag{98}$$

Since C is fixed, minimising $\mathcal{L}_{\mathrm{Cap}}$ is equivalent to minimising P_{c}. Their chosen capacity loss therefore corresponds exactly to minimising the neural discrimination budget in our model. The intuition in their work is that a good code should use this resource efficiently i.e. by occupying only part of the available neural space, the code retains capacity to represent additional rooms for example. Thus, both approaches identify centred neural energy as the relevant coding resource.

Although the underlying intuition is the same, there is of course an important difference: their capacity loss favours minimising P_{c}, whereas we wish to maximise it. Understanding the consequences of this difference would be an interesting direction for future work, as they themselves also note.

H Connection to Conformal Isometry

The conformal-isometry loss is typically defined as follows, where h is small:

$$\mathcal{L}_{\mathrm{CI}}(h) \triangleq \mathbb{E}_{\boldsymbol{q},\phi}\Big[\big(\|f(\boldsymbol{q}+h\boldsymbol{u}_\phi) - f(\boldsymbol{q})\|_2 - sh\big)^2\Big], \tag{99}$$

where $s > 0$ is the constant scaling factor. Defining

$$D_f(\boldsymbol{q},h,\phi) \triangleq \|f(\boldsymbol{q}+h\boldsymbol{u}_\phi) - f(\boldsymbol{q})\|_2, \tag{100}$$

If we now use the bias-variance identity:

$$\mathcal{L}_{\mathrm{CI}}(h) = \big(\mathbb{E}_{\boldsymbol{q},\phi}[D_f] - sh\big)^2 + \mathrm{Var}_{\boldsymbol{q},\phi}(D_f). \tag{101}$$

Thus, for fixed mean neural separation, the loss is minimised when $D_f(\boldsymbol{q},h,\phi)$ is constant over both $\boldsymbol{q}$ and ϕ. This argument is similar to the argument of theorem 6 in Xu et al. (2025), where they motivate isotropy from conformal isometry, but by also considering the different possible starting points $\boldsymbol{q}$ we also highlight how conformal isometry also favours translational covariance.

I Translation-invariant Neural Encodings

We seek continuous encoders $f : \mathbb{R}^2 \to \mathbb{R}^N$ whose inner products depend only on displacement:

$$f(\boldsymbol{q})^\top f(\boldsymbol{q}') = K(\boldsymbol{q}' - \boldsymbol{q}). \tag{102}$$

The logic of this proof is very similar to the theory underpinning Rahimi & Recht (2007). Bochner's theorem Rudin et al. (1962) states that a continuous function $K : \mathbb{R}^2 \to \mathbb{C}$ is positive definite if and only if it is the Fourier transform of a finite non-negative measure μ:

$$K(\boldsymbol{h}) = \int_{\mathbb{R}^2} e^{\,\mathrm{i}\boldsymbol{k}^\top \boldsymbol{h}}\, d\mu(\boldsymbol{k}). \tag{103}$$

Thus as our kernel (the inner product) is positive definite then it must take this form. Since f has only N coordinates, the kernel has rank at most N. Consequently, its spectral measure has finite support, and the integral in Eq. 103 reduces to a finite sum.

Moreover, because K is real-valued the imaginary parts therefore must cancel i.e. the opposite wavevectors $\boldsymbol{k}$ and $-\boldsymbol{k}$ have equal weights:

$$we^{\mathrm{i}\boldsymbol{k}^\top \boldsymbol{h}} + we^{-\mathrm{i}\boldsymbol{k}^\top \boldsymbol{h}} = 2w\cos(\boldsymbol{k}^\top \boldsymbol{h}). \tag{104}$$

Writing $a_m^2 = 2w_m$ and denoting the weight at zero frequency by c_0, we obtain

$$K(\boldsymbol{h}) = c_0 + \sum_{m=1}^{M} a_m^2 \cos(\boldsymbol{k}_m^\top \boldsymbol{h}) \tag{105}$$

Thus using the following trig identity:

$$\cos\big(\boldsymbol{k}_m^\top(\boldsymbol{q}' - \boldsymbol{q})\big) = \cos(\boldsymbol{k}_m^\top \boldsymbol{q})\cos(\boldsymbol{k}_m^\top \boldsymbol{q}') + \sin(\boldsymbol{k}_m^\top \boldsymbol{q})\sin(\boldsymbol{k}_m^\top \boldsymbol{q}'). \tag{106}$$

We can see that an encoder with the general form below, would achieve the desired inner product:

$$\boxed{f(\boldsymbol{q}) = \boldsymbol{c} + \sum_{m=1}^{M} a_m \left[\boldsymbol{u}_m \cos(\boldsymbol{k}_m^\top \boldsymbol{q}) + \boldsymbol{v}_m \sin(\boldsymbol{k}_m^\top \boldsymbol{q})\right],} \tag{107}$$

where:

$$\boldsymbol{u}_m^\top \boldsymbol{u}_n = \boldsymbol{v}_m^\top \boldsymbol{v}_n = \delta_{mn}, \qquad \boldsymbol{u}_m^\top \boldsymbol{v}_n = 0, \qquad \boldsymbol{c}^\top \boldsymbol{u}_m = \boldsymbol{c}^\top \boldsymbol{v}_m = 0, \qquad \|\boldsymbol{c}\|_2^2 = c_0. \tag{108}$$

Where the above orthogonality is needed such that the cross terms can be removed. Any other encoder with the same inner products differs only by an orthogonal change of neural coordinates, so this form is general.

Each sin cosine pair requires two neural dimensions. Hence $M \leq \lfloor (N-1)/2 \rfloor$.

It is then easy to see that the neural squared distance can be written as:

$$\|f(\boldsymbol{q}+\boldsymbol{h}) - f(\boldsymbol{q})\|_2^2 = 2\sum_{m=1}^{M} a_m^2 \left[1 - \cos(\boldsymbol{k}_m^\top \boldsymbol{h})\right]. \tag{109}$$

Thus independent of $\boldsymbol{q}$ as desired.

J Numerical Optimisation

We jointly optimise the Fourier wavevectors, power allocation, and neural mixing matrix of the parametric code as in Eq. 25. For $M = (N-1)/2$ this can be written as:

$$\boldsymbol{z}(\boldsymbol{q}) = \begin{bmatrix} \sqrt{p_0} \\ \sqrt{p_1}\cos(\boldsymbol{k}_1^\top \boldsymbol{q}) \\ \sqrt{p_1}\sin(\boldsymbol{k}_1^\top \boldsymbol{q}) \\ \vdots \\ \sqrt{p_M}\cos(\boldsymbol{k}_M^\top \boldsymbol{q}) \\ \sqrt{p_M}\sin(\boldsymbol{k}_M^\top \boldsymbol{q}) \end{bmatrix}, \qquad f(\boldsymbol{q}) = Q\boldsymbol{z}(\boldsymbol{q}). \tag{110}$$

Wavevector initialisation. Each wavevector is initialised by independently sampling its direction uniformly on the circle and its magnitude uniformly from a prescribed interval $[k_{\min}, k_{\max}]$. The wavevectors are subsequently optimised without any constraints on their magnitudes or directions.

Power constraint and non-negative initialisation. Throughout optimisation, the mode powers are obtained from trainable parameters $\{\theta_m\}_{m=0}^{M}$ using

$$p_m = C\frac{\exp(\theta_m)}{\sum_{j=0}^{M}\exp(\theta_j)}. \tag{111}$$

The softmax therefore ensures that the powers remain non-negative and satisfy the total power constraint

$$\sum_{m=0}^{M} p_m = C \tag{112}$$

at every optimisation step. Adam updates the unconstrained parameters θ_m, and the corresponding powers are recomputed whenever the encoder is evaluated. We initialise the constant component with a sufficiently large fraction of the total power, such that we start non-negative.

Orthogonal Representations Due to the requirements of our vectors in Eq. 25, the neural mixing matrix, Q must be orthogonal, hence at each optimisation step, we apply a QR decomposition to the current mixing matrix. Then using the resulting orthogonal matrix we calculate the neural responses and loss. Adam then takes a gradient step, after which the stored mixing matrix may no longer be orthogonal. We therefore apply the QR decomposition again before the next evaluation. Thus the the Q used by the encoder is always orthogonal.

Non-negativity constraint. To enforce non-negativity we use the followig loss:

$$\mathcal{L}_{\text{pos}} = \mathbb{E}_{\boldsymbol{q},i}\left[\max\left\{0, -f_i(\boldsymbol{q})\right\}\right].$$

Where the points $\boldsymbol{q}$ are sampled according to the uniform distribution on the disc, as stated in our optimisation problem, Eq. 15. These positions are periodically resampled during optimisation. Thus the complete loss is:

$$\mathcal{L} = s_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{pos}}\left(\mathcal{L}_{\text{pos}} - \epsilon_{\text{pos}}\right),$$

where ϵ_{pos} is the target positivity violation, λ_{pos} controls the strength of the constraint and $\mathcal{L}_{\text{sep}}$ is our ZZLB as in Eq. 15.

GECO-style constraint control. We do not want the optimiser to trade a large positivity violation for a lower separation loss. We therefore adapt λ_{pos} using a GECO-style procedure Rezende & Viola (2018). The signed relative constraint error is

$$e_t = \frac{\mathcal{L}_{\text{pos},t}}{\epsilon_{\text{pos}}} - 1.$$

After clipping extreme values, this error is smoothed using

$$\bar{e}_t = \alpha_{\text{pos}}\bar{e}_{t-1} + (1 - \alpha_{\text{pos}})\, e_t.$$

The positivity multiplier is then updated according to

$$\lambda_{\text{pos},t+1} = \lambda_{\text{pos},t}\exp\left(\gamma_{\text{pos}}\bar{e}_t\right).$$

Thus, the positivity penalty increases when the constraint is violated and decreases when the positivity loss is below its target. The multiplier is restricted to a fixed interval for numerical stability.

Optimisation and model selection. The wavevectors, power parameters, and unconstrained mixing matrix are jointly optimised using Adam Kingma & Ba (2014). Training positions and position pairs are periodically resampled, whereas performance is measured using a fixed validation set. To pick the hyperparameters used in the run that produced Figure 4, we performed a Cartesian hyperparameter sweep, over the Adam moment coeffs β_1 and β_2, the shared learning rate applied to the wavevectors, powers, and mixing matrix, the maximum initial wavevector magnitude $k_{\max}$, the separation-loss scale s_{sep}, the initial positivity multiplier λ_{pos}, and the GECO constraint target, and multiplier update rate.

K MSE AND ZZLB CORRELATION

Figure 5 both demonstrates the tightness of the ZZLB to the MSE, but more crucially a strong correlation between the two. Thus suggesting that encoders which perform well on ZZLB also perform correspondingly well for the MSE, justifying our choice of using the ZZLB.

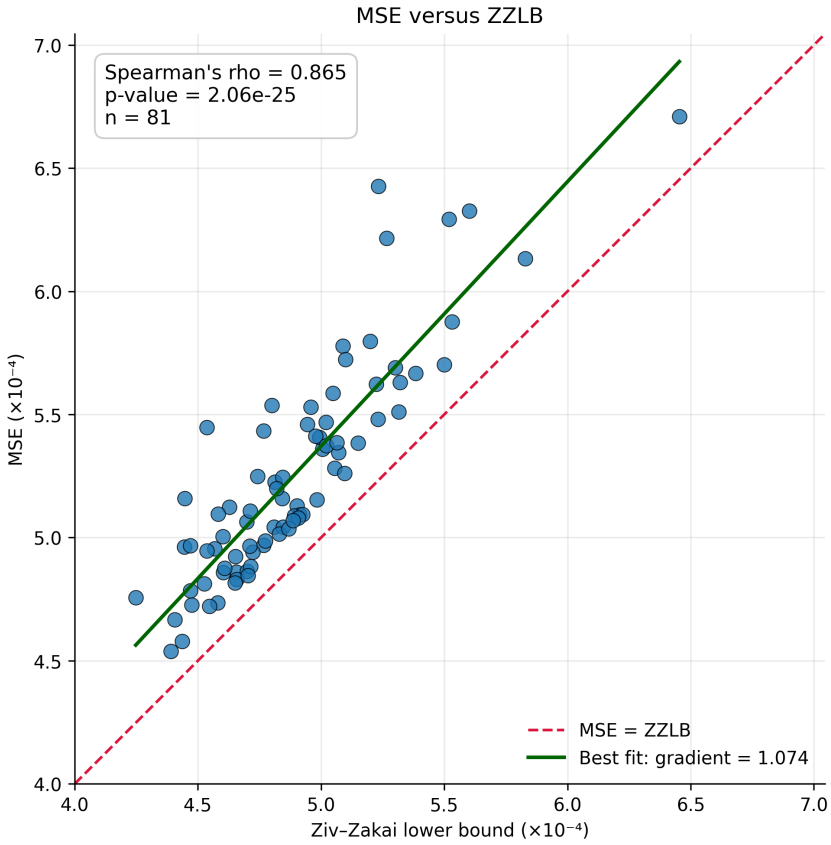


Figure 5: Scatter plot of Monte Carlo mean squared error (MSE) versus the ZZLB for 81 optimised encoders. Each point represents one encoder. The dashed red line indicates equality between MSE and ZZLB, and the solid green line shows the least-squares linear fit. Spearman's rank correlation coefficient and its two-sided p-value are shown in the plot. Both axes are expressed in units of 10^{-4}.

L RATE MAPS

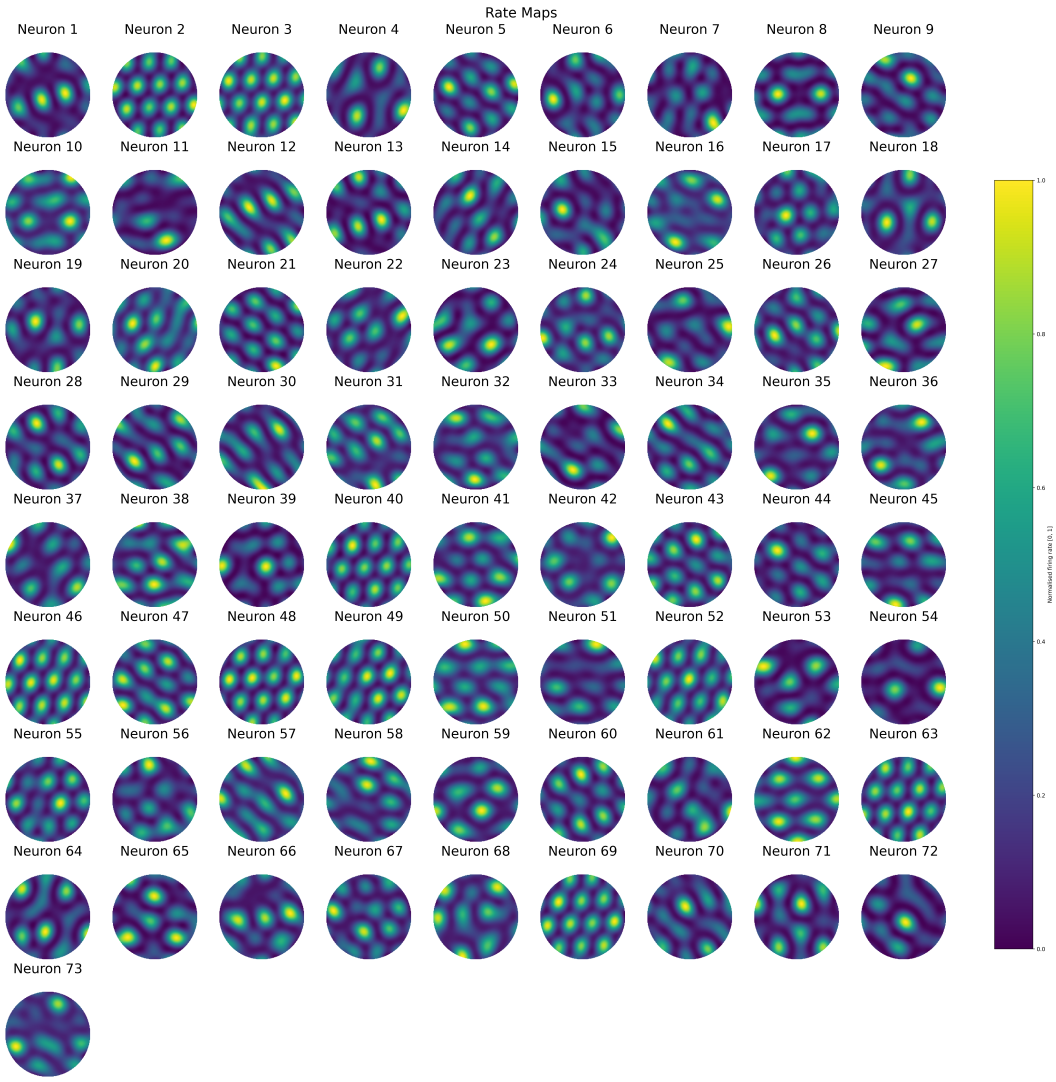


Figure 6: All rate maps for the 73 neurons corresponding to the simulation shown in the main text Figure 4

M SIMULATION PARAMETERS

The following settings were used for Figure 4. The optimisation procedure is described in Appendix J. Resampling and evaluation intervals are measured in optimisation steps. Power-parameter noise denotes the standard deviation of the Gaussian perturbations applied before the softmax at initialisation.

Parameter	Symbol	Value
Encoder		
Neurons	N	73
Domain radius	–	1
Total power	C	1
Noise standard deviation	σ_c	0.08
Initialisation		

Parameter	Symbol	Value
Wavevector magnitudes	$[k_{\min}, k_{\max}]$	$[1, 10]$
Power-parameter noise	–	0.2
Positivity safety margin	–	0
Optimisation		
Initial learning rate	η_0	8×10^{-4}
First-moment decay	β_1	0.9
Second-moment decay	β_2	0.9
Adam stabiliser	ϵ_{Adam}	10^{-8}
Gradient norm limit	–	3
Separation weight	s_{sep}	10 000
Positivity controller		
Initial multiplier	$\lambda_{\mathrm{pos},0}$	2
Constraint tolerance	ϵ_{pos}	e^{-9}
Error smoothing	α_{pos}	0.9
Multiplier update rate	γ_{pos}	5×10^{-3}
Relative-error limits	–	$[-1, 20]$
Multiplier limits	–	$[10^{-8}, 10^{8}]$
Training		
Position pairs	–	30 000
Positivity samples	–	30 000
Resampling interval	–	5
Validation		
Position pairs	–	30 000
Positivity samples	–	30 000
Evaluation interval	–	400