Spontaneous emergence of topologically robust grid cell modules: A multiscale instability theory

Mikail Khona[1,2†], Sarthak Chandra[1†], Ila Fiete[1*]

[1]Department of Brain and Cognitive Sciences and McGovern Institute,
[2]Department of Physics,
Massachusetts Institute of Technology, Cambridge, USA
[*]To whom correspondence should be addressed; E-mail: fiete@mit.edu.
[†] Denotes equal contribution

Modular structures in the brain play a central role in compositionality and intelligence, however the general mechanisms driving module emergence have remained elusive. Studying entorhinal grid cells as paradigmatic examples of modular architecture and function, we demonstrate the spontaneous emergence of a small number of discrete spatial and functional modules from an interplay between continuously varying lateral interactions generated by smooth cortical gradients. We derive a comprehensive analytic theory of modularization, revealing that the process is highly generic with its robustness deriving from topological origins. The theory generates universal predictions for the sequence of grid period ratios, furnishing the most accurate explanation of grid cell data to date. Altogether, this work reveals novel principles by which simple bottom-up dynamical interactions lead to macroscopic modular organization.

One sentence summary A novel bottom-up pathway for the self-organization of modules in biology provides quantitative match to grid cell experiments.

Introduction Modular structures are robust to localized perturbations (*1, 2*), faster to adapt if the world requires sparse or modular changes (*3*), relatively expressive because of compositionality (*4–6*), and probably for these reasons empirically favored by the arrow of evolution, which drives systems toward greater modularization (*7*). In this sense, modularity is the crux of biological organization. However, the mechanisms driving modularity are not well understood. Most models involve some form of supervision (selection in systems biology and feedback-based learning in neuroscience and machine learning) and are very slow processes that require intensive data from the world (*8–10*). These top-down models of modularization typically involve evolution in modularly changing environments (*8, 11, 12*) or learning of multiple modular tasks (*9, 10, 13, 14*)[1]. Here we propose an alternative mechanism, in the form of self-organization that emerges from spontaneous symmetry breaking dynamics: Just as simple bottom-up physical constraints serve as important shapers of morphogenesis and local pattern formation (*17–20*), they can also provide critical inductive predispositions that drive global modularization, as a complement to top-down forces.

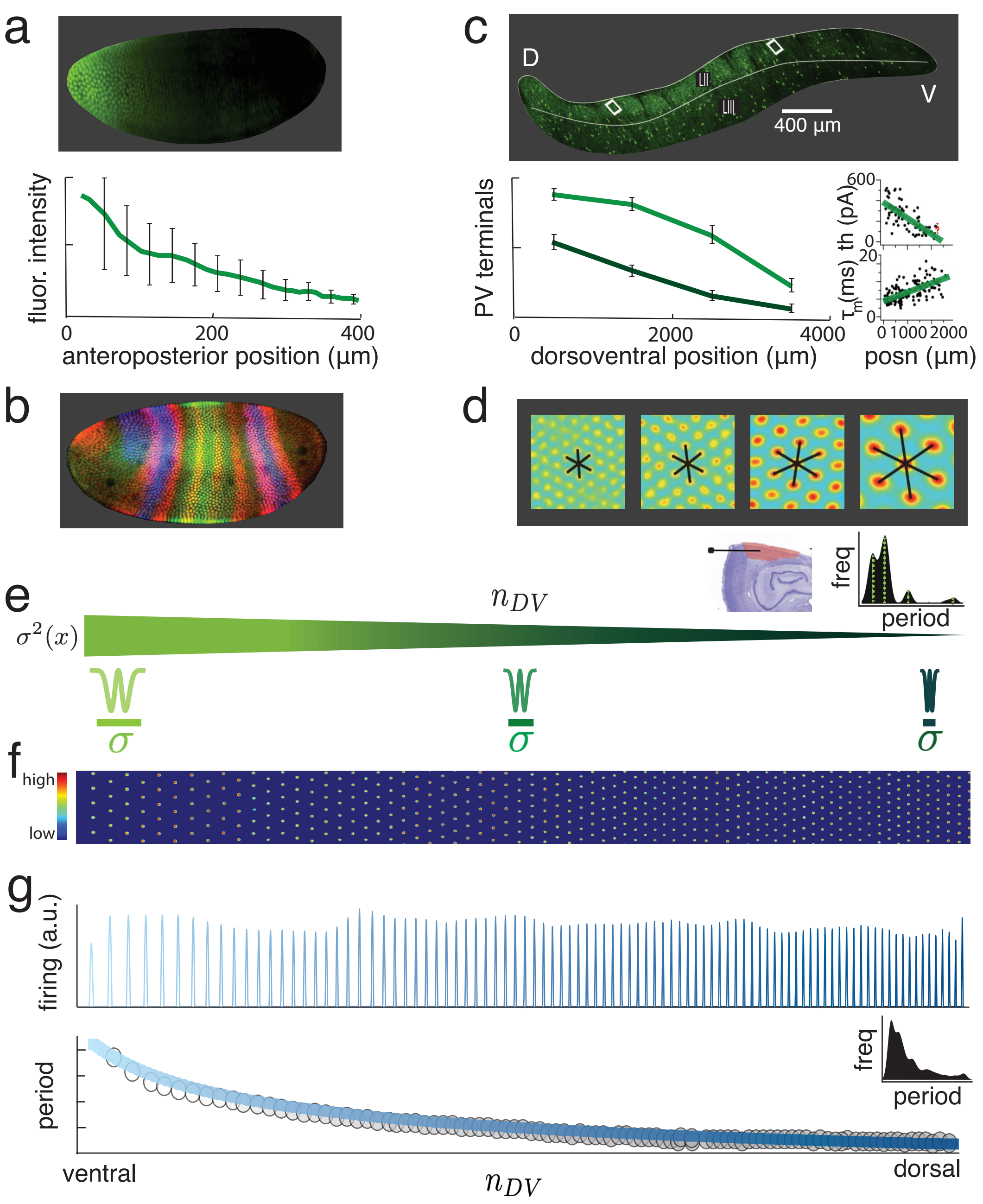
Here, we aim to provide a robust and general theory of modular structure emergence from smooth gradients via bottom-up principles, with grid cells as our concrete subject. Extensive supervised training of networks to reproduce a prespecified pattern of place cell activity with velocity inputs can result in grid-like activity patterns, with multiple spatial scales (*21–23*). The grid periods and emergence of grid patterns are, however, highly sensitive to ad-hoc modeling decisions about network size, shape, and target place cell activity (*21–23*). Two recent

[1] Some models exhibit modularity emergence through motif duplication followed by specialization (*15, 16*), though the resulting modules typically require selection or supervision under the conditions noted earlier to remain stable.

works empirically demonstrate module emergence without extensive supervision: in the first, grid tuning is derived from location-coding place cell inputs and a continuum of adaptation timescales (*24*). However, the resulting grid cells do not possess the ability to integrate velocity inputs and cell-cell relationships are not preserved across environments, inconsistent with the observed stability of pairwise grid cell relationships (*25–27*). In the second, grid module formation proceeds from a multi-sheet architecture, however it already assumes a proto-modular architecture with structured and finely tuned one-to-one linkages of neurons across the modules (*28*). In both these works, we also lack any theory of the dynamics of module emergence.

We present a novel pathway for the spontaneous emergence of modularity from smooth gradients that is robust through a topological mechanism, and does not require supervision. The mechanism admits a complete analytical description of why modules emerge, which reveals the genericity of the process and suggests that the insights should generalize from grid cells to other systems in biology and machine learning. Finally, we generate detailed predictions about the entire sequence of period ratios of grid modules that provide the best fit to data till date.

Smooth gradients as precursors to discrete modules in biology In biological systems, modular structures often originate from spatial biochemical gradients (*36–38*) that unleash a spatial patterning process (*39–42*), resulting in the emergence of complex structures. For instance, the Bicoid protein gradient in Drosophila embryos (*30*), Fig. 1a, generated by the maternal deposition of Bicoid RNA, precedes and causally guides the formation of body segments in the fly through the dynamics of a spatially modulated gene expression cascade, Fig. 1b. Mature modular systems also frequently exhibit smooth gradients in their underlying biophysical or functional properties (*32, 43–45*). Grid cells in the medial entorhinal cortex (MEC; Fig. 1c, top) form modules in which all co-modular cells share the same spatial period, with discrete jumps in period between modules (Fig 1d). The modules are arranged in order of increasing period


a
fluor. intensity
0
200
400
anteroposterior position (µm)
b
c
D
V
LII
LIII
400 µm
PV terminals
0
2000
4000
dorsoventral position (µm)
600
0
th (pA)
20
0
τ_m(ms)
0 1000 2000
posn (µm)
d
freq
period
e
n_{DV}
$\sigma^2(x)$
σ
σ
σ
f
high
low
g
firing (a.u.)
period
freq
period
ventral
n_{DV}
dorsal

along dorsoventral MEC in the brain, accompanied by a large number of cellular and network properties including cell resistance, activation threshold, time-constant, and local inhibition that vary continuously (Fig. 1c). Can these smooth underlying gradients lead to the emergence of discrete modules?

A fixed connectivity range with smooth gradients leads to modularization We build a mechanistic circuit model of grid cells based on the principle of continuous attractors (*35*). Neurons are situated along a strip, with a slow spatial gradient along the long axis in the width of the lateral interactions (W^g, with scale $\sigma(n_{DV})$). The result is the formation of hexagonal activity patterns whose period varies smoothly with the interaction gradient, and no modularization (Fig. 1f-g). We next consider the addition of a second scale in the lateral interaction (W^f, with scale d that is fixed but wider than $\sigma(n_{DV})$ along the whole neural strip), Fig. 2a-

Figure 1 *(preceding page)*: **Smooth gradients as precursors to discrete modules in biology** (a) Fluorescence imaging of the expression of the protein *bcd* early in development of the Drosophila embryo (*29*); this polarity gradient is the precursor to a gene-protein expression cascade that leads to the formation of body segmentation. (Figure adapted from (*30*).) (b) Discrete gap and pair-rule gene expression bands emerging from the initial polarity protein gradients (immunofluorescence image adapted from (*31*)). (c) Smooth gradient in inhibitory (PV-immunoreactive) axon terminals along the dorsoventral (DV) axis in layers II and III of the medial entorhinal cortex (MEC) (*32*), where grid cells are found (*33*). (d) The spatial tuning curves of grid cells (top: autocorrelation of spatial responses of four cells; intensity plot as a function of two spatial dimensions) vary in spatial period along the DV axis in MEC (region and electrode penetration angle shown in left inset, bottom), with period increasing ventrally. However, the variation is discrete: there are a few discrete modules with a few periods, and nearby cells have the same period (right inset, bottom) (adapted from (*34*)). (e) In a continuous attractor model of grid cells formed by pattern formation through lateral inhibition (*35*), we model the biophysical DV gradients in MEC by a continuous gradient in the width of lateral inhibition. (f-g) The smooth gradient of (e) results in pattern formation with smoothly varying periodicity in the population pattern, with no modularity (in (f): neural population activity pattern in a 2-dimensional neural strip; (g): top, activity pattern in a 1-dimensional neural strip; bottom, extracted period as a function of DV location; inset, histogram of extracted periods).

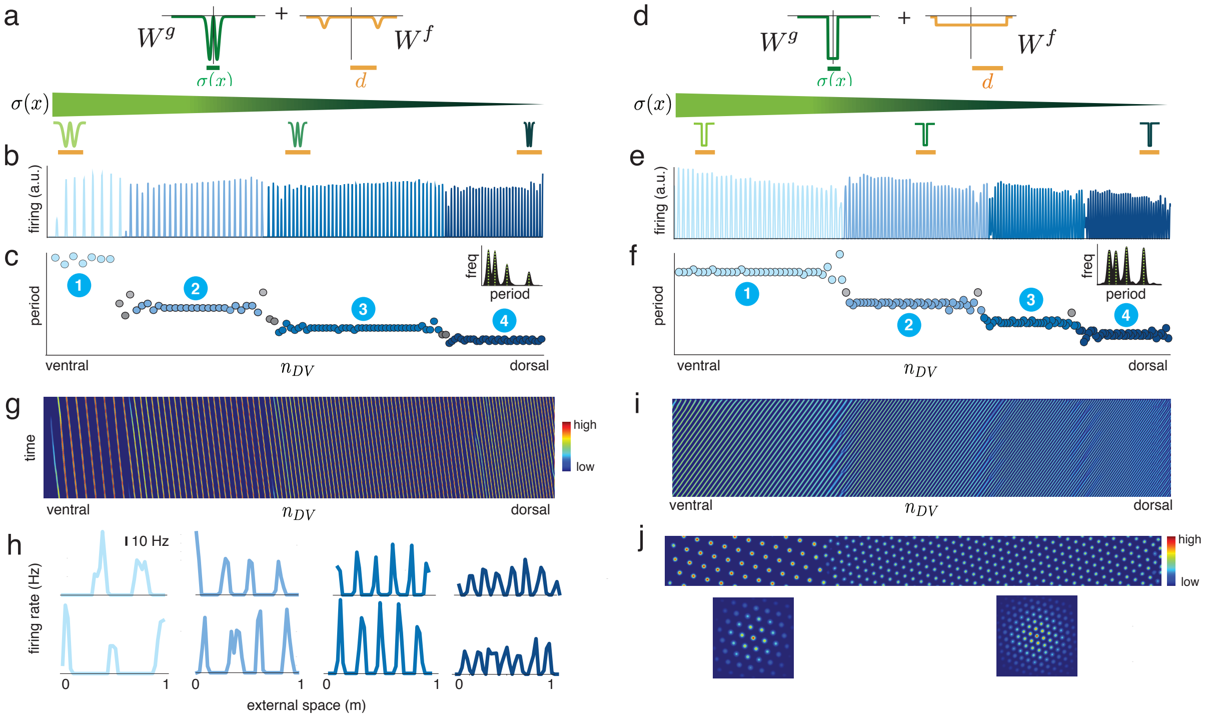


Figure 2: **A fixed connectivity range combined with smooth gradients leads to module emergence.** (a) Lateral interactions consisting of a smoothly varying component (W^g, green) and a component with fixed width (W^f, orange) along the neural strip. Widths defined as $\sigma(n_{DV}), d$, respectively. (b-c) The interactions from (a) lead to self-organized module formation (activity pattern in 1-dimensional neural strip with smaller periods more dorsally, (b); extracted periods, (c)). (d-f) Same as (a-c), but for a very different pair of W^g, W^f. (g-h) Modularity is not only present in the population pattern on the neural strip, but also in function: response of the network over time to a velocity input injected into all cells. (g) Flow of population activity in response to a constant velocity input; (h) Spatial tuning curves of 2 example cells per module (top and bottom), generated by collecting spikes in response to a sinusoidal velocity input corresponding to multiple left-right traversals of a 1-dimensional environment, and plotting the histogram of spikes as a function of position in the environment. (i) Same as (g), for the interactions in (d). (j) 2-dimensional pattern formation and module emergence for a 2-dimensional neural strip with interactions as in (a). (Bottom) 2d autocorrelation function of the local (single-module) patterns in the neural strip.

c. This fixed interaction scale could represent a general constraint on the extent of neural arborization, independent of location. With the addition of a fixed-scale interaction, the network spontaneously and robustly exhibits modular pattern formation, Fig. 2b, with spatially periodic activity patterns and a discrete number of periods with discontinuous jumps in between. Note that the formed modules are generally much larger than both interaction scales $d, \sigma(n_{DV})$. Remarkably, the results are similar across varied shapes of the smoothly varying and fixed-range interactions (cf. Fig. 2a-g with 2d-i), and regardless of whether the fixed-range interaction is spatially diffuse over d (Fig. 2b), or localized at a distance d (Fig. 2a).

When the network is driven by velocity inputs, the activity bumps move across the cortical sheet (reflecting the velocity-integration function of grid cells) and notably, despite this bump movement the modules and sharp modular boundaries persist at the same cortical locations (Fig. 2g,i). This produces spatially periodic tuning in individual cells (Fig. 2h), with the pattern phase changing at different rates in each module so that co-modular cells have the same spatial tuning curves with various phase offsets, but cells in different modules have discretely different spatial periods (Fig. 2h). Thus, the network is also fully modular in its velocity integration functionality. All the results on discrete module formation and modular function hold for 2-dimensional neural sheets as well, Fig. 2j (See SI Sec. 3.8 for additional analytical and numerical results.)

Theory of module formation: multiscale instability and topological peak selection If we initialize the network at a uniform activity state, the modularization process begins at the earliest time-steps, appearing before most neurons have crossed their nonlinear thresholds, and coincident with the process of local patterning (Fig. 3a,b). This suggests that it should be possible to explain both local pattern formation and larger-scale module emergence within a unified linear instability framework.

A multiscale analysis in Fourier space makes conceptually clear how and why this happens: If the graded interaction width varies sufficiently slowly along the neural strip, the system is locally translation-invariant. At any location n_{DV} in the neural strip, we can thus con-

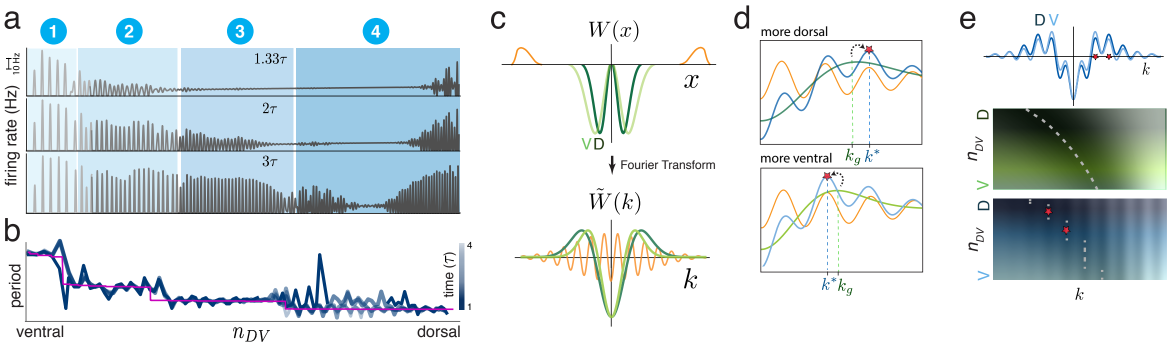


Figure 3: **Theory of module formation: multi-scale linear instability and peak selection.** (a) Snapshots of population activity in the neural strip within a few neural time-constants (τ) of running the dynamics: modules appear at the same time as the local pattern, and both appear before most neurons have hit their nonlinear thresholds. (b) Top: Schematic of the fixed-range (orange) and graded (green) interactions from two different DV locations on the neural strip (top), and bottom: their Fourier transforms. Fourier peaks of the fixed-range interaction remain unchanged dorsoventrally but the graded interaction peak slides smoothly inward when moving ventrally. (c) Peak selection process: the global maximum in Fourier space is based on the combination (blue) of the shallow graded-interaction peak (green)and the multiple narrow fixed-interaction peaks (orange). The smoothly shifting graded-interaction peak "selects" which of the local maxima from the fixed interaction is the global maximum. Thus, the global maximum jumps abruptly from one local maximum to the next. (d) The Fourier transform of the combined interactions, for a dorsal and a ventral location (top), and the Fourier-space location of the global maximum of just the graded interaction as a function of DV location (middle) and the combined interaction (bottom).

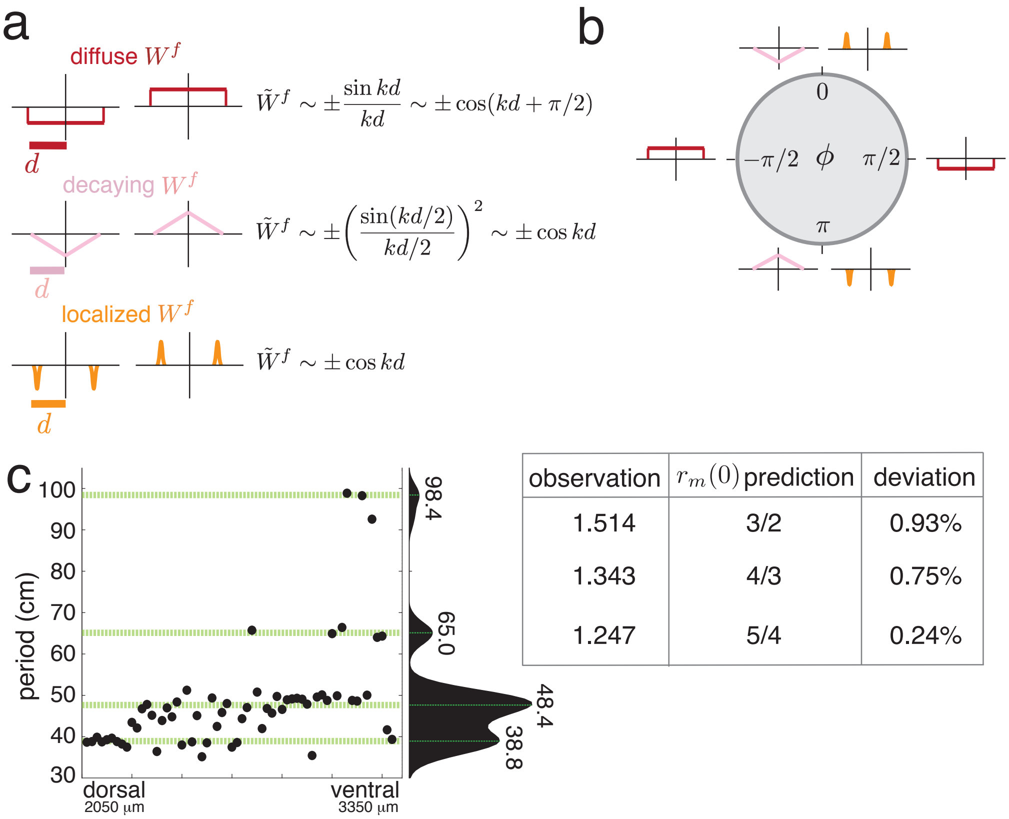
sider how local perturbations behave by decomposing them into Fourier-space activity modes (defined by $k = \frac{2\pi}{\lambda}$, see SI Fig. 6) that decay or grow exponentially, dominated by the fastest-growing mode corresponding to the peak of the combined lateral interaction ($k^*(n_{DV}) = \arg\max_k \tilde{W}(k; n_{DV}) = \arg\max_k \{\tilde{W}^f(k) + W^g(k; n_{DV})\}$) (*35, 39–42, 46*). When $\tilde{W}(k^*)$ is positive, the result is a locally patterned state with period $\lambda(n_{DV}) = 2\pi/k^*(n_{DV})$. The component W^f sets up a set of local maxima in Fourier space, with spacings $\sim 1/d$. Because this interaction does not vary along the neural strip, the set of local maxima remains the same (Fig. 3b, orange). The graded component $W^g(n_{DV})$ sets up a much broader Fourier peak (of scale $\sim 1/\sigma(n_{DV}) \gg 1/d$), which smoothly contracts more ventrally (Fig. 3b, green). Though the graded component, with its smoothly moving maximum, is ultimately responsible for any changes in spatial period along the neural sheet, the peak is too shallow to fully determine the maximum. Rather, the narrow peaks from $\tilde{W}^f(k)$ determine the set of potential locations of the global maximum, while the smoothly moving peak from $\tilde{W}^f(k)$ performs "peak selection" (cf. SI Movie 1) to determine which of the narrow peaks will form the global maximum at a given dorsoventral location (Fig. 3c). Thus, the only permitted peaks are those whose positions in Fourier space fall between $\sim 1/\max(\sigma)$ and $\sim 1/\min(\sigma)$, and moving smoothly along the neural strip causes flat steps with discrete jumps in the local periodicity of the formed activity pattern through discrete peak selection.

The analytical result further predicts the number of modules, and reveals that this number is independent of the length of the neural sheet, of the shape of the lateral interactions, and of the shape of the monotonic function that describes how $\sigma(n_{DV})$ varies across the neural sheet. Instead, it is determined only by the extremal widths $\min(\sigma), \max(\sigma)$ of the graded interaction, because it is given by how many integer multiples of $1/d$ fit into the interval $[1/\min(\sigma), 1/\max(\sigma)]$ (SI Sec. 3.7). If the extremal widths are fixed, the modularization process generates the same number of modules regardless of neural sheet size and the details of the boundary conditions. Moreover, small changes in the extremal values also have no effect on the number of modules, until a critical point at which another module would be added; thus, the number of modules is a topological invariant (*47*). As a corollary, the average module size grows linearly with network size and has little relationship to the lateral interaction widths, explaining why the formed modules in Fig. 2b,e are much larger than the length-scales of the the lateral interactions. In sum, the module formation mechanism is at its essence a topological process with each module arising as a topologically protected phase (*47*) which is independent of most details and parametric variations (SI Sec.3.7). Thus, a combination of interactions with a smoothly varying scale and a fixed larger scale leads to module formation through a robust peak selection process in Fourier space. This robustness and the genericity of the peak-selection mechanism for module emergence explains the striking similarity of results across the varied lateral interaction profiles of Fig. 2 and the additional profiles in SI (SI Fig. 6).

Finally, the peak selection process can be reformulated as minimization on an energy landscape (SI Sec. 4); thus, the principle applies beyond Fourier space and it can be used to generate modular solutions involving not just periodic patterns but a range of problems that admit multiple potential solutions (local optima), SI Sec. 5.

Detailed prediction of pairwise module period ratios The theory further generates a quantitative prediction about how different module periods scale relative to each other, another property that is invariant to the detailed shapes of lateral interactions and the shape of the gradient for the lateral interaction width. Unlike existing experimental fits in which all period ratios $r_m = \lambda_{m+1}/\lambda_m$ are described by a single parameter $\alpha \sim 1.4$, the prediction of the present model gives further detail on the sequence of successive period ratios as a function of m.

Suppose as before that W^g generates locally periodic activity patterns with a smoothly varying scale $\sigma(n_{DV})$ and that the fixed-range interaction W^f has a dominant scale $d \gg \sigma(n_{DV})$. Let us additionally assume that all other scales in W^f are $\ll \sigma(n_{DV})$. Then, the Fourier transform $\tilde{W}^f$ is typically (except for a set of measure zero; SI Sec. 3.5.1 for details) an approximately periodic oscillatory function ($\sim \cos(kd + \phi)$), with relatively evenly spaced peaks occurring every m/d for integer m (Fig. 4a). Thus, the positions of successive peaks (and thus the module periods) are specified simply by successive integers together with a single phase ϕ that depends on the shape of the fixed-range interaction, Fig 4a-b. Since module periods correspond to successive peaks of this function, it follows that the period ratio is given by: $r_m = r_m(\phi) = (m + \phi/2\pi)/(m - 1 + \phi/2\pi)$ (details in SI Sec. 3.6). The predicted average period ratio from this formula for the first three period ratios is 1.37 (averaging also over all ϕ),



observation	$r_m(0)$ prediction	deviation
1.514	3/2	0.93%
1.343	4/3	0.75%
1.247	5/4	0.24%

Figure 4: **Robust prediction of pairwise module period ratios for simple fixed-scale interaction profiles.** (a) Examples of diffuse, decaying, and localized fixed-scale lateral interactions W^f, together with the dominant terms in their Fourier transforms $\tilde{W}^f$. (b) The phases of the Fourier transforms for the fixed-scale interactions in (a). (c) Observed periods of grid cells from multiple modules (*34*) (left), successive period ratios computed from the observation (table, left column), and predicted period ratios for $\phi = 0$ (table, middle column). Ratios match predicted values with $R^2 = 0.999$ (table, right column).

in good agreement with known results (*34*), but further, successive period ratios are predicted to have distinct values with a simple relationship between them. Given the specific form of our prediction for successive period ratios, and that is nevertheless generic across different lateral interaction profiles, we compare it with experimental results, Fig. 4c. The prediction with $\phi = 0$ matches the sequence of observed period ratios strikingly well for this dataset and others (comparison with additional data presented in SI Sec. 3.9).

Discussion We have shown that the combination of fixed-range lateral interactions with smooth gradients in the lateral interaction width can lead to spatially discretely varying solutions in a pattern forming dynamical system, through discrete selection of global maxima from a set of local maxima. This is a novel recipe for bottom-up modularization in neural circuits, that could be incorporated into neural models beyond grid cells. Our analytical formulation thus simplifies and extends existing approaches that have also examined the role of smooth gradients in patterning and functional specialization (*37, 43, 48*).

We have shown that the number of formed modules is independent of the size of the neural strip and the shape of the gradient in the lateral interaction, if the extremal values of the interaction widths are held fixed. The addition of complementary mechanisms to maintain the overall gradient magnitude across the neural sheet regardless to size will thus lead to invariance with the size of the system (*49, 50*). For grid cells, our model and its sound match with experimental data suggests that MEC may contain a fixed-width interaction that co-exists with known smooth gradients in a broad range of cellular and network-level properties. The fact that $\phi = 0$ provides the best description of the data also helps to constrain possible forms of the fixed-width lateral interaction. These predictions are testable with connectomic reconstructions of the network (*51–53*).

The same principles may be relevant for biology outside the brain: just as pattern formation is a critical process both for morphogenesis and the formation of functional neural circuits (*35, 39, 54*), multi-scale interaction-driven pattern formation may be relevant not just for modularization in the brain but also for the morphogenesis of discrete structures, for example in making the process of body segmentation more robust (*7*).

References

1. M. Sales-Parda, *Science* **6347**, 128 (2017).

2. A. Limdi, A. Pérez-Escudero, A. Li, J. Gore, *Nat Commun* **9**, 2969 (2018).

3. J.-M. Park, M. Chen, D. Wang, M. W. Deem, *Phys Biol* **12**, 025001 (2015).

4. W. v. Humboldt (Cambridge University Press, 2005).

5. J. A. Fodor, Z. W. Pylyshyn, *Cognition* **28**, 3 (1988).

6. S. Sreenivasan, I. Fiete, *Nat Neurosci* **14**, 1330 (2011).

7. G. P. Wagner, M. Pavlicev, J. M. Cheverud, *Nature Reviews Genetics* **9**, 921 (2007).

8. N. Kashtan, U. Alon, *Proc Natl Acad Sci U S A* **102**, 13773 (2005).

9. F. Alet, T. Lozano-Perez, L. P. Kaelbling, *Proceedings of The 2nd Conference on Robot Learning*, A. Billard, A. Dragan, J. Peters, J. Morimoto, eds. (PMLR, 2018), vol. 87 of *Proceedings of Machine Learning Research*, pp. 856–868.

10. G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, X.-J. Wang, *Nat Neurosci* **22**, 297 (2019).

11. J. Sun, M. W. Deem, *Physical Review Letters* **99** (2007).

12. D. M. Lorenz, A. Jeng, M. W. Deem, *Phys Life Rev* **8**, 129 (2011).

13. G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, B. Schölkopf, *Proceedings of the 35th International Conference on Machine Learning*, J. Dy, A. Krause, eds. (PMLR, 2018), vol. 80 of *Proceedings of Machine Learning Research*, pp. 4036–4044.

14. J. Andreas, D. Klein, S. Levine, *International Conference on Machine Learning* (PMLR, 2017), pp. 166–175.

15. R. V. Solé, S. Valverde, *J R Soc Interface* **5**, 129 (2008).

16. F. Damicelli, C. C. Hilgetag, M.-T. Hütt, A. Messé, *Netw Neurosci* **3**, 589 (2019).

17. K. D. Irvine, B. I. Shraiman, *Development* **144**, 4238 (2017).

18. A. M. Turing, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **237**, 37 (1952).

19. K. Zhang, *J Neurosci* **16**, 2112 (1996).

20. Y. Burak, I. R. Fiete, *PLoS Comput Biol* **5**, e1000291 (2009).

21. A. t. Banino, *Nature* **557**, 429 (2018).

22. C. J. Cueva, X.-X. Wei, *International Conference on Learning Representations* (2018).

23. B. Sorscher, G. Mel, S. Ganguli, S. Ocko, *Advances in Neural Information Processing Systems* (2019), pp. 10003–10013.

24. E. Urdapilleta, B. Si, A. Treves, *Hippocampus* **27**, 1204 (2017).

25. K. Yoon, M. Buice, R. Barry, C.and Hayman, N. Burgess, I. Fiete, *Nat Neurosci* **16**, 1077 (2013).

26. S. Trettel, J. Trimper, E. Hwaun, I. Fiete, L. Colgin, *Nat Neurosci* **22**, 609 (2019).

27. R. J. Gardner, L. Lu, T. Wernle, M.-B. Moser, E. I. Moser, *Nature neuroscience* **22**, 598 (2019).

28. L. Kang, V. Balasubramanian, *Elife* **8** (2019).

29. D. St Johnston, C. Nüsslein-Volhard, *Cell* **68**, 201 (1992).

30. L. Durrieu, *et al.*, *Mol Syst Biol* **14**, e8355 (2018).

31. S. W. Paddock, E. J. Hazen, P. J. DeVries, *BioTechniques* **22**, 120 (1997).

32. P. Beed, *et al.*, *Neuron* **79**, 1197 (2013).

33. T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, E. Moser, *Nature* **436**, 801 (2005).

34. H. Stensola, *et al.*, *Nature* **492**, 72 (2012).

35. Y. Burak, I. R. Fiete, *PLoS Comput Biol* **5**, e1000291 (2009).

36. K. W. Rogers, A. F. Schier, *Annu Rev Cell Dev Biol* **27**, 377 (2011).

37. F. Schweisguth, F. Corson, *Dev Cell* **49**, 659 (2019).

38. L. Wilson, M. Maden, *Dev Biol* **282**, 1 (2005).

39. A. Turing, *Philos Trans R Soc Lond BB* **237** (1952).

40. A. Gierer, H. Meinhardt, *Kybernetik* **12**, 30 (1972).

41. M. C. Cross, P. C. Hohenberg, *Reviews of Modern Physics* **65** (1993).

42. J. D. Murray, *Mathematical Biology*, no. Ch. 16 (Springer, Berlin, 2003).

43. X.-J. Wang, *Nat Rev Neurosci* **21**, 169 (2020).

44. A. Miri, *et al.*, *Nature neuroscience* **14**, 1150 (2011).

45. L. M. Giocomo, M. E. Hasselmo, *Journal of Neuroscience* **29**, 7625 (2009).

46. J. Widloski, UT Ph.D. Theses and Dissertations.

47. D. Thouless, *Topological quantum numbers in nonrelativistic physics* (World Scientific, 1998).

48. A. L. Krause, V. Klika, T. E. Woolley, E. A. Gaffney, *J R Soc Interface* **17**, 20190621 (2020).

49. M. Almuedo-Castillo, *et al.*, *Nat Cell Biol* **20**, 1032 (2018).

50. S. Werner, *et al.*, *Phys Rev Lett* **114**, 138101 (2015).

51. S. *et al.*, *Elife* **9**, 129 (2020).

52. L. F. Abbott, *et al.*, *Cell* **182**, 1372 (2020).

53. J. A. Bae, *et al.*, *bioRxiv* (2021).

54. S. ichi Amari, *Proceedings of the IEEE* **59**, 35 (1971).

55. M. S. Goldman, A. Compte, X.-J. Wang, *Encyclopedia of neuroscience* pp. 165–178 (2010).

56. S. Deneve, P. E. Latham, A. Pouget, *Nature neuroscience* **2**, 740 (1999).

57. N. Brunel, *Nature neuroscience* **19**, 749 (2016).

58. K. Yoon, S. Lewallen, A. A. Kinkhabwala, D. W. Tank, I. R. Fiete, *Neuron* **89**, 1086 (2016).

59. Y. Gu, *et al.*, *Cell* **175**, 736 (2018).

60. K. Yoon, *et al.*, *Nat Neurosci* **16**, 1077 (2013).

61. J. J. Couey, *et al.*, *Nat Neurosci* **16**, 318 (2013).

62. E. O. Tuck, *Bulletin of the Australian Mathematical Society* **74**, 133 (2006).

63. B. Giraud, R. Peschanski, *arXiv preprint math-ph/0504015* (2005).

64. J. Krupic, M. Bauza, S. Burton, C. Barry, J. O'Keefe, *Nature* **518**, 232 (2015).

65. T. Stensola, H. Stensola, M.-B. Moser, E. I. Moser, *Nature* **518**, 207 (2015).

66. G. Chen, D. Manson, F. Cacucci, T. J. Wills, *Current Biology* **26**, 2335 (2016).

67. J. J. Hopfield, *Proc Natl Acad Sci U S A* **81**, 3088 (1984).

68. R. Chaudhuri, A. Bernacchia, X.-J. Wang, *Elife* **3**, e01239 (2014).

Acknowledgments

We are grateful to Mehran Kardar for helpful discussions and to John Widloski, Akhilan Boopathy and Ling L. Dong for comments on the manuscript. This work was supported by the MathWorks Science Fellowship to MK, the Simons Foundation through the Simons Collaboration on the Global Brain, the ONR, and the Howard Hughes Medical Institute through the Faculty Scholars Program to IRF.

Conceptualization, writing: MK, SC, IRF; Coding, analysis: MK, SC.

Supplementary materials

Materials and Methods

We use a continuous attractor network (CAN) model (*55–57*) for grid cells (*35, 58–60*), with neural dynamics obeying

$$\frac{\partial s(i,t)}{\partial t} + \frac{s(i,t)}{\tau} = \phi\left[\sum_j W_0(i,j)s(j,t) + B(i,t)\right], \tag{1}$$

where $s(i,t)$ represents the synaptic activation of neuron i at time t, $W_0(i,j)$ represents the synaptic strength of the coupling from neuron j to neuron i, $B(i,t)$ represents the feed-forward bias to neuron i, and ϕ is a non-decreasing nonlinearity, for which we use the rectification function ($\phi(z) = [z]_+ = z$ for $z > 0$ and 0 otherwise). Each neuron i has a preferred direction θ_i that is used to perform velocity integration. In the one-dimensional version of our setup, each spatial location $\mathbf{x}$ on the neural sheet has two neurons, with preferred directions $\theta = 0$ and $\theta = \pi$. Correspondingly, in the two-dimensional version of our setup, each location on the neural sheet has four neurons, with preferred directions $\theta = n\pi/4$ for $n \in \{0,1,2,3\}$. The synaptic weights $W_0(i,j)$ are defined via an interaction kernel $W(\Delta x)$ such that

$$W_0(i,j) = W(|\mathbf{x}_i - \mathbf{x}_j - \mathbf{l}(\theta_j)|), \tag{2}$$

where $\mathbf{x}_i$ represents the spatial location of neuron i, and $\mathbf{l}(\theta)$ is a vector with length l oriented parallel to the angle θ. The feed-forward bias $B(i,t;\theta)$ is given by

$$B(i,t) = b + b_{vel}|\mathbf{v}|\cos(\theta_\mathbf{i} - \psi), \tag{3}$$

where ψ is the direction of the input velocity signal and $|\mathbf{v}|$ is the speed. This results in neurons with direction preference θ driving activity in the network towards the direction of their outgoing weight shifts $\mathbf{l}(\theta)$. This mechanism is responsible for velocity integration by the network (*35*).

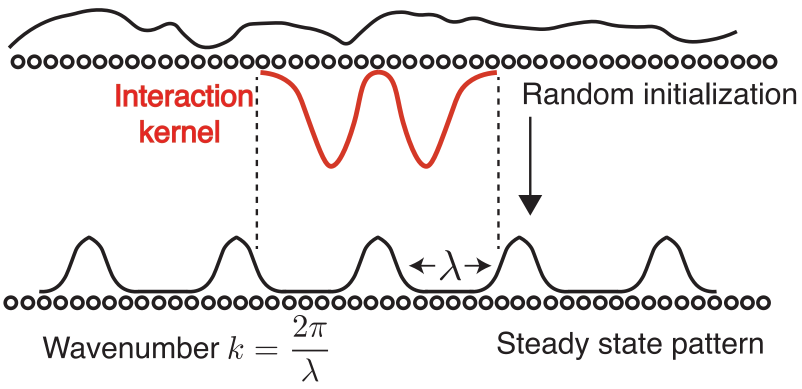


Figure 5: Local pattern formation in continuous attractor models of grid cells

As noted in the main text, the interaction weight kernel W is given by the sum of two components $W = W^g_{\mathbf{x}} + W^f$. The first, $W^g_{\mathbf{x}}$ drives local pattern formation, and has a spatial scale $\sigma(\mathbf{x})$, which varies smoothly in a gradient along the dorso-ventral axis, and the second, W^f has a fixed spatial scale d everywhere on the neural sheet. A variety of functions $W^g_{\mathbf{x}}$ can drive local pattern formation. For concreteness, we use two specific examples: the Mexican-hat profile (*35*) (used in Figs. 1-3 and SI Fig. 6)

$$W^g_{\text{mexican-hat}}(\Delta x) = \alpha_E \exp\left[-\gamma \frac{(\Delta x)^2}{2\sigma_{mh}(\mathbf{x})^2}\right] - \alpha_I \exp\left[-\frac{(\Delta x)^2}{2\sigma_{mh}(\mathbf{x})^2}\right], \tag{4}$$

and the box-function profile (*61*) (used in Fig. 2 and SI Fig. 6)

$$W^g_{\text{box}}(\Delta x) = \alpha_0 \times \mathbb{1}_{|\Delta x| < \sigma_b(\mathbf{x})} = \begin{cases} \alpha_0 & \text{if } |\Delta x| < \sigma_b(\mathbf{x}), \\ 0 & \text{if } |\Delta x| \geq \sigma_b(\mathbf{x}). \end{cases} \tag{5}$$

For the fixed-width interaction $W^f(\Delta x)$, we implement 3 main types — localized (used in

Figs. 2,3 and SI Fig. 6), diffuse (used in Fig. 2 and SI Fig. 6) and decaying (used in SI Fig. 6).

$$\begin{aligned} W^f_{\text{localized}}(\Delta x) &= \alpha_S \exp\left[-\frac{(|\Delta x| - d_{loc})^2}{2\epsilon_S^2}\right], \\ W^f_{\text{diffuse}}(\Delta x) &= \alpha_1 \times \mathbb{1}_{|\Delta x| < d_{dif}}, \\ W^f_{\text{decaying}}(\Delta x) &= \alpha_T \times [d_{dec} - |\Delta x|]_+. \end{aligned}$$

In particular,

- In Figs. 1e-g we use only a smoothly varying Mexican-hat pattern forming kernel $W = W^g_{\text{mexican-hat}}$
- In Figs. 2a-c,g,h and 3a,e we use $W = W^g_{\text{mexican-hat}} + W^f_{\text{localized}}$
- In Figs. 2d-f,i we use a 'Lincoln hat' profile $W = W^g_{\text{box}} + W^f_{\text{diffuse}}$
- and, in SI Fig. 6 we present numerical simulations of other combinations of pattern forming and fixed-scale kernels.

In Table 1 we present a list of common parameters used across all numerical simulations. Then, in Tables 2,3 we present the parameter values used for the kernels used in our numerical simulations

Supplementary Text

The supplemental information is structured as follows: We will first present the mathematical analysis for pattern formation, and demonstrate in SI Sec. 1 that simply introducing a gradient in the pattern forming kernel of the continuous attractor model is *not* sufficient to result in modularization, as demonstrated in Fig. 1 of the main text. Then, in Sec. 1.1 we show how the addition of a Gaussian localized kernel results in self-organized modularization. We then

Parameter	Value
τ	30
dt	0.05
b	$\begin{cases} 70 & \text{in 1D} \\ 1 & \text{in 2D} \end{cases}$
b_{vel}	$\begin{cases} 105 & \text{in 1D} \\ 1 & \text{in 2D} \end{cases}$
l	2

Table 1: Parameters held constant across all numerical simulations

$W^g_{\text{mexican-hat}}$ parameters	Value
α_E	1000
α_I	1000
γ	1.05
N^{1D}	3000
N_y^{2D}	100
N_x^{2D}	1000
$\sigma_{mh}(x)$	$1/\sqrt{2\beta(x)}$
$\beta(x)$	$\beta_0 + (\beta_1 - \beta_0)x/N'$
N'	$\begin{cases} N^{1D} & \text{in 1D} \\ N_x^{2D} & \text{in 2D} \end{cases}$
β_0	$\begin{cases} 2.5 \times 10^{-2} & \text{in 1D} \\ 3/676 & \text{in 2D} \end{cases}$
β_1	$\begin{cases} 2.5 \times 10^{-1} & \text{in 1D} \\ 9/338 & \text{in 2D} \end{cases}$

W^g_{box} parameters	Value
N^{1D}	5000
α_0	-40
$\sigma_b(x)$	$15 + 30x/N$

Table 2: Pattern forming kernel parameters used for numerical simulations

$W^f_{\text{localized}}$ parameters	Value
α_S	4
d_{loc}	$\begin{cases} 84 & \text{in 1D} \\ 50 & \text{in 2D} \end{cases}$
ϵ_S	$\begin{cases} 4.77 & \text{in 1D} \\ 1.6 & \text{in 2D} \end{cases}$
W^f_{diffuse} parameters	Value
α_{dif}	-0.25
d_{dif}	135
W^f_{decaying} parameters	Value
α_T	25
d_{dec}	150

Table 3: Fixed-scale kernel parameters used for numerical simulations

generalize this result in SI Sec. 2, to show that arbitrary fixed-scale kernels will almost always result in module formation as demonstrated in Fig. 3. Among arbitrary kernels, we will show in Sec. 3 that kernels with a simple features result in a simple equation describing the detailed period ratios of the formed grid modules as shown in Fig. 4. This will lead to simple estimates for the number of modules and their sizes in terms of other system parameters, which we derive in SI Sec. 3.7. After having described our results primarily for the case of one-dimensional grid cells, we then demonstrate in Sec. 3.8 that our arguments extend naturally to two dimensions, and we present numerical results demonstrating the same. In SI Sec. 3.9 we then demonstrate that our results and predictions of grid period ratios are consistent with available data sources to a large extent. We then end with some brief remarks relating our result to the context of energy minimization in neural networks (SI Sec. 4), and providing broader perspectives of our results in the contexts of general loss optimization (Sec. 5) and eigenvector localization (SI Sec. 6).

We will largely restrict our analysis to the continuum limit of a large number of neurons (a neural field model) for mathematical convenience. However, as shown in the main text, the results hold for and accurately predict the dynamics of networks of discrete neurons. We will use a continuous index to represent position along the neural sheet. In this limit, the dynamics of rate-based neurons is given by:

$$\frac{\partial s(\mathbf{x},t)}{\partial t} + \frac{s(\mathbf{x},t)}{\tau} = \phi\left[\int_{-\infty}^{+\infty} W(\mathbf{x},\mathbf{x}')s(\mathbf{x}',t)d\mathbf{x}' + B(\mathbf{x})\right], \tag{6}$$

where $s(\mathbf{x})$ is the synaptic activation of the neuron at the vector position $\mathbf{x}$ on the neural sheet, τ is the biophysical time-constant of individual neurons, ϕ is a monotonic non-decreasing nonlinearity, and $B(\mathbf{x})$ is the feedforward input to the neuron. For simplicity, we will use the the rectification function ($\phi(z) = [z]_+ = z$ for $z > 0$ and 0 otherwise) as the nonlinearity in all of our results.

In nonlinear continuous attractor models the interaction weights $W(\mathbf{x},\mathbf{x}')$ are chosen to have a continuous symmetry, usually a translation-invariant symmetry such that $W(\mathbf{x},\mathbf{x}') = K(|\mathbf{x}-\mathbf{x}'|)$, where the weight between neurons at locations $\mathbf{x}$ and $\mathbf{x}'$ depends only on their separation, and not on their absolute locations. One example of such a weight K is the so-called Mexican-hat function described by a difference-of-Gaussians, in which neurons excite their immediate neighbors and inhibit those slightly further away; there is no interaction between faraway neurons:

$$W(\mathbf{x},\mathbf{x}') = W(\mathbf{\Delta x}) = \alpha_E \exp\left(-\frac{\mathbf{\Delta x}^2}{2\sigma_E}\right) - \alpha_I \exp\left(-\frac{\mathbf{\Delta x}^2}{2\sigma_I}\right). \tag{7}$$

Continuous attractor grid cell models of individual modules rely principally on a local inhibitory interaction (the excitatory center interaction is dispensable) (*35, 61*).

Motivated by the experimental observations described in the main text, we modify the Mexican-hat function to introduce a smooth gradient in the characteristic interaction widths σ_E, σ_I.

$$W^g_{\mathbf{x}}(\mathbf{\Delta x}) = \alpha_E \exp\left(-\frac{\mathbf{\Delta x}^2}{2\sigma_E(\mathbf{x})}\right) - \alpha_I \exp\left(-\frac{\mathbf{\Delta x}^2}{2\sigma_I(\mathbf{x})}\right), \tag{8}$$

where $\sigma_E(\mathbf{x})$ and $\sigma_I(\mathbf{x})$ are now functions that depend on position in the neural sheet, and encode the smoothly varying characteristic scale of the Mexican-hat interaction, say

$$\sigma_{E/I}(\mathbf{x}) = \sigma_{E/I} + \sigma'_{E/I}(0) \cdot \mathbf{x}. \tag{9}$$

For such graded kernels, we will use $W(\mathbf{x}, \mathbf{x}')$ and $W_{\mathbf{x}}(\mathbf{x} - \mathbf{x}') = W_{\mathbf{x}}(\mathbf{\Delta x})$ interchangeably.

1 Pattern formation with graded kernels

In the limit of very slow changes in the length-scale of the interaction kernel $W_{\mathbf{x}}(\mathbf{\Delta x})$ along the dorso-ventral axis of the MEC, we can analyze the pattern formation process at position $\mathbf{x}$ by assuming that $W_{\mathbf{x}}(\mathbf{\Delta x})$ is locally constant.

Under this approximation, we perform a linear stability analysis of the neural dynamics, to identify the the growing periodic modes locally at the position on the neural sheet $\mathbf{x}$.

We first identify an unstable steady-state solution to Eq. (6), which we denote as $s_0(\mathbf{x})$. This solution satisfies

$$\frac{s(\mathbf{x})}{\tau} = \phi\left[\int_{-\infty}^{+\infty} W_{\mathbf{x}}(\mathbf{x} - \mathbf{x}')s_0(\mathbf{x}')d\mathbf{x}' + B(\mathbf{x})\right]. \tag{10}$$

In the limit of very slowly varying changes in $W_{\mathbf{x}}(\mathbf{\Delta x})$ as a function of $\mathbf{x}$, the unstable steady state solution will be

$$s_0(\mathbf{x}) = \frac{\tau\bar{B}}{1 - \tau\bar{W}}, \tag{11}$$

where $\bar{B} = \int B(\mathbf{x})d\mathbf{x}$ and $\bar{W} = \int W_{\mathbf{x}}(\mathbf{x} - \mathbf{x}')d\mathbf{x}'$.

We then consider a perturbative analysis, by examining the evolution of $s(\mathbf{x}, t) = s_0(\mathbf{x}) + \epsilon(\mathbf{x}, t)$. We apply our analysis to the early time evolution of this initial condition, such that $\epsilon(\mathbf{x}, t) \ll s_0(\mathbf{x})$. Inserting our form of $s(\mathbf{x}, t)$ in Eq. (6), we obtain

$$\frac{\partial\epsilon(\mathbf{x}, t)}{\partial t} + \frac{\epsilon(\mathbf{x}, t)}{\tau} = \phi'(\gamma\bar{W}s_0(\mathbf{x}) + B)\int_{-\infty}^{\infty} W_{\mathbf{x}}(\mathbf{x} - \mathbf{x}')\epsilon(\mathbf{x}', t)d\mathbf{x}' \tag{12}$$

Since $W_{\mathbf{x}}(\mathbf{x} - \mathbf{x}')$ is a local kernel, we approximate the above integral with one evaluated over the region $\{\mathbf{x}' : |\mathbf{x} - \mathbf{x}'| < L\}$, with L much larger than the length-scale of the kernel $W_{\mathbf{x}}$ at all

$\mathbf{x}$. Over this interval, we posit that $\epsilon(\mathbf{x}', t) = \epsilon e^{i\mathbf{k}\cdot\mathbf{x}' + \alpha(\mathbf{k})t}$, where $\alpha(\mathbf{k})$ denotes the growth rate of this ϵ perturbation. Inserting this form into Eq. (12) yields,

$$\alpha(\mathbf{k}) + 1/\tau = \phi'(\bar{W}s_0(\mathbf{x}) + B)\int_{\infty}^{\infty} W_{\mathbf{x}}(\mathbf{x}-\mathbf{x}')e^{-i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}d\mathbf{x}', \tag{13}$$

$$= \phi'[\bar{W}s_0(\mathbf{x}) + B]\mathcal{F}[W_{\mathbf{x}}(\mathbf{x}-\mathbf{x}')], \tag{14}$$

$$= \phi'[\bar{W}s_0(\mathbf{x}) + B]\mathcal{F}W_{\mathbf{x}}(\mathbf{k}) \tag{15}$$

where $\mathcal{F}[W_{\mathbf{x}}(\mathbf{x}-\mathbf{x}')] = \mathcal{F}W_{\mathbf{x}}(\mathbf{k})$ is the Fourier transform of the interaction kernel corresponding to position $\mathbf{x}$ on the neural sheet.

Note that since $W_{\mathbf{x}}(\boldsymbol{\Delta}\mathbf{x})$ is a kernel, it is a radially-symmetric real function, and hence the Fourier transform $\mathcal{F}W_{\mathbf{x}}(\mathbf{k})$ will also be real function that is radially-symmetric in k. Thus, for simplicity, we will only focus on the magnitude of $\mathbf{k}$, which we denote as $k = |\mathbf{k}| \geq 0$ (In this context, for the two-dimensional case, one may re-interpret the radial component of the Fourier transform of $W_{\mathbf{x}}(\boldsymbol{\Delta}\mathbf{x})$ as the Hankel transform of $W_{\mathbf{x}}(|\boldsymbol{\Delta}\mathbf{x}|)$).

By definition, the magnitude of the wave vector k^* that corresponds to the fastest growing mode locally around position $\mathbf{x}$ on the neural sheet will be the $\mathbf{k}$ that maximizes $\alpha(\mathbf{k})$. Under the approximation of slow changes in the length-scale of the interaction kernel $W_{\mathbf{x}}(\boldsymbol{\Delta}\mathbf{x})$, we see from Eq. (15) that

$$k^*(\mathbf{x}) = \arg\max_k \mathcal{F}W_{\mathbf{x}}(k), \tag{16}$$

since $W_x(\boldsymbol{\Delta}\mathbf{x})$ (and hence $s_0(\mathbf{x})$) has been assumed to have a negligible dependence on $\mathbf{x}$.

For $W_{\mathbf{x}}(\boldsymbol{\Delta}\mathbf{x})$ given by Eq. (8), i.e., without any additional fixed-scale interaction, we obtain from Eq. (16)

$$[k^*(\mathbf{x})]^2 = \frac{2}{\sigma_E(\mathbf{x})^2 - \sigma_I(\mathbf{x})^2}\log\left(\frac{\alpha_E\sigma_E(\mathbf{x})^3}{\alpha_I\sigma_I(\mathbf{x})^3}\right). \tag{17}$$

If we assume that $\sigma_{E/I}(\mathbf{x}) = \eta_{E/I}\sigma(\mathbf{x})$, where η_E and η_I are $\mathbf{x}$-independent constants, then we obtain

$$k^*(\mathbf{x}) \propto 1/\sigma(\mathbf{x}), \tag{18}$$

and hence

$$\lambda^*(\mathbf{x}) \propto \sigma(\mathbf{x}), \tag{19}$$

where $\lambda * (\mathbf{x})$ is the periodicity of the grid pattern formed locally around position $\mathbf{x}$. This results in a smooth change of grid period, corresponding to the observation in Fig. 1g of the main text.

Note that this result is generally true for any pattern forming kernel $W^g_{\mathbf{x}}(\mathbf{\Delta x})$ that has a Fourier transform with at least one local maximum, and does not rely on the specific form of a Mexican-hat interaction. Indeed, Eq. (19) holds for any kernel $W^g_x(\Delta x)$ that depends on a length-scale $\sigma(x)$. As an example, we present the corresponding analysis for the box-shaped kernel employed for pattern formation in Ref. (*61*). In this case

$$W^g_{\mathbf{x}}(\mathbf{\Delta x}) = -W_0 \nVdash_{\mathbf{\Delta x} \leq \sigma(\mathbf{x})}. \tag{20}$$

As discussed above, the quantity of interest is $\mathcal{F}W^g_{\mathbf{x}}(\mathbf{k})$

$$\mathcal{F}W^g_{\mathbf{x}}(\mathbf{k}) = \int_{-\infty}^{\infty} -W_0 \nVdash_{|\mathbf{x}| \leq \sigma(\mathbf{x})} e^{i\mathbf{k}\cdot\mathbf{x}} d\mathbf{x} \tag{21}$$

$$= -W_0 \int_{|\mathbf{x}| \leq \sigma(\mathbf{x})} e^{i\mathbf{k}\cdot\mathbf{x}} d\mathbf{x}. \tag{22}$$

The above integral can be calculated in a one-dimensional setup to obtain

$$\mathcal{F}W^g_x(k) = -2W_0 \frac{\sin(k\sigma(x))}{k} \tag{23}$$

and can be calculated in a two-dimensional setup to obtain

$$\mathcal{F}W^g_{\mathbf{x}}(k) = -2\pi W_0 \sigma(\mathbf{x}) \frac{J_1(k\sigma(\mathbf{x}))}{k}. \tag{24}$$

In both of the above cases, note that $k^* \propto 1/\sigma(\mathbf{x})$ since $\sigma(\mathbf{x})$ is the only length-scale characterizing the kernel $W^g_{\mathbf{x}}$. In particular, numerical maximization yields

$$k^* \approx \begin{cases} 4.493/\sigma(x) & \text{if } x \text{ is one-dimensional, and} \\ 5.136/\sigma(\mathbf{x}) & \text{if } \mathbf{x} \text{ is two-dimensional.} \end{cases} \tag{25}$$

1.1 Fixed-scale interactions and modularization

We now claim that the addition of a fixed-scale kernel, $W^f(\mathbf{\Delta x})$ is sufficient to result in modularization of grid periods, with discrete changes in grid period as a function of spatial position along the dorso-ventral axis. This set of interactions can effectively be implemented by two populations of interneurons - one with fixed arborization and weaker synaptic connections and one with varying arborization length and stronger synaptic connections.

For simplicity, we shall present the specific Fourier transform computations for the one-dimensional problem, although we note that all of the qualitative results hold in two dimensions as well, with the Fourier transforms of the relevant functions replaced with their Hankel transforms (as shown in Sec. 3.8).

We include an additional weak interaction term W^f that critically does *not* depend on the neural sheet position x. For reasons that will become apparent soon, we choose kernels $W^f(\Delta x)$ such that the Fourier transform changes sign a sufficiently large number of times. We hypothesize that this requirement is not particularly restrictive, and will demonstrate that this holds for most kernels W^f.

The entire interaction profile is then given by

$$W_x(\Delta x) = W_x^g(\Delta x) + W^f(\Delta x). \tag{26}$$

We first demonstrate our result with an example of a simple kernel, to justify how Eq. (16) leads to the emergence of discrete grid modules. Consider the localized excitatory interaction

$$W^f(\Delta x) = \alpha_S \exp\left(-\frac{(\Delta x - d)^2}{2\sigma_S^2}\right) + \alpha_S \exp\left(-\frac{(\Delta x + d)^2}{2\epsilon_S^2}\right). \tag{27}$$

Corresponding to our interpretation of $W^f(\Delta x)$ above being a localized kernel, we choose $\epsilon_S \ll d$.

This choice of $W_x(\Delta x) = W_x^g(\Delta x) + W^f(\Delta x)$ leads to the the Fourier transform,

$$\begin{aligned}\mathcal{F}W_x(k) &= \mathcal{F}W_x^g(k) + \mathcal{F}W^f(k), \qquad (28)\\ &= \sqrt{2\pi}\left[\alpha_E\sigma_E(x)\exp\left(-\frac{\sigma_E(x)^2k^2}{2}\right) - \alpha_I\sigma_I(x)\exp\left(-\frac{\sigma_I(x)^2k^2}{2}\right) + 2\alpha_S\epsilon_S\cos(kd)\exp\left(-\frac{\epsilon_S^2k^2}{2}\right)\right. \qquad (29)\end{aligned}$$

In our model, the magnitude of the $W^f(\Delta x)$, i.e., α_S, is chosen to be smaller than the magnitude of the Mexican-hat interaction. Thus we interpret $\mathcal{F}W^f(k)$ in Eq. (29) as being a small perturbation to the Fourier transform of the standard Mexican-hat interaction, $\mathcal{F}W_x^g(k)$. Further, since d is assumed to be much larger than the scale of the Mexican-hat, $\sigma_{E/I}$, then the term $\cos(kd)$ in $\mathcal{F}W^f(k)$ oscillates at a k-scale much smaller than the relevant scales of $\mathcal{F}W_x^g(k)$ (see Fig. 3b-c of the main text). Additionally, since $\epsilon_S \ll d$, the gaussian envelope multiplying the rapidly oscillating term has a scale $1/\epsilon$, which is much larger than the periodicity $1/d$.

Thus, in k-space, the rapidly oscillating term, $\mathcal{F}W^f(k)$ can be thought of as predefining a set $S = \{k_1, k_2, \ldots\}$ of local maxima. Under the approximations made above, the addition of the smoother function $\mathcal{F}W_x^g(k)$, will not change the position of the local maxima. This results in the *local* maxima of $\mathcal{F}W_x(k)$ also being the same set S. Importantly, we note that since S was predefined purely via $\mathcal{F}W^f(k)$, *there is no x dependence on the set S.*

Following Eq. (16), the wave-vector corresponding to the pattern formation at point x on the neural sheet corresponds to the *global* maxima of $\mathcal{F}W_x(k)$. Thus, at all points, the pattern formation corresponds to one of the discrete set of choices of wave vectors, $S = \{k_1, k_2 \ldots\}$. As can be seen from Fig. 3c, the smoothly varying gradient in the Mexican-hat term, $\mathcal{F}W_x^g$ as a function of x picks different choices of k_i depending on the position x — the $k \in S$ that is nearest to the maxima of $\mathcal{F}W_x^g(k)$ will be chosen as the global maxima, and will be the wave vector corresponding to the pattern at x. We refer to this mechanism as "peak selection".

For our particular choice of $W^f(x)$ made in Eq. (27), we obtained

$$\mathcal{F}W^f(k) = 2\alpha_S \epsilon_S \cos(kd) \exp\left(-\frac{\epsilon_S^2 k^2}{2}\right). \tag{30}$$

We can then approximate the local maxima of $\mathcal{F}W^f(k)$ as occurring at

$$S = \left\{ \frac{2m\pi}{d} \middle| m \in \mathbb{Z}^+ \right\}. \tag{31}$$

This immediately indicates that the ratios of periods of successive grid modules will be given by

$$\frac{\lambda_{m+1}}{\lambda_m} = \frac{m+1}{m}. \tag{32}$$

Thus, the addition of a fixed-scale interaction, W^f such as Eq. (27) results in discrete grid modules. We now show that this peak-selection mechanism, and hence modularization, occurs for arbitrary choices of the fixed-scale interaction kernel $W^f(\Delta x)$.

2 Kernels that lead to modularization

The peak-selection modularization mechanism described above arises naturally from the presence of the rapidly oscillating term in $\mathcal{F}W^f(k)$. In fact, for discrete grid modules to occur, the only constraints imposed on the fixed-scale kernel W^f are: (a) the Fourier transform $\mathcal{F}W^f(k)$ must have a sufficiently large number of maxima (at least 4 maxima, corresponding to the 4 grid modules observed in experimental observations); and, (b) these maxima must be at scales smaller than $1/\sigma$ in k-space. Here we argue that this is generally true for arbitrary kernels, modulo a single scaling parameter.

We hypothesize and give support, without formal proof, that almost every arbitrarily chosen kernel $W^f(\Delta x)$ will have a Fourier transform with multiple maxima satisfying condition (a). We will then argue that this kernel can always be scaled to satisfy condition (b).

To motivate our hypothesis, we first note that it is actually possible to construct specific kernels $W^f(\Delta x)$ whose Fourier transform does not present multiple maxima. For example, the Gaussian kernel, $W_{gauss}(\Delta x) = \exp[-(\Delta x)^2/2]$, results in a Fourier transform that is unimodal. However, we hypothesize that such functions are rare in the space of all continuous functions in L^2. Indeed, we can construct a function that is arbitrarily close to the Gaussian kernel whose Fourier transform will have an infinite number of maxima: Let $f_0(\Delta x) = \nVdash_{[-1,1]}$ be the box function. Define

$$f_n = f * f_{n-1}$$

for all $n \geq 1$, where $f * g$ represents the convolution of functions f and g. By the central limit theorem, $\sqrt{n} f_n(\sqrt{n}\Delta x)$ will approach $W_{gauss}(\Delta x)$. However,

$$\mathcal{F} f_n(k) = [2\sin(k)/k]^n, \tag{33}$$

which clearly has an infinite number of maxima. Thus, even though the Gaussian kernel has a unimodal Fourier transform, we can construct a function $g_n(\Delta x) = \sqrt{n} f_n(\sqrt{n}\Delta x)$ that is arbitrarily close to the Gaussian kernel (for sufficiently large n) but has a Fourier transform that presents an infinite number of maxima.

In this context, we claim that almost every arbitrarily chosen kernel $W^f(\Delta x)$ will have a Fourier transform with multiple maxima. This may be intuited as follows: First note that Fourier space is a dual space, and hence instead of considering arbitrary kernels in real space we may equivalently choose arbitrary kernels in Fourier space. Further assuming that $\mathcal{F}W^f(k)$ is a smooth function, we hypothesize that generically smooth functions that are in L^2 will almost always have multiple maxima and minima.

Thus condition (a) may be satisfied for arbitrary kernels $W^f(\Delta x)$. Next, note that scaling a function in real space results in an inverse scaling of the Fourier transform, i.e., $\mathcal{F}[W^f(a\Delta x) = \mathcal{F}W^f(k/a)$. Hence, we can always scale the function $W^f(\Delta x)$ to obtain a Fourier transform with maxima that are within any desired scale, allowing condition (b) to be satisfied.

In Fig. 6, we show examples of modularization arising from different combinations of graded pattern forming kernels (W^g) and fixed-scale kernels (W^f). In each case, we also present the expected periodicity in each module as a function of spatial position as given by the perturbative analysis Eq. (16). The analytical result based on linear stability provides an excellent prediction of the pattern periods per module (see also Main text, Fig. 3e). It also predicts the locations of the module boundaries (see also Main text, Fig. 3e) though not as accurately: module boundary predictions tend to be slightly but systematically offset relative to the simulated dynamics, due to the effects of nonlinearity in the later stages of pattern formation.

3 Simple kernels and period ratios

What kinds of fixed-scale interactions might be present in the medial-Entorhinal cortex? As described in the main text, in the context of biology, we might expect *simple* interaction kernels W^f to be relevant i.e., the fixed-scale interaction profile W^f has the following characteristics: (a) there exists a *single* length-scale d that primarily characterizes the shape of W^f; (b) any other length-scales relevant to W^f, say scales $\epsilon_1, \epsilon_2, ...$ are each much smaller than the primary length scale d. Further, we assume that the primary length-scale associated with the fixed-scale interaction is larger than the length-scales of the pattern forming kernel, i.e., $d \gg \sigma_{E/I}(x)$.

We will demonstrate that *simple* fixed-scaled interaction kernels result in analytic expressions for grid periods that are characterized by a single angular variable ϕ

$$\frac{\lambda_{m+1}}{\lambda_m} = \frac{m+1+\phi/(2\pi)}{m+\phi/(2\pi)}. \tag{34}$$

Before filling in the details of our argument, we present an intuitive explanation of the general idea:

Consider the following basic classes of *simple* kernels that satisfy the above-described cri-

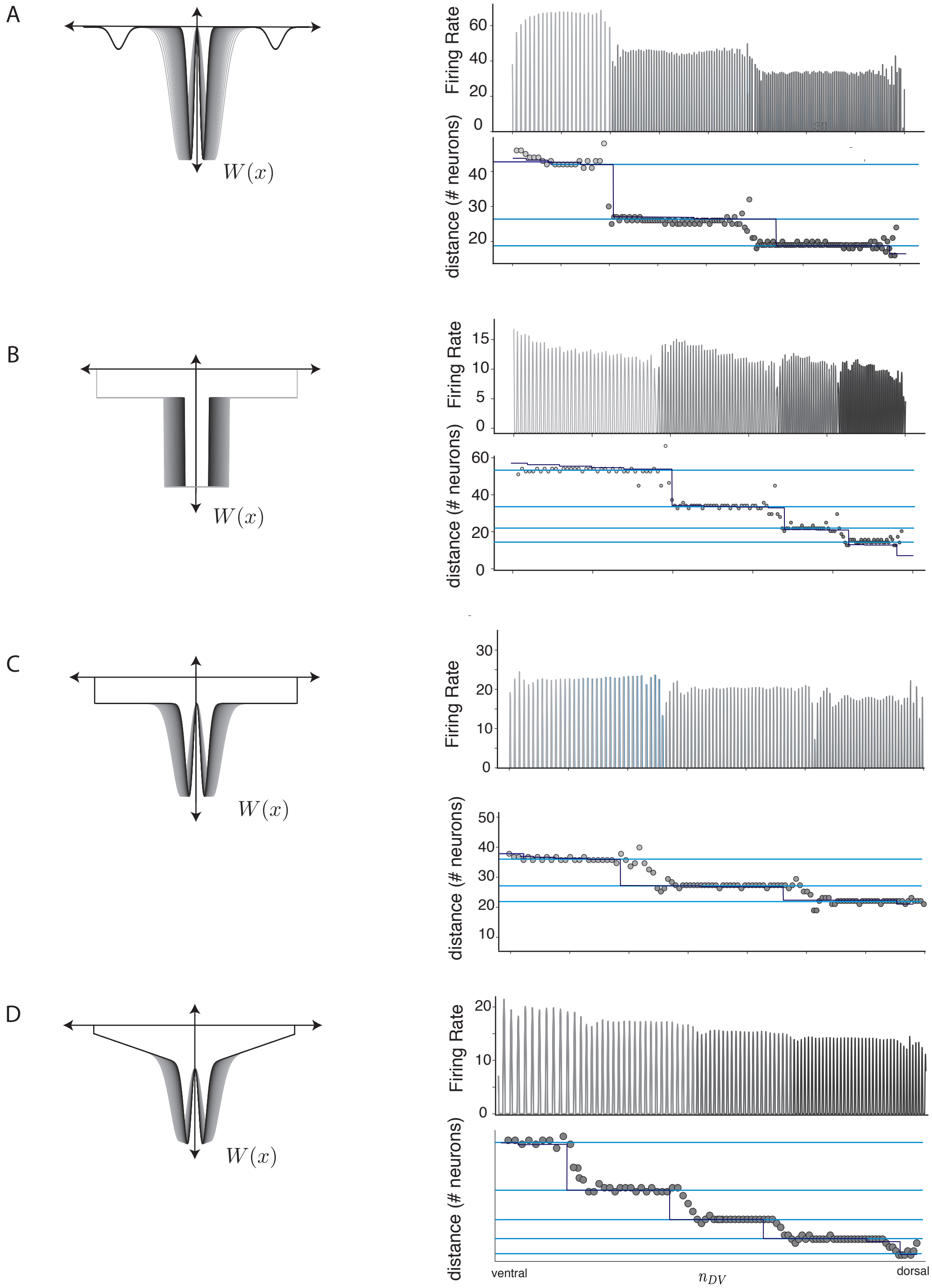


Figure 6: Examples of modularization and (right column) population activity with (left column) various pattern forming and fixed interactions.

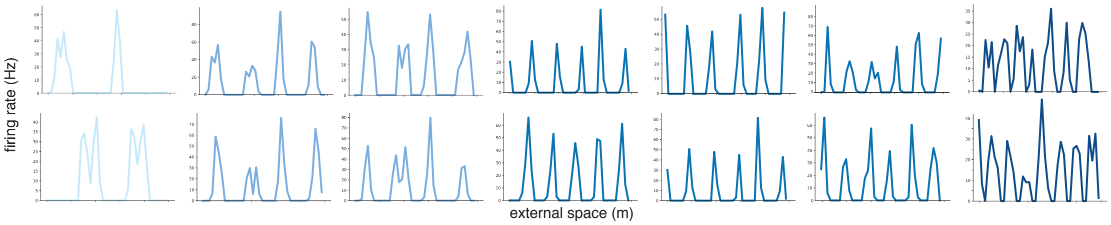


Figure 7: Sample tuning curves from several neurons in all modules from the network of Fig 2a.

teria corresponding to a length-scale d:

(a) $g(|\Delta x| - d)$, for arbitrary functions $g(\rho)$ that are nonzero only over scales $|\rho| < \epsilon_i$ (a *localized* kernel), and,

(b) A constant term, that is uniform everywhere up to $\Delta x = d$, after which it falls to zero (a *diffuse* kernel),

(c) A decaying term, that decreases from a constant value at $\Delta x = 0$ to zero at $\Delta x = d$ (a *decaying* kernel).

We also define *short-range* kernels, as any arbitrary function $h(\Delta x)$ that is nonzero only over scales $|x| < \epsilon_i$.

Any *simple* kernel $W^f(\Delta x)$ can be generally constructed as a linear combination of the above basic classes. In addition, *simple* kernels may also contain an added component of a *short-range* kernel.

To see that *simple* kernels will generally result in grid period ratios corresponding to Eq. (34), we will examine the approximate Fourier transform structure for each component of the linear combination of *simple* kernels corresponding to a given length-scale d. We first demonstrate that each of the basic *simple* kernels will result in Fourier transforms that are sinusoidal functions with phase shifts and decaying envelopes and hence each basic *simple* kernel will satisfy Eq. (34). We then show that short-range kernels present Fourier transforms that vary only at large scales, and can be ignored in our analyses of *simple* kernels. We then use these results to demonstrate that all *simple* kernels constructed as the above-described linear combination will have sinusoidal Fourier transforms and will satisfy Eq. (34).

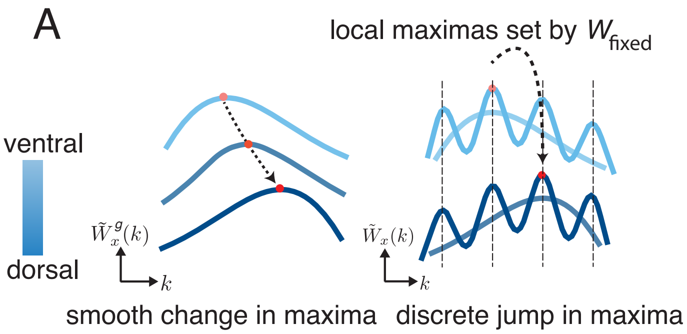


Figure 8: **Peak selection process** demonstrating how the additional of the oscillatory fourier transform of the fixed interaction leads to discrete jumps in maxima despite smooth movement of the primary peak.

3.1 Localized kernels

For a general localized kernel $W^f(\Delta x) = g(|\Delta x| - d)$ we obtain

$$\mathcal{F}W^f(k) = \Re[e^{-ikd}\mathcal{F}g(k)]. \tag{35}$$

Since $g(x)$ is supported over a scale ϵ, the Fourier transform $\mathcal{F}g(k)$ will only vary at scales $k \sim 1/\epsilon \gg 1/d$. Thus for $1/d \ll k \ll 1/\epsilon$, we can approximate Eq. (35) as

$$\mathcal{F}W^f(k) = |\mathcal{F}g(k)| \cos\left(kd - \psi\right), \tag{36}$$

where $\psi = \arg[\mathcal{F}g(k)]$. The local maxima of $\mathcal{F}W^f(k)$ will then occur at

$$S = \left\{ \frac{2m\pi + \psi}{d} \middle| m \in \mathbb{Z}^+ \right\}, \tag{37}$$

resulting in period ratios described by

$$\frac{\lambda_{m+1}}{\lambda_m} = \frac{m + 1 + \psi/(2\pi)}{m + \psi/(2\pi)}, \tag{38}$$

which is identical to Eq. (34) for $\phi = \psi$. We also note that we can now ascribe an interpretation to the phase angle ϕ — it is the phase difference between $\mathcal{F}W^f(k)$ and $\cos(kd)$.

3.2 Diffuse kernels

We model a diffuse interaction kernel $W^f(x)$ as

$$W^f(x) = -W_0 \mathbb{1}_{[-d,d]} = \begin{cases} -W_0 & \text{if } |x| \leq d \\ 0 & \text{if } |x| > d \end{cases}. \tag{39}$$

Corresponding to the discussion above, we look at the Fourier transform $\mathcal{F}W^f(k)$

$$\mathcal{F}W^f(k) = \int_{-\infty}^{+\infty} -W_0 \mathbb{1}_{[-d,d]} e^{ikx} dx = \int_{-d}^{+d} -W_0 e^{ikx} dx \tag{40}$$

$$= -2W_0 \frac{\sin(kd)}{k} = -2W_0 d \,\text{sinc}(kd). \tag{41}$$

Note that once again, similar to Eqn. (30), we obtain a functional form consisting of a periodic function ($\sin(kd)$) that is multiplied by a decaying envelope $1/(kd)$. Ignoring the effects of the envelope function, the maxima of this function occur at

$$S \approx \left\{ \frac{2m\pi - \pi/2}{d} \middle| m \in \mathbb{Z}^+ \right\}, \tag{42}$$

which immediately results in period ratios of the form

$$\frac{\lambda_{m+1}}{\lambda_m} \approx \frac{m + 1 - 1/4}{m - 1/4}, \tag{43}$$

which corresponds to the result in Eq. (34) for $\phi = \pi/2$.

More precisely, the extrema of $\mathcal{F}W^f(k)$ occur at $k_m d = q - 1/q - 2/3q^3 + O(q^{-5})$ where $q = \left(m + \frac{1}{2}\right)\pi$. Notably, the errors decay approximately as $1/(\pi m)$, and thus for modules generated corresponding to $m \gtrsim 2$ will result in period ratios that approximate Eq. (34) closely.

3.3 Decaying kernels

Decaying kernels with a scale d may be modeled as any monotonically decreasing function that decays from some constant W_0 at $\Delta x = 0$, to zero, at $\Delta x = d$. For simplicity, we consider the simplest linear approximation to such a kernel, modeled as a triangular kernel. For additional subtleties in the treatment of other decaying kernels, see 3.5.1 The triangular kernel can be written as:

$$W^f(\Delta x) = \begin{cases} W_0(\Delta x - d)/d & \text{if } \Delta x < d \\ 0 & \text{if } \Delta x \geq d \end{cases} \tag{44}$$

This function can be written as the convolution of 2 diffuse box functions:

$$W^f(\Delta x) = (-W_0 \mathbb{1}_{[-d/2,d/2]}) * (W_0 \mathbb{1}_{[-d/2,d/2]}).$$

Thus, its Fourier transform is:

$$\begin{aligned} \mathcal{F}W^f(k) &= -W_0^2 d^2 \left(\frac{\sin(kd/2)}{(kd/2)} \right)^2 \\ &= -\frac{2W_0^2}{k^2}[1 - \cos(kd)]. \end{aligned}$$

Once again, we obtain a simple trigonometric function, with maxima at

$$S \approx \left\{ \frac{2m\pi}{d} \middle| m \in \mathbb{Z}^+ \right\}, \tag{45}$$

which immediately results in period ratios of the form

$$\frac{\lambda_{m+1}}{\lambda_m} \approx \frac{m+1}{m}, \tag{46}$$

which corresponds to the result in Eq. (34) for $\phi = 0$.

3.4 Short-range kernels

For the case of a short-range kernel $W^f(\Delta x)$ that extends upto a scale ϵ, we note from the Fourier uncertainty principle that the characteristic k-scales of $\mathcal{F}W^f(k)$ will $\sim 1/\epsilon \gg 1/d$.

Thus, unlike the three other types of simple kernels discussed above, short range kernels do not have structure at the scale of $1/d$. Since all relevant scales are much larger than $1/d$, adding short range kernels to any of the other types of *simple* kernels will *not* change the structure of local maxima at scales of $1/d$.

3.5 Arbitrary simple kernels

We now consider a general form for *simple* kernels, by constructing linear combinations of the above described three basic classes of *simple* kernels each corresponding to the same length scale d and additional short-range kernels.

$$W^f = a_{\text{local}} W^f_{\text{local}} + a_{\text{diffuse}} W^f_{\text{diffuse}} + a_{\text{decaying}} W^f_{\text{decaying}} + a_{\text{short}} W^f_{\text{short}}. \tag{47}$$

As demonstrated in the preceding sections, the Fourier transform $\mathcal{F}W^f(k)$ will be given as

$$\mathcal{F}W^f(k) = a_{\text{local}}|\mathcal{F}g(k)|\cos(kd-\psi) - 2W_0 a_{\text{diffuse}}\sin(kd)/k - 2W_0^2 a_{\text{decaying}}(1-\cos(kd))/k + \mathcal{F}h(k) \tag{48}$$

$$= H_0(k) + \sum_{i=0}^{3} H_i(k)\cos(kd+\phi_i) \tag{49}$$

for some constants ϕ_i, and some envelope functions $H_i(k)$ for $i = 0, 1, 2, 3$ that are slowly varying for $kd \gtrsim \mathcal{O}(1)$. Under this approximation, $\mathcal{F}W^f(k)$ is simply the sum of multiple sinusoidal waves with different phases and identical frequencies. Thus,

$$\mathcal{F}W^f(k) \approx \cos(kd-\phi) \tag{50}$$

for some ϕ and $kd \gtrsim \mathcal{O}(1)$. Hence, the maxima of $\mathcal{F}W^f(k)$ occur at

$$S \approx \left\{ \frac{2n\pi+\phi}{d} \middle| n \in \mathbb{Z}^+ \right\}, \tag{51}$$

which immediately results in period ratios of the form Eq. (34). Note that the approximations made above imply that there may be deviations from our results for the maxima corresponding

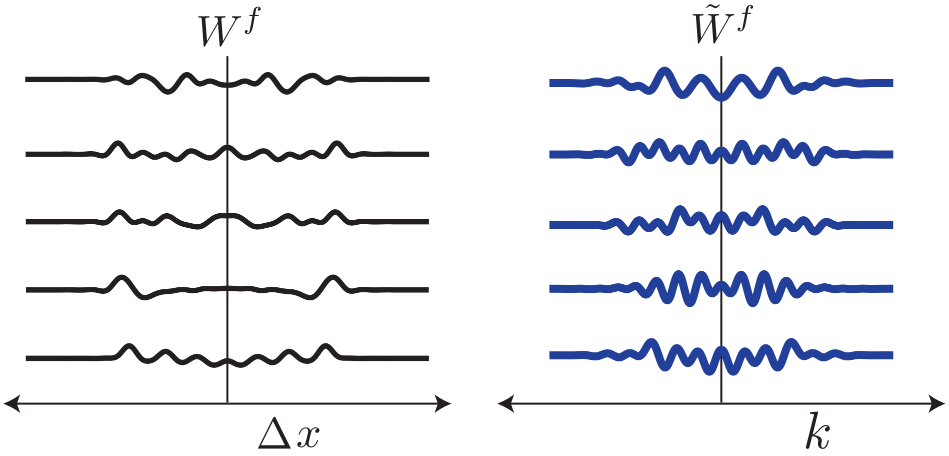


Figure 9: Randomly constructed fixed-scale interactions (left column) and their fourier tranforms (right column), in addition to the hand-designed ones in Fig.4a, that give $\phi = 0$.

to small k values — this may manifest as deviations in the largest period grid module away from Eq. 34.

3.5.1 Caveats

Clearly there exist *simple* kernels with Fourier transforms that are not given by $\mathcal{F}W^f(k) \approx \cos(kd - \phi)$. For example the Gaussian kernel, $W^f(\Delta x) = \exp[-\Delta x^2/(2d^2)]/(d\sqrt{2\pi})$ is a *simple* decaying kernel (since it has only a single scale d). Yet, its Fourier transform is simply $\mathcal{F}W^f(k) = \exp[-k^2d^2/2]$, which has only a single maximum! However, as we have shown earlier, there exist kernels that are arbitrarily close to the Gaussian kernel, whose Fourier transforms are given by powers of trigonometric functions, and hence have multiple regularly-spaced maxima with a spacing of $\sim 1/d$. Similarly, there exist additional *simple* functions (*62, 63*), $f(\Delta x)$, (like the Gaussian kernel) whose Fourier transforms $\mathcal{F}f(k)$ have a small number of maxima. We hypothesize that for all such functions $f(\Delta x)$ there exist *simple* kernels $g(\Delta x)$ that are arbitrarily close to $f(\Delta x)$ and possess regularly spaced maxima.

3.6 Period ratios

Having demonstrated analytically that *simple* kernels result in a sequence of period ratios given by Eq. (34), we now address the question of the mean period ratio over the sequence and over different values of ϕ. In the main text we have demonstrated that setting $\phi = 0$ results in a detailed period ratio sequence that is in close agreement with the sequence of experimentally observed values. Here we consider the period ratios obtained for other values of ϕ, to demonstrate that the experimental observation of mean period ratios being approximated by 1.4 (*34*) emerges naturally from our setup.

From Eq. (34), we obtained that the period ratio, $r_m = \lambda_{m+1}/\lambda_m$ can be written as

$$r_m = 1 + 1/(m + f), \tag{52}$$

where $f = \phi/(2\pi)$. We ignore $m = 1$, since that results in a period ratio close to 2, which does not correspond to experimental observations. Averaging the period ratio over the next 4 modules (corresponding to r_m for $m \in \{2 \dots 4\}$) results in

$$\langle r_m \rangle_m = 1 + \frac{1}{3}\left(\frac{1}{\phi+2} + \frac{1}{\phi+3} + \frac{1}{\phi+4}\right) \tag{53}$$

As can be seen in Fig. 10, this mean period ratio lies in the range [1.3,1.45], indicating that at all values of ϕ, the period ratio obtained from Eq. (34) matches well with experimental observations. The average of these period ratios over all values of ϕ can also be calculated as

$$\langle r_m \rangle_{\phi,m} = 1 + \frac{1}{3}\left[\log\left(\frac{5}{3}\right) + \log\left(\frac{7}{5}\right) + \log\left(\frac{9}{7}\right)\right] \tag{54}$$

which is approximately equal to 1.37.

3.7 Module size; number of modules as a topological quantity

As discussed in the main text, peak-selection for modularization is a highly robust mechanism that is largely indifferent to system parameters such as the the particular forms of the fixed-scale interaction and the shape of the gradient. Here we provide an analysis of the number

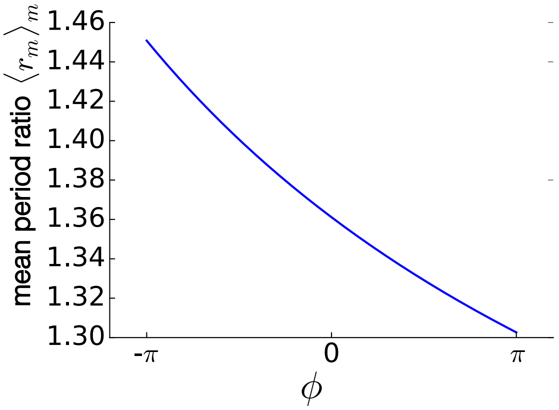


Figure 10: **Mean grid-period ratios** Ratios of grid periods averaged over 4 modules as a function of the phase shift ϕ in Eq. (34)

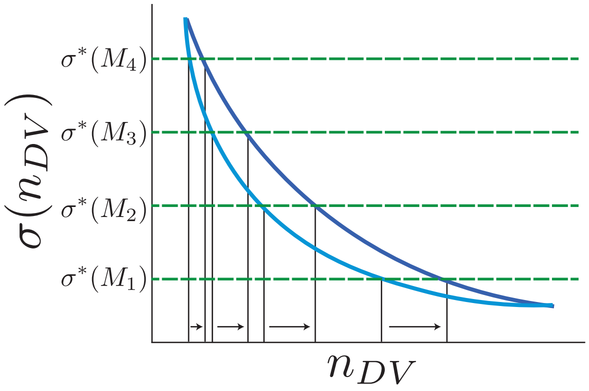


Figure 11: **The number of modules is a topological invariant and hence invariant to the gradient shape (the function** $\sigma(x)$**)** since it is set only by the fixed interaction width and the extremal/endpoint values of $\sigma(x)$, even though changing the gradient shape shifts the locations of module boundaries.

of modules, the scaling of module sizes, and the positions of module boundaries, which also exhibit the same robustness. Further, we also describe how this robustness may be interpreted as arising from a topological origin, similar to topological robustness in other physical systems like the quantum hall effect.

Recall that for the continuously graded kernel $W^g_x(\Delta x)$ with characteristic spatial scale $\sigma(x)$ at position x, the wave-vector of the formed pattern was proportional to $1/\sigma(x)$:

$$k^{*g}(x) = \eta/\sigma(x), \tag{55}$$

where η is an x-independent constant that depends on only the particular form of the graded kernel. Let the spatial extent of the system be $x \in [0, L]$, with $\sigma(x)$ monotonic such that $\sigma_{\text{min}} = \sigma(0) \leq \sigma(x) \leq \sigma(L) = \sigma_{\text{max}}$.

We assume for simplicity that the fixed-scale lateral interaction is a *simple* kernel, such that $\mathcal{F}W^f(k) \sim \cos(kd + \phi)$. Thus, the local maxima generated by $\mathcal{F}W^f(k)$ occur at $k_n \approx (2n\pi - \phi)/d$, where n are the natural numbers. As discussed in the main text, each of these local maxima is 'selected' in turn by the moving broad peak of the Fourier transform of the graded kernel, whose position according to Eq. 55 occurs at $k^{*g}(x) = \eta/\sigma(x)$.

Notably, the selected maximum k_m will be robust to small perturbations in the selection function $\mathcal{F}W^g_x(k)$, since k_m will remain quantized to one of the discrete values prespecified by the set $\{k_n \| n \in \mathbb{N}\}$. In this sense, the chosen maximum k_m (and hence the corresponding module) presents the hallmarks of a topologically protected state (*47*). The topological number corresponding to a given module is the module number m, which is a topological invariant similar to a winding number (*47*)[2].

The set of modules expressed through the length of the system corresponds to the set of local maxima k_n that lie within the range $[\eta/\sigma_{\text{max}}, \eta/\sigma_{\text{min}}]$ that is delineated by the range of

[2] Note that in our convention the module number m is ordered such that the largest grid period module is the first module. This is opposite to the numbering usually used in the literature, such as in (*34*).

peak positions of the graded interaction. It follows that the maxima selected by the graded interaction obey:

$$\frac{\eta}{\sigma_{\text{max}}} \leq \frac{2n\pi - \phi}{d} \leq \frac{\eta}{\sigma_{\text{min}}}. \tag{56}$$

Thus, the set of formed modules are determined by the set of integers n that fit in the following interval:

$$\frac{\phi + \eta d/\sigma_{\text{max}}}{2\pi} \leq n \leq \frac{\phi + \eta d/\sigma_{\text{min}}}{2\pi} \tag{57}$$

and hence the number of modules ν_{mod} is:

$$\text{\# modules} \equiv \nu_{mod} = \left\lfloor \frac{\phi + \eta d/\sigma_{\text{min}}}{2\pi} \right\rfloor - \left\lceil \frac{\phi + \eta d/\sigma_{\text{max}}}{2\pi} \right\rceil = \left\lfloor \frac{\phi + k^{*g}(0)d}{2\pi} \right\rfloor - \left\lceil \frac{\phi + k^{*g}(L)d}{2\pi} \right\rceil \tag{58}$$

where $\lfloor\ \rfloor$, $\lceil\ \rceil$ indicate the floor and ceiling operations, respectively.

The above result leads to the following observations: First, the central quantity essential for determining the number of modules is the difference in the integer ratios of the fixed-scale interaction width to the extremal lateral interaction widths, $d/\sigma_{\text{min}}, d/\sigma_{\text{max}}$. Second, the number of modules depends only on the end-point values $\sigma_{\text{min}}, \sigma_{\text{max}}$ of the smoothly varying width $\sigma(x)$ the graded interaction; notably, it does not depend on the detailed shape of $\sigma(x)$. Moreover, if $\sigma_{\text{min}}, \sigma_{\text{max}}$ are varied smoothly (while d is held fixed), or if d is varied smoothly (while $\sigma_{\text{min}}, \sigma_{\text{max}}$ are held fixed), the number of modules will remain fixed, until the change becomes large enough to accommodate one additional or one less module. Thus, the number of modules is also a topological invariant of the system, through the module number m. Third, the number of modules does not depend on the system size L, or the number of neurons n_{DV} the system is discretized into (cf. SI Fig. 11). Fourth, since the average module size will be L/ν_{mod}, the module sizes are extensive in L. Thus, for sufficiently large L, the module sizes can be orders of magnitude larger than the scales of the lateral interaction d and σ.

Note that the above argument on topological robustness of the modularization of the system is not restricted to the case of *simple* fixed-scale kernels. Indeed, for any fixed-scale interaction W^f, the topological number m for any given expressed module will correspond to selecting the m^{th} maximum of $\mathcal{F}W^f(k)$, for $k > 0$.

3.7.1 Module boundary locations

Following the peak-selection arguments made earlier, the module boundaries will occur at spatial locations that have $k^{*g}(x)$ in between k_n and k_{n+1} (the specific location will depend on the particular forms of the kernels). As a zeroth order approximation, we can assume that the module boundaries will occur near $(k_n + k_{n+1})/2$,

$$k^{*g}(x_{\text{boundary}}) \approx \frac{(2n+1)\pi - \phi}{d} \tag{59}$$

and thus

$$x_{\text{boundary}} \approx \sigma^{-1}\left(\frac{\eta d}{(2n+1)\pi}\right). \tag{60}$$

where σ^{-1} is the inverse function of $\sigma(x)$, $\sigma^{-1} \circ \sigma(x) = x$. Thus, while the specific positions of the module boundaries are dependent on the shape of the gradient $\sigma(x)$, qualitative features such as the number of modules, module periods and module sizes are indifferent to the particular forms of the gradient (cf. Fig. 11).

3.8 2D analysis

We have presented a majority of the above analysis for the case of one-dimensional grid cells. Here we briefly present the analogous computations for the Fourier transforms in two dimensions. We first demonstrate a classical result relating the Fourier transform of radially symmetric functions to the Hankel transform, which we shall then use to compute the relevant transforms.

Consider the Fourier transform of a function $f(\mathbf{x}) = f(x, y)$

$$\mathcal{F}f(\mathbf{k}) = \int f(\mathbf{x})e^{i\mathbf{k}\cdot\mathbf{x}}d\mathbf{x}$$
$$\mathcal{F}f(k_x, k_y) = \int f(x, y)e^{ik_x x + ik_y y}dxdy.$$

Define polar coordinates in real and Fourier space such that:

$$x = r\cos\theta$$
$$y = r\sin\theta$$
$$k_x = k\cos\phi$$
$$k_y = k\sin\phi$$

This leads to the dot product $\mathbf{k}\cdot\mathbf{x}$ to be simplified as

$$\begin{aligned} k_x x + k_y y &= rk(\cos\theta\cos\phi + \sin\theta\sin\phi) \\ &= rk\cos(\theta - \phi) \end{aligned}$$

Thus,

$$\mathcal{F}f(k_x, k_y) = \mathcal{F}f(k, \phi) = \int_0^\infty \int_0^{2\pi} rdrd\theta f(r, \theta)e^{ikr\cos(\theta-\phi)}$$

In all cases of interest, the function f is a kernel, and is hence a radially-symmetric real function $f(r, \theta) = f(r)$. Similarly, the Fourier transform $\mathcal{F}f$ will also be a real radially-symmetric function $\mathcal{F}f(k, \phi) = \mathcal{F}f(k)$. Thus

$$\begin{aligned} \mathcal{F}f(k) &= \int_0^\infty \int_0^{2\pi} rdrd\theta f(r)e^{ikr\cos(\theta-\phi)}, && (61) \\ &= \int_0^\infty rdrf(r)\int_0^{2\pi} e^{ikr\cos(\theta-\phi)}d\theta, && (62) \\ &= 2\pi\int_0^\infty rf(r)J_0(kr)dr, && (63) \end{aligned}$$

where J_0 is the Bessel function of the first kind, defined by

$$J_0(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{ix\cos(\theta-\phi)}d\theta.$$

Equation (63) defines the Hankel transform (of order zero) of $f(r)$ — the radial component of the Fourier transform of the kernel $f(\mathbf{x})$ is simply the Hankel transform of $f(|\mathbf{x}|)$.

For the localized gaussian secondary interaction, we can calculate the Fourier transform analytically.

$$\begin{aligned}\mathcal{F}W_{\text{local}}(k) &= 2\pi\int_0^{\infty} r\left[\alpha_E e^{-r^2/2\sigma_E^2} - \alpha_I e^{-r^2/2\sigma_I^2} + \alpha_S e^{-(r-d)^2/2\sigma_S^2}\right] J_0(kr)dr \\ &= 2\pi\left[\alpha_E\sigma_E^2 e^{-k^2\sigma_E^2/2} - \alpha_I\sigma_I^2 e^{-k^2\sigma_I^2/2} + \alpha_S J_0(kd)\sigma_S^2 e^{-k^2\sigma_S^2/2}\right]\end{aligned}$$

We can also analytically calculate the Fourier transform for a box-like interaction:

$$\begin{aligned}\mathcal{F}W_{\text{diffuse}}(k) &= 2\pi W\int_0^{d} rJ_0(kr)dr \\ &= \frac{2\pi W}{k^2}\int_0^{kd}\rho J_0(\rho)dr \\ &= \frac{2\pi W}{k^2}\left[kdJ_1(kd)\right] \\ &= \frac{2\pi W d^2 J_1(kd)}{kd}\end{aligned}$$

We can similarly also define a two-dimensional equivalent of the decaying kernel, as the convolution of the half-sized circular box kernel with itself. Thus, by applying convolution theorem to the result on diffuse kernels we obtain

$$\mathcal{F}W_{\text{decaying}}(k) = \left[\frac{\pi W dJ_1(kd/2)}{k}\right]^2.$$

Note that $J_0(x)$ and $J_1(x)$ display qualitatively similar behavior to $\cos(x)$ and $\sin(x)$ respectively, apart from an amplitude modulation of the peaks — particularly, we note that the Bessel functions display approximately periodic maxima, which was the central property required for

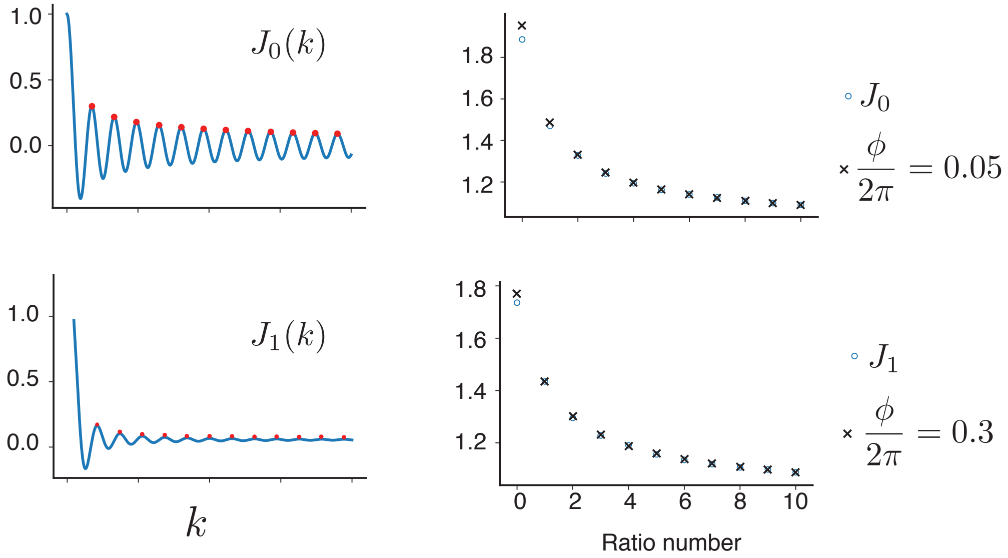


Figure 12: Bessel functions (left column) and period ratios for Bessel function maxima (right column) with their best-fit values of ϕ for the period ratios corresponding to Eq. (34)

all of our results on modularization and peak selection to apply. We demonstrate this in Fig.12, where we show that the maxima of the Bessel functions are approximately periodic, and fit the form of Eq. (34) well. In particular, note that the best-fit value of ϕ for $J_0(k)$ is approximately 0, which is similar to $\cos(k)$, and the best-fit value of ϕ for $J_1(k)$ is approximately $\pi/4$, which is similar to $\sin(k)$.

We implemented a 2d simulation that generates 3 discrete modules as shown in Figure 14. For computational feasibility, the simulation was performed in 2 parts: one with $x \in [0, 0.6N_x^{2d}]$ and the other with $x \in [0.6N_x^{2d}, N_x^{2d}]$. The weight matrices for each network were of size 100x1000 each. The weight matrix for a single large 100x2000 network would have contained $4x10^{10}$ elements, which we found prohibitively difficult and slow to run.

Fig 15(a) shows another instance of a modular 2d network, the only difference being the value of d_{loc}, which changed from 50 to 45. Fig 15(b) shows the same simulation with 2 distinct random initializations. The pair of resulting modules in each simulation have different

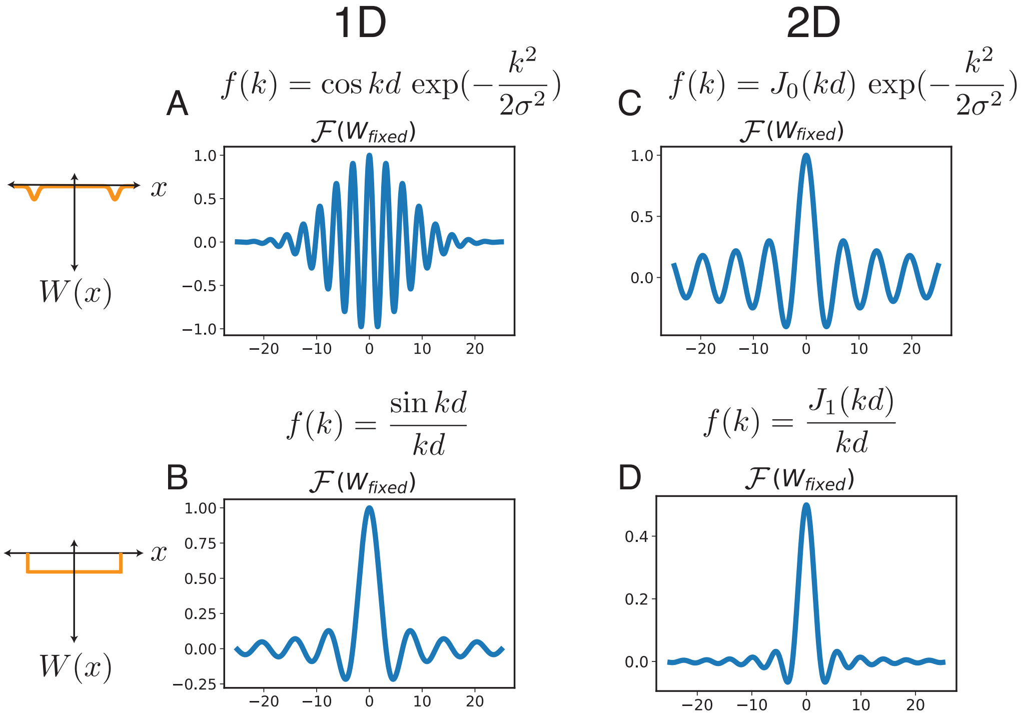


Figure 13: Fixed interactions(left, in orange) and their oscillatory Fourier transforms in 1D (left column) and 2D (right column).

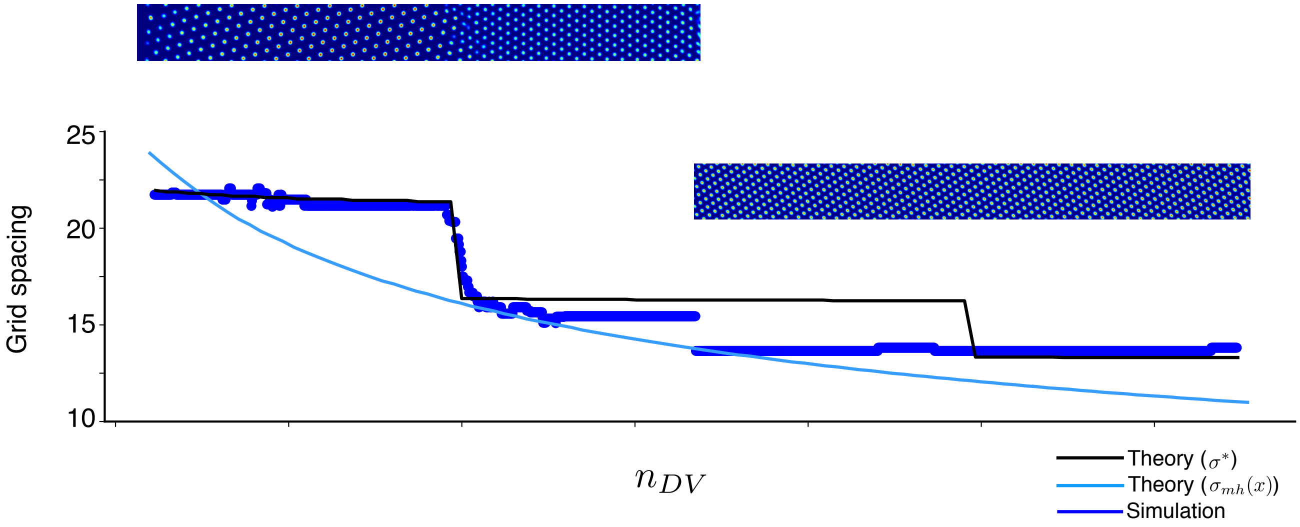


Figure 14: **2d simulation with 3 modules:** (top) Snapshots of population activity showing 3 discrete 2d grid modules, (bottom) plot of grid spacing and comparision with Hankel transform predictions. Grid spacing determined by calculating the (neural) spatial auto-correlation of the population firing activity.

relative orientations. Because finite size effects from our simulations also partially constrain the orientations of the modules (data not shown), we cannot make predictions about the relative orientations of the grid modules found in experiments (*34*).

3.9 Comparison of experimental observations with predicted period ratios

The general mechanism of peak-selection presented above describes how discrete modules can spontaneously arise in the presence of continuous gradients, by consideration of an additional fixed-scale lateral interaction W^f. However, this mechanism does not provide any testable predictions for the ratio of grid periods unless additional assumptions are made. If indeed we assume that W^f is a *simple* kernel, i.e., W^f is primarily defined by a single spatial scale, then we demonstrated in SI Sec. 3 that the period ratios will be given by the simple formula, Eq. 34. In this section, we show that experimental observations of grid periods largely appear to match

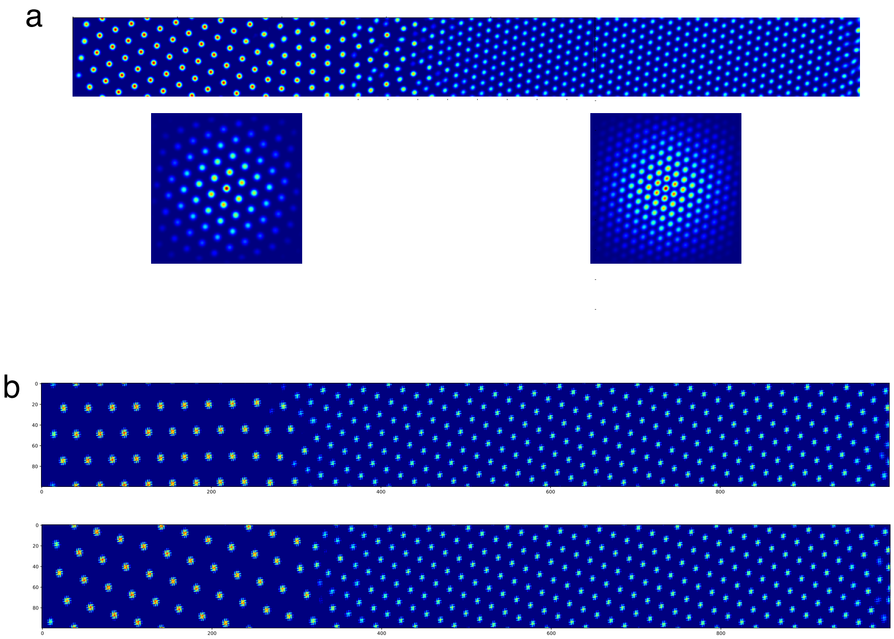


Figure 15: (a) Another instance of a spontaneously formed two dimensional network with parameters given in Table 4. (b) Two different random initializations of the network from Fig 2h show different relative orientations between the 2 formed modules.

our predicted period ratios for *simple* kernels with $\phi = 0$.

For verification of our main results on the predicted form of period ratios, we examine the literature for grid period measurements for multiple simultaneously measured grid modules in rats (*34, 64–66*). We note that a large fraction of experimental observations of grid cells with more than one module measure only two modules. For a single pair of grid periods λ_1 and $\lambda_2 > \lambda_1$, we can always explicitly solve for ϕ and m in Eq. (34), to obtain

$$\frac{\phi}{2\pi} = \left\{ \frac{\lambda_2}{\lambda_1 - \lambda_2} \right\}, m = \left\lfloor \frac{\lambda_2}{\lambda_1 - \lambda_2} \right\rfloor, \tag{64}$$

where $\{x\}$ represents that fractional part of x, and $\lfloor x \rfloor = x - \{x\}$ represents the integer part of x. Thus, a single ratio, because it can always be fit by Eq. (34), imposes no constraints on the accuracy of the expression.

It is possible to obtain a value of ϕ from Eq. (34) and a single pair of periods; however, the estimate obtained from a single pair is not robust: r_m depends too sensitively on ϕ. For example, in (*34*), Rat 13388 exhibits grid periods of ≈ 53.24 cm and ≈ 43.00 cm (as estimated from SI Fig. 12b in (*34*)); Eq. (34) then yields $\phi/(2\pi) = 0.199$. Assuming a very small measurement error of $\sim 0.5cm$ in the larger period, such that if it were 53.75 cm instead of 53.24, would yield ϕ exactly equal to zero. A simple sensitivity analysis of the magnitude of error in estimating ϕ can be performed from Eq. (64):

$$\delta\phi = 3\epsilon \frac{\lambda_2}{\lambda_1 - \lambda_2} \approx 3\epsilon m, \tag{65}$$

where ϵ represents the fractional error in the estimate of grid period. Thus, particularly for smaller grid periods (corresponding to larger m), even small errors in grid period estimation can result in a large error in ϕ, making the errorbars in the estimation of ϕ from a single pair of periods large.

To obtain results with significant statistical certainty, we focus our analysis on published experimental studies that measure at least 50 grid cells per animal, spanning at least 3 distinct

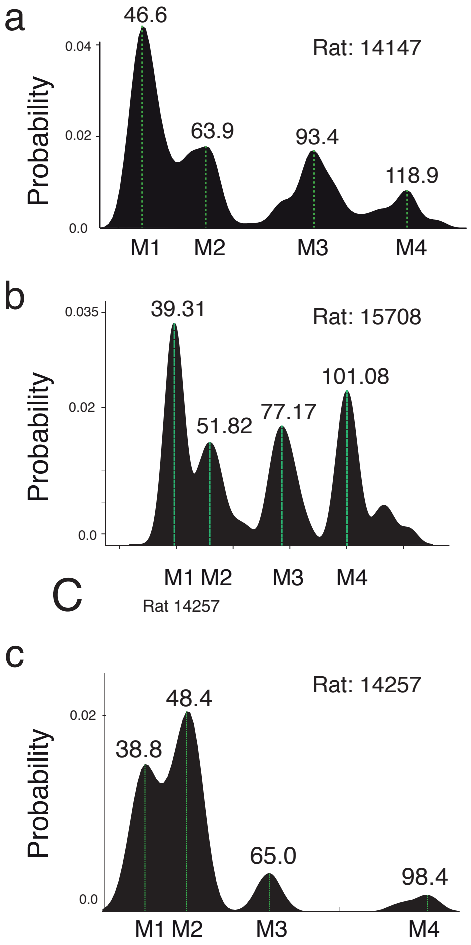


Figure 16: The 3 rats from Stensola *et al.* with 4 modules and their corresponding periods.

modules. This restriction results in grid period data sets for three rats — we present kernel density estimates of the module periods for each of them in Fig. 16 (Fig. 16c corresponds to the data presented in the main text in Fig. 4).

We have already demonstrated in Fig. 4 that Rat 14257 presents an extremely accurate match to the period ratio prediction for $\phi = 0$ (i.e., predicted period ratios of 2, 3/2, 4/3, 5/4,...); in addition, Rat 14147 (observed period ratios of 1.27 , $1.46 \approx 3/2$, $1.37 \approx 4/3$) and Rat 15708 (observed period ratios of 1.31, $1.49 \approx 3/2$, $1.32 \approx 4/3$) also match $\phi = 0$ very well (R^2 values of 0.999, 0.979, and 0.968 for Rats 14257, 15708, 14147 resp.) for all grid modules except for the module with the largest period.

Why is there an observed discrepancy for the largest grid module? Our predictions for grid period ratios Eq. (34) are for the case of *simple* kernels that have a single spatial scale. A discrepancy at only the largest grid module may thus be suggestive of fixed-scale interactions that are primarily described by a single scale, with an additional low frequency perturbation at a larger spatial scale. Alternately, this discrepancy may be a result of the approximation made in Sec. 3.5 that would only affect the largest grid period module. However, note that (particularly for Rats 14147 and 14257) there are relatively few grid cells observed from this largest period module, and the resulting uncertainty in period estimation may instead contribute to the error. In sum, apart from the possibility of some additional low frequency perturbations, the experimental data for rats with several simultaneously observed grid modules is largely consistent with the predicted period ratios for simple kernels with $\phi = 0$.

Skipped modules: Sometimes, neural recordings can miss a module. This can cause a large deviation from our predictions. For example, for a set of 5 modules following period ratios M4/M5 = 1.20, M3/M4 = 1.25, M2/M3 = 1.33, M1/M2 = 1.5. If recordings had missed module M4, the measured ratios would be M1/M2 = 1.5, M2/M3 = 1.33, M3/M5 = 1.5.

However, we do note that available data on multiple modules with a statistically large number of grid cells per module are quite sparse. To obtain further verification of our theoretical results, including the prediction of Eq. (34) and even more specifically the hypothesis that ϕ is close to zero, additional data with multiple simultaneously observed grid modules will be important.

4 Lyapunov Function

The energy function of continuous time neural networks can be written as (*67*):

$$E(\mathbf{s}) = -\frac{1}{2}\sum_{ij} s(i)W_{ij}s(j) + \sum_{i}\frac{1}{R_i}\int_0^{s(i)} \phi^{-1}(s)\,ds - \sum_i I_i s(i), \tag{66}$$

where s represents a vector of the synaptic activation at each neuron in the network, and I_i is the input bias to neuron i. For simplicity and since linear analysis does a remarkably good job in predicting the formed modules, let us restrict ourselves to the case of $\phi(x) = x$. Also, since the system is locally translationally invariant, we know that the dominant modes are going to be periodic. Hence, we may evaluate the energy function of the network dynamics (in the linearized regime) by assessing the energy of the periodic neural activity modes:

$$\mathbf{s}_k(\mathbf{x}) = A\sin(\mathbf{k}\cdot\mathbf{x} + \delta) + B, \tag{67}$$

where $\mathbf{k} = k\hat{\mathbf{k}}$ is an arbitrary Fourier space vector, and A,B and δ are arbitrary constants. For these modes, we can write the energy function in the continuum limit as:

$$E[s_{\mathbf{k}}(\mathbf{x})] = -\frac{1}{2}\int d\mathbf{x}d\mathbf{x}'W(\mathbf{x},\mathbf{x}')s_{\mathbf{k}}(\mathbf{x})s_{\mathbf{k}}(\mathbf{x}') + \frac{1}{2}\int d\mathbf{x}s_{\mathbf{k}}(\mathbf{x})^2$$

Assuming that the system size L is large,

$$\begin{aligned}
2E[s_{\mathbf{k}}(\mathbf{x})] &= -\int W(\mathbf{x}-\mathbf{x}')[A\sin(\mathbf{k}\cdot\mathbf{x}+\delta)+B][A\sin(\mathbf{k}\cdot\mathbf{x}'+\delta)+B]d\mathbf{x}d\mathbf{x}' + \int[A\sin(\mathbf{k}\cdot\mathbf{x}+\Delta)+B]^2 \\
&= -A^2\int d\mathbf{u}d\mathbf{v}W(\mathbf{u})\cos(\mathbf{k}\cdot\mathbf{u}) + A^2\int d\mathbf{u}d\mathbf{v}W(\mathbf{u})\cos(2\mathbf{k}\cdot\mathbf{v}+\delta) + B^2\int d\mathbf{x}d\mathbf{x}'W(\mathbf{x}-\mathbf{x}') + L \\
&= -A^2L\int d\mathbf{u}e^{i\mathbf{k}\cdot\mathbf{u}}W(\mathbf{u}) + A^2\int d\mathbf{u}W(\mathbf{u})\int d\mathbf{v}\cos(2\mathbf{k}\cdot\mathbf{v}+\delta) + B^2\int d\mathbf{u}d\mathbf{v}W(\mathbf{u}) + L(A^2/2 + \\
&= -A^2L\tilde{W}(k) + LB^2\bar{W} + L(A^2/2 + B^2), \\
&= -\text{constant}_1 \times \tilde{W}(k) + \text{constant}_2
\end{aligned}$$

where have used the simple trigonometric identity, $2\sin(C)\sin(D) = \cos(C-D) - \cos(C+D)$, and a change of variables, $\int d\mathbf{x}d\mathbf{x}' = (1/2)\int d(\mathbf{x}-\mathbf{x}')d(\mathbf{x}+\mathbf{x}') = \int d\mathbf{u}d\mathbf{v}$, with $\mathbf{u} = \mathbf{x}-\mathbf{x}'$ and $\mathbf{v} = \frac{1}{2}(\mathbf{x}+\mathbf{x}')$.

Thus, we obtain that the energy function $E[s_{\mathbf{k}}]$ is a simple linear function of the Fourier transform $\tilde{W}(k)$ of the recurrent weight matrix. The minimum energy solution corresponds to the Fourier mode that maximizes $\tilde{W}(k)$. In other words, the dynamics is dominated by the k^* that maximizes $\tilde{W}(k)$. This result, derived from an energy landscape perspective, is equivalent to the result in Eq. (16), which we obtained earlier via perturbation analysis.

5 General formulation of module formation dynamics: Discrete peak selection via loss minimization

In Sec. 4, we demonstrated how the pattern formation on the neural sheet can be derived via an energy minimization approach. Here, we use an energy landscape view to describe how loss function minimization results in modular solutions.

The key components for spatially modular solutions to arise from energy minimization are as follows: 1) A spatially-independent loss function $f(\theta)$ with multiple local maxima and minima; 2) A gradient in a spatially-dependent variable, $\theta_0(x)$; and 3) A coupling between the system parameters θ and θ_0, that results in a combined loss function

$$L(\theta, x) = f(\theta) + \alpha||\theta - \theta_0(x)||^2 \tag{68}$$

Under appropriate constraints on $f(\theta)$, solving the following optimization at each x

$$\theta^*(x) = \arg\max_\theta L(\theta, x) \tag{69}$$

will produce discrete, step-like changes as a function of x. This happens because the smooth minimum given by the $||\theta - \theta_0(x)||^2$ term effectively selects one of the local minima in $f(\theta)$ as the global minimum. As the function $||\theta - \theta_0(x)||^2$ slides smoothly along with x,the peak of $f(\theta)$ selected as the global minimum remains the same for some time, then jumps abruptly. These step-like changes are modular solutions to the global optimization problem. The energy function defined in Eq. (68) can be viewed as a regularized optimization problem, with the spatially-dependent regularizer $||\theta - \theta_0(x)||^2$ acting as a prior that selects one of the minima of $f(\theta)$ at each location (Fig. 17).

The correspondence of this general picture with the peak selection mechanism described in the main text follows directly with the following identifications: the spatially independent nonlinear loss function $f(\theta)$ with the fixed-scale interaction W^f; the spatially varying parameter prior $\theta_0(x)$ with the graded scale $\sigma(x)$ of the pattern-forming kernel; the combined loss $L(\theta, x)$ with the full kernel W_x; and the spatially-varying, multi-step-like set of optima $\theta^*(x)$ with the grid periods $\lambda^*(x)$, respectively. Similar to peak selection for grid cells, the formed modules in this generalized setting will also inherit topological robustness and stability.

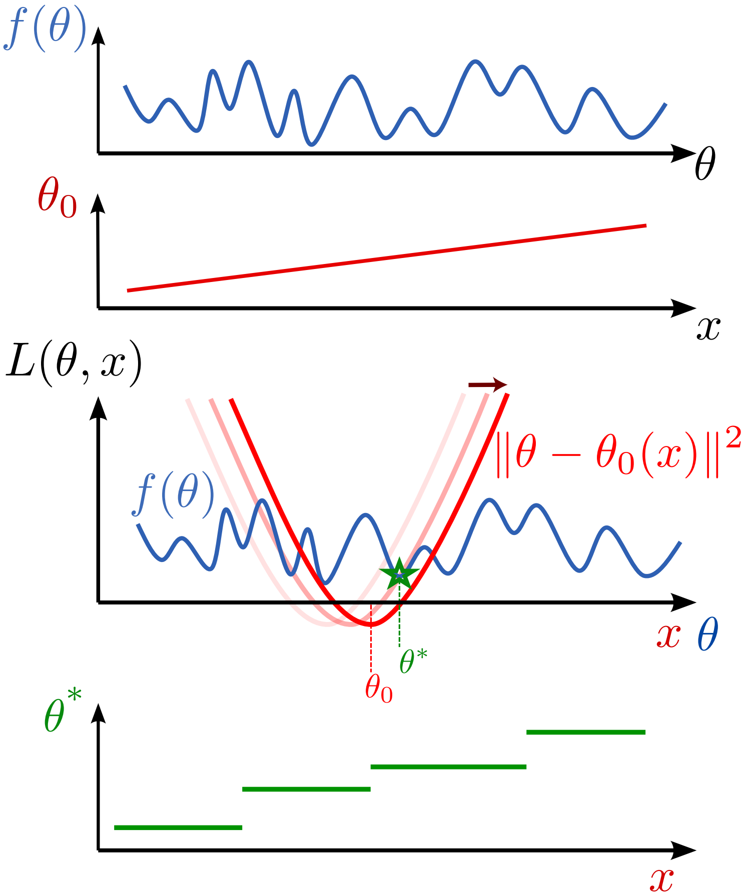


Figure 17: **A general setting for peak-selection** Assuming a loss function $f(\theta)$ (blue) and a spatially dependent quantity θ_0 (red), a combined loss function $L(\theta, x)$ can be constructed such that the x-dependent optimizer of $L(\theta, x)$ will be modular (green), since it will be constrained to correspond to one of the minima of $f(\theta)$.

6 The emergence of modules corresponds to the formation of localized eigenvectors

As has been observed before (*68*), a neural network endowed with slowly varying local interactions shows diverse timescales that are spatially localized: different parts of the network respond with disparate temporal dynamics.We also find a localization of eigenvectors in our multi-module grid network, Fig. 18A. Similar to (*68*), our interaction matrix has a locally circulant form (due to the slowly varying gradient in lateral inhibition width).

We find that in the resulting set of localized eigenvectors, each has a different but constant period, Fig. 18B. These periods exactly match the spatial periods of the modules formed in steady state. In sum, the locally circulant matrix gives rise to eigenvector localization, and the localized eigenvectors correspond to the modules.

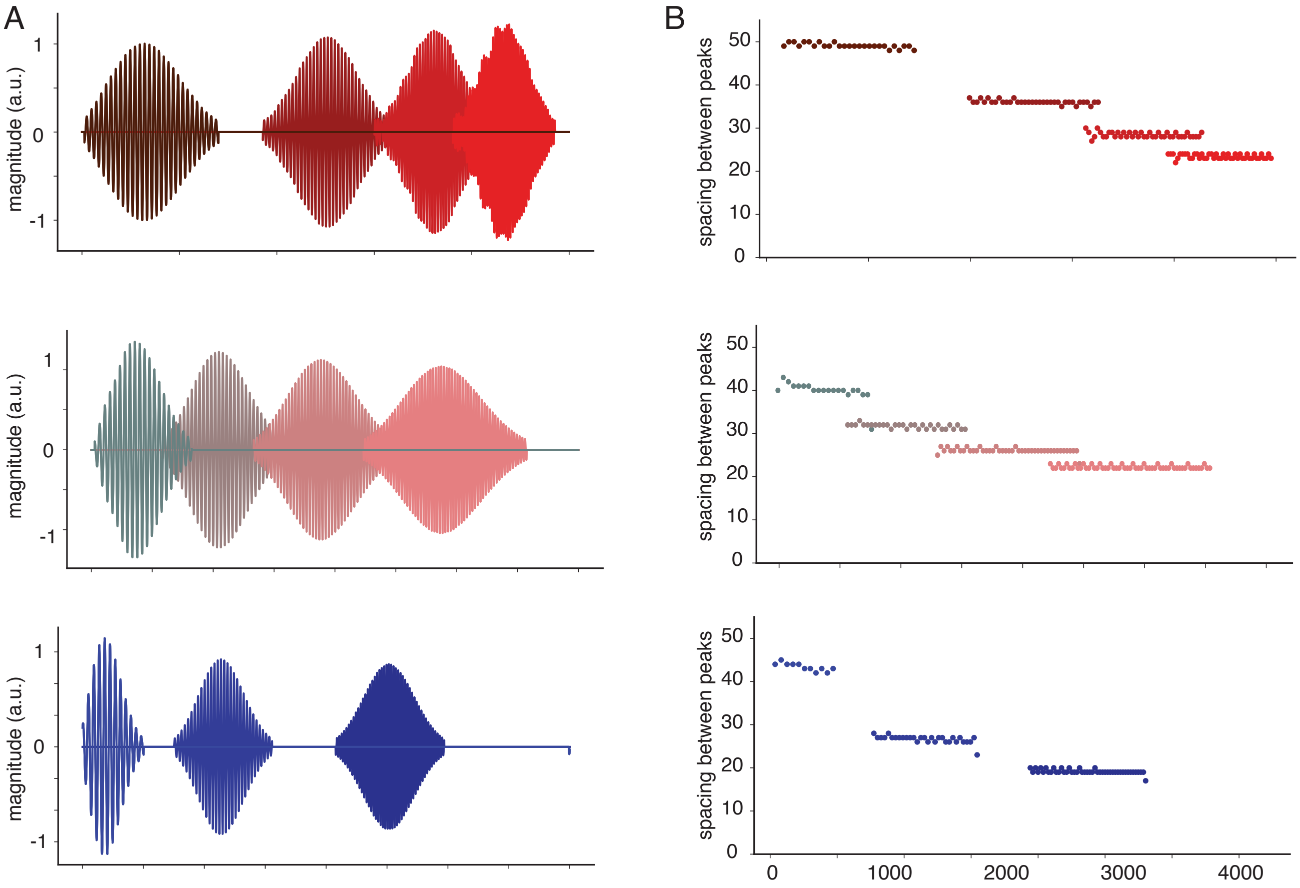


Figure 18: **Localization of eigenvectors**: A) Eigenvectors of various one-dimensional interaction weight matrices along with the corresponding inter-peak spacings are localized, B) The periodicity within an eigenvector is constant.